Language, time, and the mind:
Understanding human language processing using continuous-time deconvolutional regression

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy
in the Graduate School of The Ohio State University

By
Cory Shain

Graduate Program in The Department of Linguistics
The Ohio State University
2021

Dissertation Committee:
Professor William Schuler, Advisor
Professor Micha Elsner
Professor Subhadeep Paul

Abstract

The predictions of theories of incremental human sentence processing are often cached out in word-by-word measures, but the mind is a dynamical system that responds to language in real time. As a result, there may be a complex alignment between the properties of words in language and the influence those properties exert on measures of human cognition. One possible aspect of this alignment is *temporal diffusion*, whereby sentence processing effects are realized in a delayed manner (Mitchell, 1984). For example, because of real-time bottlenecks in human information processing (Mollica and Piantadosi, 2017), encountering a surprising word may increase cognitive load not only at that word, but also on subsequent words as the rest of the experiment unfolds (Smith and Levy, 2013).

In this thesis, I argue that effect timecourses are of direct or indirect importance to many central questions in psycholinguistics, that failure to account for these timecourses can have large impacts on the results of scientific hypothesis tests, and that existing discrete-time approaches to estimating and controlling for effect timecourses are not well adapted to many experimental designs in psycholinguistics, which involve *non-uniform* time series in which events (words) have variable duration. I define and implement an analysis technique that addresses these concerns: continuous-time deconvolutional regression (CDR). CDR estimates continuous-time functions that describe the shape and extent of a predictor's influence on the response over time, thus directly illuminating and controlling for temporally diffuse effects. I show empirically that CDR accurately recovers ground-truth models from synthetic data and provides plausible and detailed estimates of temporal structure in human data that generalize better than estimates obtained using existing techniques.

I apply CDR to measures of naturalistic sentence processing in order test several theoretical questions in psycholinguistics. In one study, I present a CDR analysis that challenges

an existing hypothesis that human reading times exhibit distinct effects of a word's overall frequency vs. its predictability from context (Staub, 2015), instead finding no evidence for such a distinction in naturalistic reading. In another study, I present a CDR analysis showing evidence from naturalistic functional magnetic resonance imaging (fMRI) data that human predictive mechanisms for language processing are sensitive to the syntactic features of sentences, and that these predictive mechanisms reside primarily in regions of the brain that are selective for language processing, rather than in regions involved in domain-general executive control. In a final study, I reanalyze the fMRI data with respect to theory-driven measures of working memory retrieval difficulty and report significant retrieval effects over strong controls for word predictability, but only in language-selective regions. This result supports a core role for language-specific working memory resources in typical language comprehension.

Finally, I define and implement a deep neural generalization of CDR — the continuous-time deconvolutional regressive neural network (CDRNN) — that relaxes many of CDR's simplifying assumptions while retaining its deconvolutional interpretation. I show empirically that CDRNN generalizes better than CDR and other baselines on human data, and that it supports novel insights into the functional form of effects and effect interactions over time that are difficult to obtain using other methods. Based on these results, I advocate both increased attention to the temporal dimension in psycholinguistic regression analyses and the use of CDR to understand the dynamics of human sentence processing.

Acknowledgements

This thesis is the result of the time, effort, expertise, and friendship of many other people. While I could never pay back their investment, I hope this small token of thanks can stand in for my heartfelt appreciation of all of them.

I'd first like to thank my inimitable advisors, William Schuler and Micha Elsner. From his welcoming response to my initial cold-call email asking if I could help out in his lab to "explore computational linguistics", to his patient advice and assistance with all issues academic and personal over years of weekly meetings, William modeled professionalism and skill in advising, empowering me to pursue the research directions that interested me while also honing and shaping those directions to ensure I stayed on track. I also benefited greatly from Micha's unparalleled clarity of thinking and interdisciplinary breadth of knowledge, which both strengthened my research and inspired my own development as a scholar. I'd also like to thank my "unofficial" third advisor, Ev Fedorenko, who invited me early in my career into collaborations that were beyond my expertise, and who has patiently and encouragingly supported my forays into cognitive neuroscience.

I'd also like to thank the other faculty members who served on my committees, especially Subhadeep Paul, who greatly helped expand my awareness of connections between my work and related approaches in statistics, and Cynthia Clopper, whose feedback consistently sharpened the clarity of my writing and strengthened connections to related ideas.

I'm so thankful for the support and insight of the various collaborators I've been privileged to work with during my time as a graduate student. During my first year, Marten van Schijndel, at the time in his final year of Ohio State's linguistics PhD, took me on as an unofficial mentee, and his kindness, humor, and dedication to first-rate science had a profound impact on my own aspirations as a scholar. Ted Gibson graciously allowed

himself to be roped into multiple collaborations based on my out-of-the-blue email without even meeting me. And Evan Jaffe has allowed me to join him on a journey of discovery into the complexities of incremental coreference processing. I wouldn't have made it this far without the friendship and support of many other people with whom I've worked on projects, written papers, or simply shot the breeze, including Evan Thomas, Idan Blank, Richard Futrell, Lifeng Jin, Greta Tuckute, and Frank Mollica. I am additionally grateful for the contributions to my work and professional development of my master's thesis advisors Judith Tonhauser and Peter Cullicover, members of the Clippers and CaCL discussion groups at Ohio State, Andrea Sims and my classmates in the Spring 2019 seminar on computational morphology, and scientists from other institutions who have shown an interest in and provided feedback on the research presented in this thesis, including Tal Linzen, Roger Levy, and Stefan Frank, as well as members of their lab groups.

Finally, I owe the biggest debt of gratitude to the members of my family who have made profound sacrifices to enable my doctoral training, including my parents, Dave and Kay Shain, and my mother-in-law Janice Craig, who were all-hands-on-deck to give childcare assistance and encouraging words during our busiest times. And I'm most especially thankful for the love and understanding of my beautiful kids — Simon, Manny, Solenne, and Oscar — who are a constant font of life and joy, and for my tireless spouse and best friend Rachel, who bent over backward to love me, support me, and help me meet deadlines and handle conference travel despite the fact that she was herself a medical student/resident throughout my PhD. Words can't express my love, affection and gratitude for you.

Graduate funding acknowledgments:

- Presidential Fellowship, The Ohio State University
- National Science Foundation grant #1422987 to Micha Elsner
- Google Faculty Research Award to Micha Elsner

Vita

PhD, The Ohio State University	2021
MA, The Ohio State University	2009
BA, The Ohio State University	2009

Publications

Shain, C.; and Schuler, W. (to appear). Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*.

Wehbe, L.; Blank, I.; Shain, C.; Futrell, R.; Levy, R.; von der Malsburg, T.; Smith, N.; Gibson, E.; and Fedorenko, E. (to appear). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *Cerebral Cortex*.

Shain, C.; and Elsner, M. (2020). Acquiring language from speech by learning to remember and predict. In *Proceedings of the 24th Conference on Computational Natural Language Learning*: 195-214.

Jaffe, E.; Shain C.; and Schuler, W. (2020). Coreference information guides human expectations during natural reading. In *Proceedings of the 28th International Conference on Computational Linguistics*: 4587-4599.

Shain, C.; Blank, I.; van Schijndel, M.; Fedorenko, E.; and Schuler, W. (2020). fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138: 107307.

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*: 4086-4094.

Shain, C.; and Elsner, M. (2019). Measuring the perceptual availability of phonological features during language acquisition using unsupervised binary stochastic autoencoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*: 69-85.

Elsner, M.; Sims, A. D.; Erdmann, A.; Hernandex, A.; Jaffe, E.; Jin, L.; Johnson, M. B.; Karim, S.; King, D. L.; Lamberti Nunes, L.; Oh, B.; Rasmussen, N.; Shain, C.; Antetomaso, S.; Dickinson, K. V.; Diewald, N.; McKenzie, M.; and Stevens-Guille, S (2019). Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modeling*, 7(1): 53-98.

Shain, C.; and Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*: 2679-2689.

Shain, C.; van Schijndel, M.; and Schuler, W. (2018). Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Jaffe, E.; Shain, C.; and Schuler, W. (2018). Coreference and Focus in Reading Times. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*: 1-9.

Elsner, M.; Shain, C. (2017). Speech segmentation with a neural encoder model of working memory. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*: 1070-1080.

Mahler, T.; Cheung, W.; Elsner, M.; King, D.; de Marneffe, M.; Shain, C.; Stevens-Guille, S.; and White, M. (2017). Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*: 33-39.

Shain, C.; Bryce, W.; Jin, L.; Krakovna, V.; Doshi-Velez, F.; Miller, T.; Schuler, W.; and Schwartz, L. (2016). Memory-Bounded Left-Corner Unsupervised Grammar Induction on Child-Directed Input. In *Proceedings of the 26th International Conference on Computational Linguistics*: 964-975.

Shain. C.; van Schijndel, M.; Futrell, R.; Gibson, E.; and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*: 49-58.

Shain, C., and Tonhauser, J. (2010). The synchrony and diachrony of differential object marking in Paraguayan Guaraní. *Language Variation and Change*, 22(03), 321346.

Fields of Study

Major Field: Linguistics

Table of Contents

List of Figures

List of Tables

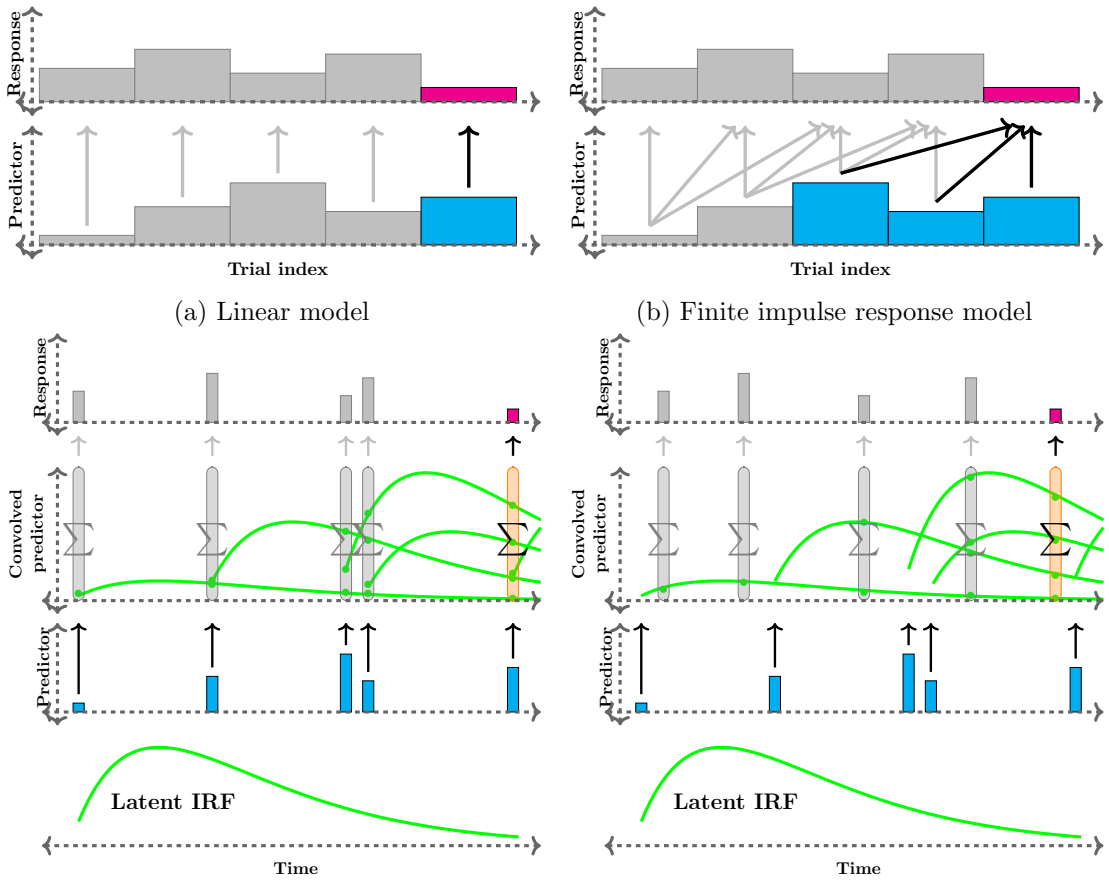
Part I

Language Processing and Time

Chapter 1

Introduction

Central questions in psycholinguistics concern the mental processes involved in incremental human language understanding: which representations are computed when, by what mental algorithms (Frazier and Fodor, 1978; Just and Carpenter, 1980; Abney and Johnson, 1991; Tanenhaus et al., 1995; Almor, 1999; Gibson, 2000; Coltheart et al., 2001; Hale, 2001; Lewis and Vasishth, 2005; Levy, 2008, *inter alia*)? Such questions are often studied by caching out a theory of language processing in an experimental stimulus, collecting human responses, and fitting a regression model to test whether measures show the expected effects (e.g. Grodner and Gibson, 2005). Because the human mind operates in real time and experiences computational bottlenecks of various kinds (Bouma and De Voogd, 1974; Ehrlich and Rayner, 1981; Mitchell, 1984; Mollica and Piantadosi, 2017), delayed effects in human sentence processing may be pervasive. In light of this consideration, this thesis advocates recasting many kinds of psycholinguistic regression analyses as a form of *dynamical systems modeling* (Ljung and Glad, 1994) that seeks to uncover how variables of interest influence human experimental measures over time. I will argue that widely-used analysis methods in psycholinguistics struggle to capture processing effects that linger over time (henceforth, *temporal diffusion*), with potential implications for interpretation, testing, and (consequently) scientific theory selection. I will propose a novel time series model — continuous-time deconvolutional regression (CDR) — that addresses this concern, and I will show empirically that it both closely recovers ground truth dynamics in synthetic data and finds detailed and plausible estimates of dynamics in human data from multiple experimental modalities, with better generalization error than standard statistical models.



(a) Linear model

(b) Finite impulse response model

(c) CDR (synchronous data)

(d) CDR (asynchronous data)

Figure 1.1: **Visual comparison of time series models.** In linear models (a), the response (y-axis) is independent of previous events, while in FIR models (b), previous events are assumed to be equidistant in time (x-axis). In CDR models (c and d), the response is a weighted sum of all previous events, with weights provided by the IRF as a function of continuous time. Because the IRF is continuous, the response can be queried at any point, permitting direct application to both synchronous (c) and asynchronous (d) stimulus and response measures.

I will additionally describe scientific insights afforded by CDR analysis into the role of lexical retrieval, prediction, and working memory in human sentence processing. Finally, I will describe and empirically evaluate a deep neural extension of CDR that relaxes many simplifying assumptions and permits detailed insights into cognitive dynamics that are difficult to obtain using existing methods. The conceptual and empirical advantages of CDR make

it a powerful new tool for studying the mind.

Temporal diffusion has been carefully studied in some psychological subfields. For example, a sizeable literature on fMRI has investigated the structure of the *hemodynamic response function* (HRF), which is known to govern the relatively slow response of blood oxygenation to neuronal activity (Boynton et al., 1996; Friston et al., 1998b; H. Glover, 1999; Ward, 2006; Lindquist and Wager, 2007; Lindquist et al., 2009). The HRF is an instantiation of the more general notion of *impulse response function* (IRF) from the field of signal processing (Madisetti, 1997), where the response $h * g$ of a dynamical system as a function of time is described as a convolution over time of an impulse h with an IRF g as shown in eq. 1.1, where τ is bound by the integral operation and ranges over the time interval $[0, t]$, and $h(\tau)$ is the impulse at time τ:[1]

$$(h * g)(t) = \int_0^t h(\tau)g(t-\tau)d\tau \tag{1.1}$$

The process of *deconvolution* seeks to infer the structure of g (the IRF) given that the impulses h (stimuli) and responses $h * g$ (experimental measures) are known.

CDR recasts the sequences of stimuli (predictors) and responses as convolutionally-related signals whose temporal relationship is mediated by one or more continuous-time IRFs with shape estimated from data. By convolving predictors with their estimated IRFs, as in eq. 1.1, the model can condition its predictions on the entire history of stimuli encountered up to a given point in an experiment, rather than e.g. on the properties of the current word alone. And by estimating the shape of the IRF from data, the model can reveal fine-grained patterns of temporal structure that are otherwise difficult to obtain.

Established techniques for IRF identification, including *finite impulse response* (FIR)

[1] Throughout this paper, vectors and matrices are notated in **bold** lowercase and uppercase, respectively (e.g. $\mathbf{u}$, $\mathbf{U}$). Objects with indexed names are designated using subscripts (e.g. $\mathbf{v}_r$). Vector and matrix indexing operations are notated using subscript square brackets, and slice operations are notated using $*$ (e.g. $\mathbf{X}_{[*,k]}$ denotes the k^{th} column of matrix $\mathbf{X}$). Hadamard (pointwise) products are notated using $\odot$. The notations $\mathbf{0}$ and $\mathbf{1}$ designate conformable column vectors of 0's and 1's, respectively.

(also known as *distributed lag*, or DL) models (Koyck, 1954; Griliches, 1967; Sims, 1971; Robinson, 1975; Neuvo et al., 1984; Saramaeki et al., 1993) and *vector autoregressive* (VAR) models (Sims, 1980), implicitly assume that the time series is sampled at a fixed frequency. This assumption is often ill-suited to language research because words in natural language have variable duration, whether spoken or read. The number of parameters in discrete-time deconvolutional models is also linear (or super-linear) on the length of the history window, which can easily lead to overparameterization. These objections equally apply to the common technique in psycholinguistics of injecting "spillover" regressors into linear models (i.e. adding coefficients for predictors associated with preceding events, e.g. Erlich and Rayner, 1983; Mitchell, 1984), which turns out to be FIR/DL by a different name (§2.2).

By contrast, CDR defines IRFs as parametric functions of continuous time and applies the same continuous IRF to all events in the history, yielding a model that can be applied to non-uniform time series (such as language) without distorting the temporal or featural structure of the stimulus sequence, with constant parametric complexity on the length of the history window. The continuous-time nature of CDR also allows the estimated response to be queried at any timepoint without reliance on post-hoc techniques for interpolation or extrapolation.

A visual comparison of CDR and FIR models is given in Figure 1.1. An ordinary least-squares model (Figure 1.1a) considers the response to be independent of all preceding stimulus events:[2]

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{b}, \sigma^2) \tag{1.2}$$

FIR models relax this independence assumption at the expense of model complexity by including additional weights for a fixed number of preceding events (Figure 1.1b).[3] Note

[2] $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the design matrix of predictors, $\mathbf{b}$ is the vector of coefficients, and σ^2 is the variance of the error distribution.

[3] FIR is a special case of the linear model given in eq. 1.2, with a structured design matrix $\mathbf{X}$ containing O values for each predictor representing a history of O timesteps into the past (see §2.2).

that FIR assumes that events are equidistant in time, or, equivalently, that variation in temporal spacing is inconsequential (an assumption implicitly made by spillover models in psycholinguistics). By contrast, CDR retains real-valued timestamps of the stimulus events and fits continuous IRFs that govern the influence of each predictor on the response as a function of time (Figures 1.1c and 1.1d). Unlike linear models, CDR is agnostic as to whether the stimuli and responses are measured at the same time (Figure 1.1c) or at different times (Figure 1.1d), a useful property for applications like fMRI modeling in which stimuli have variable duration but responses are measured at a fixed frequency (see §4.3 for discussion).

This thesis defines the CDR model and evaluates it empirically, assessing the influence of factors such as noise, multicollinearity, IRF misspecification, and hyperparameter selection on the final estimates, and additionally using CDR for IRF discovery from reading data and for hemodynamic response (HRF) discovery from fMRI data. The latter application is an appealing use case for continuous-time deconvolution because of the asynchrony between stimuli (words) and responses (brain scans) in naturalistic fMRI studies of language (§4.3).

The structure of the thesis is as follows. Chapter 2 (1) motivates continuous-time deconvolutional modeling in light of ongoing methodological challenges and theoretical questions in psycholinguistics and (2) relates CDR conceptually to existing deconvolutional methods. Chapter 3 defines the CDR model, exposits some of its mathematical properties, and describes a documented open-source Python implementation. Chapter 4 empirically evaluates CDR on synthetic data as well as human subjects data from self-paced reading, eye-tracking during reading, and fMRI studies of human language comprehension. Chapters 5–7 deploy CDR to address outstanding research questions in human sentence processing: whether word retrieval and prediction are supported by distinct cognitive mechanisms (Chapter 5), whether predictive processing in naturalistic language comprehension is structure-sensitive (Chapter 6), whether naturalistic language processing exhibits working memory effects (Chapter 7), and whether neural resources for predictive coding and working memory in the

language domain are shared with other domains of cognition (Chapters 6 and 7). Chapter 8 (1) defines the continuous-time deconvolutional regressive neural network (CDRNN), an extension of CDR that implements the IRF as a deep neural projection relating predictors and their timestamps to response estimates and their error distributions, thereby relaxing many of the remaining simplifying assumptions of CDR and thereby potentially allowing more flexible discovery of complex non-linear dynamics in human sentence processing, and (2) shows empirically that CDRNN also recovers ground truth models from synthetic data and compares favorably to CDR and other baselines on human-generated data, while also supporting detailed insights into underlying dynamics that cannot easily be obtained using alternative approaches. Chapter 9 concludes.

Chapter 2

Motivating the CDR Approach

2.1 Temporal Diffusion in Psycholinguistics: Methodological and Theoretical Considerations

The issue of temporal diffusion of effects has been recognized for decades by psycholinguists as being both methodologically and theoretically important (Morton, 1964; Bouma and De Voogd, 1974; Mitchell, 1984; Vasishth, 2006), although the kind and extent of this diffusion in human sentence comprehension is a matter of debate.

From a methodological standpoint, much behavioral methodology in psycholinguistics is based on the assumption that online behavioral measures like response times from self-paced reading or fixation durations from eye-tracking reveal local fluctuations of processing load in the language comprehension system that are closely time-locked to the linguistic signal. An early and influential theoretical characterization of this cognitive-behavioral link is the eye-mind hypothesis of Just and Carpenter (1980), which in its strong form posits a strict temporal alignment between attention and eye-movements (i.e. you look at what you are thinking about). This position predicts virtually no temporal diffusion of effects in reading measures, since information processing related to a particular word should fully terminate before attention is shifted to another word in the form of a *saccade* (a large-scale jump of the eyes to a new position on the screen). Although some early studies of eye movements indeed suggested little influence of preceding words on current fixation durations (Carpenter and Just, 1983), a great deal of evidence has subsequently accumulated both for the possibility of covert attention to unfixated objects (Posner, 1980; Posner et al.,

1987; Carrasco and McElree, 2001; Engbert and Kliegl, 2003, *inter alia*) and for temporal diffusion of various psycholinguistic effects, including word length (Kliegl et al., 2006; Pollatsek et al., 2008; Pynte et al., 2008), word frequency (Rayner and Duffy, 1986; Pollatsek et al., 2008; Staub and Clifton, 2006; Staub, 2011; Findelsberger et al., 2019), word predictability (Rayner et al., 2004b; Ashby et al., 2005; Smith and Levy, 2013), and memory retrieval cost (Warren and Gibson, 2002; Grodner and Gibson, 2005; Van Dyke, 2007). The proper interpretation of this evidence is a subject of ongoing theoretical debate (Reichle et al., 1998; Engbert et al., 2005; Reichle et al., 2009; Radach and Kennedy, 2013; Reichle and Drieghe, 2014, *inter alia*), especially in the reading literature, where some prominent computational models of reading (e.g. Reichle et al., 1998) require serial allocation of attention and thus disallow covert attention to preceding words. Although advocates of this position have argued that diffuse effects reported during eye-tracking are driven by measurement error (Reichle and Drieghe, 2014), this explanation cannot account for well known effects of temporal diffusion in other experimental paradigms, such as "spillover" effects in self-paced reading (e.g. Grodner and Gibson, 2005) and event-related potential (ERP) components from electroencephalography (EEG, Kutas and Hillyard, 1984), which occur over long enough time intervals (up to 900ms depending on the component of interest) that the responses to multiple words regularly overlap during naturalistic sentence comprehension (Smith and Kutas, 2015). For these reasons, the existence of temporally diffuse effects is largely taken for granted by psycholinguists, and controlling for it statistically is considered best practice (McDonough and Trofimovich, 2012).

The methodological importance of the temporal diffusion problem depends not only on latency in the underlying cognitive mechanisms but also on latency in the experimental measure. For example, response times in eye-tracking studies contain latencies due to planning and executing a saccade, over and above latency in information processing at the neural substrate (Travis, 1936; Hallett, 1978; Theeuwes et al., 1998; Yang et al., 2002; Altmann, 2010, *inter alia*). Hemodynamic measures like functional magnetic resonance

imaging (fMRI) and functional near-infrared spectroscopy (fNIRS) are a more extreme example of this problem, since the measurable response of blood oxygenation to neuronal activity peaks at a latency of around 6s and continues for over 30s (Boynton et al., 1996). Although such latencies are particular to the response measure and may have little direct bearing on cognitive theories, it is nonetheless important to develop methodology to control for them, especially in neuroimaging applications where measurement latency is known to have much more sluggish dynamics than the underlying neuronal timecourses of interest.

Temporal diffusion is relevant to a number of questions in psycholinguistic theory. As suggested by the foregoing discussion, one such question concerns the role of attention in language processing. Numerous studies have advocated the existence of a mental "buffer" that allows information processing to lag behind perception (Morton, 1964; Bouma and De Voogd, 1974; Mitchell, 1984; Sharkey and Sharkey, 1987; Tabor et al., 2004; Jacquemot and Scott, 2006; Mollica and Piantadosi, 2017), and numerous additional studies implicitly appeal to such a construct in explaining "spillover" effects as residual processing of previously encountered words (Grodner and Gibson, 2005; Vasishth and Lewis, 2006; Smith and Levy, 2013, *inter alia*). Under this view, words enter the language processing system and impose a processing burden that may not be fully discharged by the time subsequent words are encountered. A short-term buffer system that supports such lags by holding perceptual units in memory until they can be processed may be particularly adapted to real-time speech processing settings in which comprehenders cannot control the rate of input (Dahan, 2010; Mollica and Piantadosi, 2017). This proposal is ideally suited to continuous-time deconvolutional analysis such as CDR, which can directly estimate the rates at which different kinds of processing unfold.

Temporal diffusion is also of indirect theoretical importance to psycholinguistic theories that make temporally fine-grained predictions about processing effects. One high profile example concerns the debate between memory-based (Miller and Chomsky, 1963; Gibson, 2000; Lewis and Vasishth, 2005) and expectation-based (Hale, 2001; Levy, 2008) theories

of comprehension effort in human sentence processing. Memory-based theories predict processing cost to be proportional to the difficulty of retrieving referents from working memory in order to construct syntactic dependencies, with difficulty inversely related to the recency of the referent's latest mention. Thus, longer dependencies are expected to be more costly than shorter dependencies. Expectation-based theories predict processing cost to be inversely related to the contextual predictability of words. In the case of object-extracted relative clauses like the following from Levy (2008), both theories predict a local increase in processing difficulty due to the object extraction but disagree as to *where* in the sentence it will occur:

The reporter who **the** photographer **sent** to the editor hoped for a good story.

Memory-based accounts predict difficulty at *sent*, since this is where a long syntactic dependency (to *reporter*) must be resolved, while expectation-based accounts predict difficulty at *the*, since object extractions are less frequent than subject extractions in English, making the subject noun phrase *the photographer* unexpected. Although evidence from English indicates that the cost resides primarily at *sent* (Gordon et al., 2001; Grodner and Gibson, 2005; Staub, 2010), favoring memory-based models, the ensemble of cross-linguistic evidence on this question is more mixed, leading some to advocate the co-existence of both kinds of mechanisms (Levy et al., 2013). However, assumptions about timecourse (i.e. that costs are paid primarily on the costly word) are of crucial importance in interpreting these patterns. If processing costs actually diffuse across time, the predictions of these theories become harder to disentangle, since the processing cost at both locations is determined not only by the cost of processing the critical word but also by any residual cost of processing its sentential prefix. This is in fact one explanation considered by advocates of expectation-based models of language comprehension (Levy, 2008).

A related theoretical debate that hinges critically on temporal diffusion concerns the predicted influence of syntactic dependencies on sentence processing within memory-based

accounts. An influential memory-based theory of sentence comprehension effort is the dependency locality theory (DLT) of Gibson (2000). The DLT predicts processing difficulty proportional to the number of discourse referents that intervene in a syntactic dependency at the point at which it can be computed. In the example above, the DLT assigns an integration cost of 1 to the verb *sent* because one discourse referent (*photographer*) intervenes between *sent* and its extracted object (*reporter*). While this *locality effect* prediction has been borne out using constructed stimuli in English (Gibson and Ko, 1998; Grodner and Gibson, 2005), the opposite pattern (i.e. an *anti-locality effect*, or *decreased* processing cost for dependencies with many intervening discourse referents) has also been attested, both using naturalistic stimuli in English (van Schijndel and Schuler, 2013) as well as using constructed stimuli in head-final languages like German (Konieczny, 2000), Japanese (Nakatani and Gibson, 2010), and Hindi (Vasishth, 2003). These countervailing sources of evidence motivated Vasishth (2006) to re-evaluate the Grodner and Gibson (2005) data from English and show that locality effects are no longer significant under better control for temporal diffusion (i.e. considering "spillover" effects from preceding words).

In light of the above considerations, temporal diffusion may also play an underappreciated role even in studies of phenomena that do not directly concern effect timecourses. For example, while many studies in psycholinguistics (such as those cited above) include spillover effects in analyses, there are also many that do not, even in naturalistic settings where the rate of presentation is not controlled (e.g. Smith and Levy, 2008; Boston et al., 2008; Roark et al., 2009; Frank and Bod, 2011; Fossum and Levy, 2012; van Schijndel and Schuler, 2015) and diffusion might be especially pronounced (§2.1.3). In general, such studies are not directly concerned with effect timecourses, and omission of spillover regressors embodies an implicit belief that an *in situ* model with no controls for diffusion is "good enough" to permit discovery of the relevant patterns. However, controlling for temporal diffusion is important regardless of whether the central research question directly concerns timecourses because failing to do so can lead to false negatives (e.g. failing to detect an effect

because it occurred later than expected) or misattribution of variance due to temporally diffuse effects.

For example, using LME to evaluate the sensitivity of self-paced reading in the Natural Stories corpus (Futrell et al., 2018) to theory-driven predictors of memory retrieval cost shows significant main effects of syntactic constituent wrap-up (p = 2.33e-14) and syntactic dependency length (p = 4.87e-10; Shain et al., 2016). However, spilling over one control variable (probabilistic context-free grammar surprisal) one position (from *in situ* to spillover-1) causes the effects of interest to vanish ($p = 0.816$ for constituent wrap-up and $p = 0.370$ for dependency length).

The contrast between these two sets of results comes down to different assumptions about timecourse that are both reasonable *a priori* (i.e. whether the locus of surprisal effects should fall on the current word or spill over into the following one). For this particular dataset and model definition, it turns out that spilling over surprisal produces a stronger baseline that ultimately casts doubt on results obtained using a baseline inspired by preceding work. Such an outcome can be difficult to anticipate in advance. The possibility of such discrepant results based solely on assumptions about timecourse should motivate increased attention to diffusion of effects in psycholinguistic modeling.

The importance of controlling for temporal diffusion is of course dependent on experimental design. For example, there may be little impact from diffusion in a lexical decision task based on words presented in isolation with long intervals in between, while there is almost certainly a large influence of diffusion in fMRI scans of subjects listening to running speech. Psycholinguists and cognitive scientists are increasingly using naturalistic experiments in order to improve ecological validity and minimize task artifacts from artificially constructed designs (Demberg and Keller, 2008; Hasson and Honey, 2012; Campbell and Tyler, 2018). As suggested by the discussion of Shain et al. (2016) above, controlling for temporal diffusion may be of particular importance in such a setting, since measurements are taken from subjects carrying out rapid incremental sentence comprehension and multi-

ple word fixations may take place within a short span of time (Morton, 1964; Kolers, 1976). It is nonetheless possible that even experiments with carefully constructed stimuli might benefit from improved control of temporal diffusion. For example, even holding prefixes of linguistic stimuli fixed up to a critical region cannot entirely control the influence of temporal diffusion; the same prefix can be fixated differently in different presentations, both within and across subjects, potentially leading to variation in patterns of diffuse processing that may affect the response in the critical region.

In summary, the existing psycholinguistic literature indicates that temporal diffusion is an important factor in the study of human language processing, both because of methodological difficulties it presents for data analysis and interpretation, and because many psycholinguistic theories make different predictions about effect timecourses.

2.1.1 Existing Methods for Handling Temporal Diffusion in Psycholinguistics

Psycholinguists currently use one of two classes of methods for controlling and/or measuring temporal diffusion, one experimental and one statistical. The experimental method is to design the study in a way that minimizes the potential influence of temporal diffusion. For example, some cognitive processes can be studied by presenting participants with isolated words, as is done in the widely-used lexical decision task (McKoon and Ratcliff, 1979), in which participants judge whether a string presented on the screen constitutes an existing word of their language, and their reaction times are measured. In these kinds of studies, diffuse processing effects can reasonably be assumed to vanish during the inter-stimulus interval (ISI). Another option is to directly control the rate of stimulus presentation using rapid serial visual presentation (RSVP), in which sentences are presented to participants one word at a time at a fixed rate. RSVP is the dominant presentation method in EEG studies of language processing, since it permits clear separation of ERP components (Kutas and Hillyard, 1980, 1984). RSVP has also been used in many behavioral studies of language

processing (Lawrence, 1971; Potter, 1984; Kanwisher, 1987; McElree, 2000; McElree et al., 2003; Mollica and Piantadosi, 2017, *inter alia*).

Although such experimental manipulations reduce the influence of temporal diffusion, this reduction comes at the expense of both the ecological validity and the generality of the experimental paradigm. First, as discussed in more detail in §2.1.3, isolated word processing and RVSP are not the circumstances for which the human language comprehension system has been primarily adapted. Using these paradigms may distort the underlying cognitive processes of interest, and findings using them should ideally be replicated in more naturalistic settings (Hasson and Honey, 2012). Second, isolated word presentation or fixed-rate presentation (RSVP) are limited in the range of psycholinguistic hypotheses that they can be used to test. Isolated word presentation cannot be used to test theories about the influence of context on human sentence comprehension (such as the memory-based and expectation-based models cited above), and RSVP rules out the use of word-level reaction times (e.g. eye-tracking during reading or self-paced reading) as indicators of incremental comprehension difficulty, since it eliminates participants' control over word-by-word processing rates. This is a serious limitation because word-level reaction times are an inexpensive and widely-used source of evidence about incremental language processing.

Because of these issues, psycholinguistic study design often uses statistical rather than experimental control over temporal diffusion. Overwhelmingly, this is done using finite impulse response (FIR) modeling — typically referred to by psycholinguists as regression with "spillover" (Mitchell, 1984) — as described in §2.2.[1] FIR analyses can be estimated using ordinary least squares (OLS) (Grodner and Gibson, 2005) or linear mixed-effects (LME) (Demberg and Keller, 2008), and non-linear generalizations of FIR can be estimated

[1]Measurement specific impulse response functions have also been used, when available, most notably the use of preconvolution with the canonical hemodynamic response (Boynton et al., 1996) in fMRI studies of language processing (Willems et al., 2015; Brennan et al., 2016; Henderson et al., 2016, *inter alia*). A limitation of this approach is that the response function is based on group-level brain-wide averages and cannot adapt to individual participants or brain regions. It can therefore be a poor fit to the true underlying response in the experimental sample (Handwerker et al., 2004).

using generalized additive (GAM) models (Smith and Levy, 2013).[2]

An FIR regression of order o is a linear regression model that conditions the response at time t not only on properties of the event at time t but also on the properties of the $o-1$ preceding events at time $t-1,\ldots,t-o+1$ (see §2.2). When events are words, this allows the model to condition the response (e.g. eye-tracking fixation duration) not only on the properties of the word currently being fixated but also on the properties of the words fixated during the preceding $o-1$ fixations. If effects are temporally diffuse, then the coefficients of effects at preceding words should on average be non-zero, enabling statistical control over temporal diffusion of effects. The principal drawback of FIR for psycholinguistic data analysis is that psycholinguistic time series often consist of sequences of discrete events (e.g. word fixations) that have variable duration, and existing distributed lag models cannot directly handle such data. To apply FIR to such time series, psycholinguistic studies usually ignore the amount of clock time elapsed between events and consider only the relative order of events in the time series. If event durations are variable and the underlying response is a function of real time rather than relative event index, FIR modeling can be distortionary. For example, the corpus frequency of the preceding word may impact the current word's fixation duration differently depending on whether the preceding word was fixated 100ms ago vs. 1000ms ago. FIR/spillover cannot account for this possibility, while CDR can (see §2.2).

2.1.2 Types of Temporally Diffuse Effects in Language Processing

This section briefly reviews what is currently known about temporal diffusion of cognitive processes in psycholinguistic measures. For the purposes of this discussion, I set aside hemo-

[2] GAM models (Hastie and Tibshirani, 1986) relax the linearity requirement of linear models, allowing the predictors to be related to the response via arbitrary smooth functions. A GAM with a Gaussian linking function has the following form:

$$\mathbf{y} \sim \mathcal{N}(b_0 + f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \ldots + f_k(\mathbf{x}_k), \sigma^2) \tag{2.1}$$

dynamic measures like fMRI because latency in the measurement technique is so extreme that it swamps differences in latency between types of cognitive processes (Boynton et al., 1996). I focus instead on evidence from behavioral (e.g. eye-tracking) and electrophysiological (e.g. EEG) paradigms, which have high temporal resolution. Research in these domains has provided evidence of latency in (1) perceptual and motor processes, (2) lexical access, (3) working memory retrieval, and (4) predictive processing.

The linguistic signal is first conveyed to the mind by sensory systems (typically visual or auditory, although tactile systems are implicated in e.g. braille reading), and, in the case of reading, the rate of this conveyance is also mediated by comprehenders' motor processes. These sensory and motor processes can exhibit latencies that affect psycholinguistic response measures. EEG research sheds light on the timecourse of responses to visual and auditory stimulation. Although the bulk of these responses are quite rapid (within 100ms of stimulus onset, Woodman, 2010; Pratt, 2011), certain higher-level perceptual processes can last substantially longer, including phoneme recognition (Näätänen et al., 1997, 2007) and visual word recognition (Hauk et al., 2006). Psycholinguistic investigations of reading typically control for visual perception costs through orthographic variables like word length (in characters). Word length effects have been shown to be temporally diffuse (Kliegl et al., 2006; Pollatsek et al., 2008; Pynte et al., 2008), suggesting that visual processing latencies can register in the reading record. Response times in human reading are also affected by temporally diffuse processes in the motor system. For example, as discussed above, it takes time (around 130ms) to plan and execute a visual saccade (Travis, 1936; Hallett, 1978; Theeuwes et al., 1998; Yang et al., 2002; Altmann, 2010), although such latencies may interact with other processing latencies in non-linear ways (Altmann, 2010). Larger saccades are also known to increase fixation durations, not only on the current word but also on the subsequent one (Kliegl et al., 2006). These findings suggest that even low-level perceptual and motor processes may diffuse across time to such an extent that they can measurably affect processing of subsequent words.

Diffuse word frequency effects are one of the most widely attested forms of temporal diffusion in the psycholinguistic literature. Many studies have reported that the cost of retrieving rare words from the mental lexicon (as measured by corpus-based lexical frequencies) spills over onto the subsequent word (Inhoff and Rayner, 1986; Rayner and Duffy, 1986; Ashby et al., 2005; Kliegl et al., 2006; Staub and Clifton, 2006; Demberg and Keller, 2008; Pollatsek et al., 2008; Pynte et al., 2008; Staub, 2011; Findelsberger et al., 2019). Since word frequency is often used as a proxy for lexical retrieval difficulty under the hypothesis that frequency of access modulates the strength of a word's representation in memory (Staub, 2015), these findings are typically interpreted as driven by lexical access processes that do not fully finish by the time processing of the subsequent word begins (Ashby et al., 2005).

It is widely believed in psycholinguistics (1) that incremental sentence comprehension recruits working memory resources in order to store and update structured representations of the unfolding sentence and (2) that the computational demands placed on these memory systems vary as a function of the structural properties of the sentence (Miller and Chomsky, 1963; Gibson, 1998; McElree, 2000; Warren and Gibson, 2002; Lewis and Vasishth, 2005; Rasmussen and Schuler, 2018). Studies that explore such hypotheses regularly find that memory-related processing difficulty spills over onto the subsequent word. Erlich and Rayner (1983) and Jaffe et al. (2018) report that the cost of resolving an anaphor to low-activation referents is realized primarily on the word following the anaphor. Warren and Gibson (2002) and Grodner and Gibson (2005) report that costs related to constructing long dependencies during self-paced reading can be detected on the word *after* the dependency is complete. Van Dyke (2007) shows that a form of syntactic similarity-based interference, whereby the difficulty of constructing a syntactic dependency is increased by the presence of an intervener with syntactically similar features to the cue target, spilled over onto the following word. Pearlmutter et al. (1999) and Wagers et al. (2009) find that costs related to constructing a subject-verb agreement dependency in the presence of an attractor could

be observed following the verb. The ensemble of this evidence suggests that processing of subsequent words can begin before memory retrieval operations of various kinds have completed, leaving a portion of the retrieval cost to be paid on future words.

The contextual predictability of words is known to be an important determinant of processing difficulty, such that highly expected words contribute a small processing burden and highly unexpected words contribute a large one (Hale, 2001; Levy, 2008). Many studies have found that this word predictability effect can spill over into the processing of subsequent words (Balota et al., 1985; Rayner et al., 2004a; Ashby et al., 2005; Frisson et al., 2005; Smith and Levy, 2013; Delogu et al., 2017). One of the most extensive explorations of the temporal diffusion of word predictability is Smith and Levy (2013), discussed in §2.1.3, who find predictability effects up to three words after the unexpected word. These findings suggest that costs associated with unexpected material (or, conversely, facilitations associated with expected material) may be realized in psycholinguistic measures over a fairly long span of time.

Together, the results reviewed here indicate that the effects of many psycholinguistic phenomena diffuse across time and stress the importance of controlling for this possibility in the statistical model. While the resolution of the timecourse insights afforded by most of these studies is very coarse (typically, how much the next fixation is affected), the interpretations of these diffusion patterns regularly appeal to continuous-time notions: processing burdens associated with words or structures are thought to be discharged in the mind at a rate that may not fully keep pace with the rate at which words are encountered, leading to processing costs that bleed into future words (Morton, 1964; Mitchell, 1984; Ashby et al., 2005, *inter alia*). Given this, CDR stands to shed light on underlying sentence processing mechanisms, thanks to its ability to go beyond word-discretized measures of diffusion and directly estimate the shape of the response function, by exploiting naturally occurring variation in word duration.

2.1.3 Temporal Diffusion in Naturalistic Vs. Constructed Psycholinguistic Experiments

Observer's paradox is a serious obstacle in the study of human cognition: the mind is a complex system that adapts to the experimental environment in unpredictable ways (Miller and Cohen, 2001; Sreenivasan et al., 2014; D'Esposito and Postle, 2015; Campbell and Tyler, 2018). The dominant experimental paradigm in psycholinguistics is to construct stimuli that directly manipulate a linguistic variable of interest and assess the impact of this manipulation on a response measure (e.g. Ferreira and Clifton, 1986; MacDonald et al., 1994; Trueswell et al., 1994; Tanenhaus et al., 1995; Gibson, 1998; Kutas and Federmeier, 2000, among many others). This *constructed stimulus* paradigm is well suited to testing causal relationships between the manipulated variables and the measured response. However, human language is so complex that it is usually impossible to fully isolate the variable of interest from possible confounds using experimental design alone (Cutler, 1981), and therefore some amount of statistical control over confounding variables is usually necessary over and above even the most careful experimental design. More fundamentally, the more control is exercised over the experimental conditions, the more dissimilar the experiment becomes from the functional conditions for which human language capacity evolved, and the greater the chance that results may reflect the activity of cognitive mechanisms that usually play little or no role in language comprehension (Hasson and Honey, 2012; Richlan et al., 2013; Hasson et al., 2018; Campbell and Tyler, 2018). While this possibility is difficult to evaluate in the context of global behavioral measures, recent neuroscientific evidence shows that domain-general executive control regions activate during the processing of some artificially constructed language stimuli (Kaan and Swaab, 2002; Kuperberg et al., 2003; Novick et al., 2005; January et al., 2009) but fail to activate during the processing of naturalistic stimuli (Blank and Fedorenko, 2017). Such results have led some to argue that artificially constructed experimental stimuli may increase general cognitive load by coercing comprehension into problem solving, thereby engaging mechanisms that play little role

in everyday sentence processing (Campbell and Tyler, 2018; Diachek et al., 2020; Wehbe et al., 2020). While constructed experiments are an invaluable source of evidence about human language comprehension, patterns revealed by such experiments should ideally also be detectable in more naturalistic settings. It is therefore important to pursue questions about language processing using both approaches in parallel (Hasson et al., 2018), and a growing number of studies in psycholinguistics and cognitive neuroscience are using context-rich naturalistic language stimuli (e.g. stories and informative pieces) in their designs (Speer et al., 2007; Demberg and Keller, 2008; Yarkoni et al., 2008; Speer et al., 2009; Whitney et al., 2009; Frank and Bod, 2011; Fossum and Levy, 2012; Wehbe et al., 2014; Frank et al., 2015; Hale et al., 2015; van Schijndel and Schuler, 2015; Henderson et al., 2015; Brennan et al., 2016; Henderson et al., 2016; Huth et al., 2016; de Heer et al., 2017; Bhattasali et al., 2018; Brennan and Hale, 2019, *inter alia*).

To my knowledge, no previous study has directly investigated the degree of temporal diffusion in naturalistic psycholinguistic experiments relative to constructed psycholinguistic experiments. That said, there are several reasons to suppose that controlling for temporal diffusion may be of particular importance in a naturalistic paradigm. First, naturalistic experiments tend to be used to evaluate "broad-coverage" features that are expected to influence the response at every word (e.g. Demberg and Keller, 2008; Smith and Levy, 2013),[3] while this is not necessarily the case in constructed experiments. While some constructed experiments indeed examine the response at every word (e.g. Grodner and Gibson, 2005), others are interested in the response at a *critical region* where the manipulation is expected to produce an effect (e.g. Rayner et al., 1983). If the research design provides a critical region and the prefix to the critical region is carefully controlled, the effects of temporal diffusion are plausibly minimal, and differences at or shortly after the critical region can be attributed to the experimental manipulation (Rayner et al., 1989). Otherwise, any com-

[3]This generalization is not exceptionless; see e.g. Jaffe et al. (2018), who analyze a subset of words in a naturalistic study that were relevant to coreference resolution.

parison between responses at two different words of the stimulus should take into account potential differences in processing demands imposed by those words' contexts, especially if effects diffuse across time (Vasishth, 2006). Since word-by-word analyses of naturalistic sentence processing data seek to infer aggregate effect sizes from many words whose contexts all differ, they are plausibly more susceptible to confounds from temporal diffusion than controlled experiments with a critical region.

Second, the hypothesis that the processing of neighboring words will overlap during naturalistic sentence comprehension is suggested by prior knowledge about both (1) the neural response to language and (2) the rate at which words typically enter the language comprehension system. It is well established that scalp potentials can be deflected up to 900ms or more after word onset (Kaan, 2007), especially in the case of words that are particularly difficult to process (Kaan et al., 2000). Word durations in natural language are typically much shorter than this, with observational studies generally reporting average word durations on the order of 200-400ms, whether the presentation is visual (Rayner et al., 1989; Futrell et al., 2018) or auditory (Tauroza and Allison, 1990). These two findings suggest that (1) the neural response to a word can last more than three times longer than the word's presentation to the perceptual system and therefore that (2) the processing of previously-encountered words will continue even as future words are encountered (Smith and Kutas, 2015). In word-by-word analyses such as those that are typical of naturalistic studies, this diffusion may have a measurable impact on results.

Third, while many analyses of constructed experiments consider short spillover windows (typically 0 or 1 previous words, see e.g. studies reviewed in Rayner, 1998), at least one previous naturalistic reading study in psycholinguistics (Smith and Levy, 2013) has considered longer windows and found evidence of quite diffuse effects, with significant word predictability effects up to 3 words into the future in a self-paced reading experiment. Furthermore, the fitted impulse response to word predictability in Smith and Levy (2013) is not monotonic; there is relatively little effect on the costly word, a large effect on the sub-

sequent word, and an attenuated effect on the remaining two words in the discrete impulse response kernel. This indicates not only that the processing cost of word predictability can diffuse across time but that it can do so to such an extent that it is primarily realized on other words. Such a diffuse, late-peaking response pattern is not typically reported by constructed studies. This possibly indicates a more pronounced influence of temporal diffusion in the naturalistic paradigm, since participants are carrying out rapid incremental sentence comprehension where multiple words are processed within a short span of time (Morton, 1964; Kolers, 1976).

In summary, the naturalistic experimental paradigm may be more influenced by temporal diffusion than the constructed paradigm, both because lack of control over word contexts increases the need to control for a diffuse influence of these contexts, and because naturalistic studies tend to focus on word-by-word modeling of responses under conditions that plausibly do not fully separate the processing costs of individual words.

2.1.4 Example: Temporal Diffusion of Word Predictability Effects

One of the most robust findings in psycholinguistics is that human language processing mechanisms are sensitive to how predictable a word is given its context (Ehrlich and Rayner, 1981; Balota et al., 1985; Rayner et al., 2004a; Frisson et al., 2005; Frank and Bod, 2011; Rayner et al., 2011; Smith and Levy, 2013; Willems et al., 2015; Delogu et al., 2017, *inter alia*). For example, in isolation, the word *bottle* is more frequent and will be less costly to process than the word *kettle*, but when prefixed by *the pot calling the ...*, the word *kettle* is more expected and will be less costly to process than the word *bottle*. Predictability effects are also widely reported to spill over into subsequent words (Balota et al., 1985; Rayner et al., 2004a; Ashby et al., 2005; Frisson et al., 2005; Smith and Levy, 2013; Delogu et al., 2017). For instance, according to studies like the above, in the previous example, the word following *bottle* will also on average be processed more slowly than the word following *kettle*.

Although the existence of temporally diffuse predictability effects is fairly uncontroversial, different theories of the mechanisms that underlie them make different predictions about the shape of the response function (Reichle and Drieghe, 2014). For example, *E-Z Reader* (Reichle et al., 1998), an influential computational model of eye movement control during skilled reading, hypothesizes that attention is allocated serially word by word, and that spillover effects are due to parafoveal processing. Under this account, predictability facilitates lexical access of the fixated word, allowing the reader to begin to access the following word's properties even while it is in the parafoveal region and thus facilitating processing of the subsequent word when it is fixated. As a result, *E-Z Reader* predicts that any spillover effects — preditability-related or otherwise — will decrease rapidly and monotonically, and will have little impact two or more words later because of parafoveal limits (Reichle and Drieghe, 2014). Buffer-based accounts of temporal diffusion in language processing (Morton, 1964; Mitchell, 1984; Mollica and Piantadosi, 2017) do not necessarily make this prediction. Instead, the response function is thought to depend on both (1) the speed at which sentence processing mechanisms can carry out various tasks (e.g. wordform recognition vs. lexical access) and (2) the sequential ordering between tasks (e.g. the extent to which wordform recognition must precede lexical access, Hauk et al., 2006), and the response is generally assumed to be a function of continuous time (Mollica and Piantadosi, 2017) rather than driven by the spatial arrangement of words on a page. Under such a view, the response function is not necessarily expected to decay rapidly (if the relevant computations take a lot of time relative to the rate of stimulus presentation), nor is it necessarily expected to decay monotonically (if some kinds of processing, e.g. syntactic structure building, must wait to begin until other kinds have finished, e.g. lexical access). Since predictability effects are generally regarded as driven by late-stage processing (Reichle et al., 2003), a buffer-based account might predict more extensive diffusion of predictability effects. The shape of impulse response functions in human sentence processing therefore bears on debates about the structure of the sentence processing architecture, and there does not yet

appear to be a clear consensus on this question. In some eye-tracking studies, reported predictability effects are indeed larger at the critical word than at subsequent words (Rayner et al., 2004a, more consistent with the parafoveal account), while in others predictability effects at subsequent words are as large as or larger than those at the critical word (Smith and Levy, 2013, more consistent with the buffer account). The improved detail of impulse response estimation afforded by CDR could be used to shed light on this question.[4]

Beyond the allocation of attention underlying predictability effects, there is also debate about *why* more predictable words are easier to process and less predictable words are harder (see Kuperberg and Jaeger, 2016, for review). Of the many interesting questions in this domain, one that has recently risen to prominence concerns whether predictability effects are primarily driven by a *cost* or a *facilitation*. Influential information-theoretic models of the role of expectation in human sentence comprehension (Hale, 2001; Levy, 2008), henceforth *surprisal* theory, argue for the existence of a processing cost proportional to the Shannon information (Shannon, 1948) or *surprisal* of a word, which is equal to the negative log probability of a word given its context according to some sequential probability model. Subsequent work has argued that the assumption of highly incremental predictive processing leads asymptotically to this predicted logarithmic relationship between contextual probability and processing cost, and has provided experimental evidence that the relationship is indeed logarithmic (Smith and Levy, 2013). Thus, in this view, a logarithmic relationship between predictability and processing effort is not simply a convenient mathematical assumption, but a direct consequence of assuming highly incremental processing mechanisms that reallocate resources between competing interpretations of the unfolding sentence (Levy, 2008). This view also primarily frames predictability effects as a

[4]Note that this debate concerns the role of parafoveal processing in contributing to temporal diffusion of predictability effects and thus is specific to the study of eye movements during reading. Yet temporally diffuse effects are attested in many other experimental modalities for which such an explanation is not available, including self-paced reading (Mitchell, 1984), word-by-word lexical decision (Remington et al., 2018), and EEG (Smith and Kutas, 2015). Therefore, temporal diffusion of effects appears to be a more fundamental characteristic of human cognition than the eye movement debates might suggest.

cost: fully predictable words impose no cost (because they contribute zero bits of Shannon information), and costs scale up (in principle, to infinity) in inverse proportion to the logarithm of the word's contextual probability. An alternative (and older) view of predictability effects, henceforth *pre-activation* theory, is that they are driven by pre-activation by context of upcoming words and structures, making processing of those words and structures easier when they are eventually encountered (Kuperberg and Jaeger, 2016; Brothers and Kuperberg, 2021). Under this view, processing a particular word in a particular context imposes a cost on the language processing system (e.g. retrieving the word and its properties from the mental lexicon, constructing syntactic dependencies, updating semantic representations, etc.) that can be partially paid in advance if the context renders the word predictable, but must be paid in full once the word is encountered if the word could not be predicted. This view primarily frames predictability effects as a *facilitation*: successful prediction eases future processing, but processing can nonetheless be carried out even if the system entirely failed to predict the word (in contrast to surprisal theory, which in principal assigns infinite cost to words with a human subjective probability of 0 — whether this edge case is ever relevant is part of the debate). Since differences in degree of low contextual probability are not predicted to have much impact on processing cost (since the word will mostly fail to be pre-activated), this view also predicts a *linear* rather than logarithmic relationship between predictability and frequency, as supported by recent evidence (Brothers and Kuperberg, 2021) in contrast to Smith and Levy (2013).

As shown in Smith and Levy (2013), predictability effects can be temporally diffuse, and the extent of this diffusion can depend on experimental modality (in that study, self-paced reading vs. eye-tracking during reading). CDR can be used to localize predictability effects in time more precisely than alternative methods, thereby improving inferences that can be made about their influence on the response under e.g. an assumed linear vs. logarithmic effect shape. In addition, as shown in Chapter 8, a deep neural relaxation of CDR can jointly estimate both the timecourse and functional form of effects, potentially providing

more direct insight into this question.

Furthermore, although to my knowledge differences in predictions about temporal diffusion between these two theories have not been previously explored, I would argue that surprisal theory and pre-activation theory not only make different predictions about the functional form of the relationship between predictability and processing cost, but also about how predictability effects will diffuse across time. In particular, under pre-activation theory, predictability effects arise epiphenomenally from anticipatory partial execution of the lexical and structural operations required to process a word in context. Predictability or surprisal should therefore not carry a unique impulse response, but should simply modulate the timing and extent of anticipatory processing. By contrast, under surprisal theory, the primary work of sentence comprehension is a reallocation of resources between competing interpretations of the sentence at each new word, with cost proportional to the Kullback-Liebler divergence (Kullback and Leibler, 1951) of the conditional distribution over strings given context before and after observing the word, plus possibly other costs related to memory and integration (Levy et al., 2013); that is, the distributional update itself contributes a unique cost (Aurnhammer and Frank, 2019). Although it is not pursued further in this thesis, a possible future CDR-based test of these two theories could be to compare models in which word predictability simply weights the impulse responses to other measures of processing demand (e.g. measures of integration cost), as predicted by pre-activation theory, against models in which word predictability has its own impulse response, as predicted by surprisal theory.

In addition to these theoretical considerations, it is important to consider the possibility that temporal diffusion of predictability effects may influence the results of psycholinguistic data analysis in ways that may be unforeseen by existing theory. This is the case even if predictability is not a quantity of direct interest, as evidenced by the sensitivity of previously-reported memory retrieval effects to the assumed timecourse of predictability-related covariates (discussed above with respect to Shain et al., 2016). CDR improves the

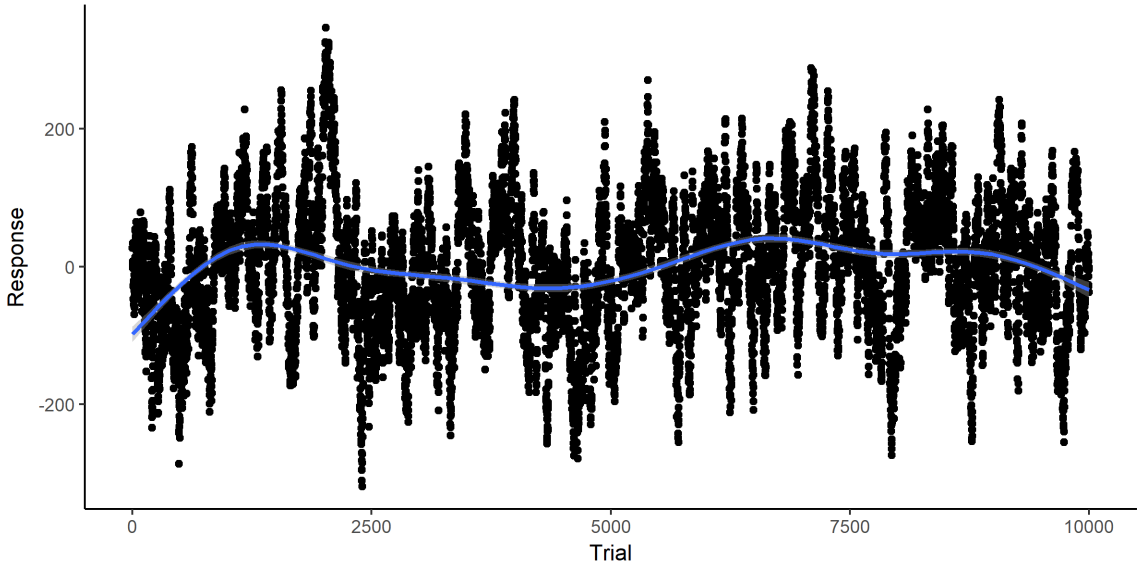


Figure 2.1: Synthetic responses produced by convolving *i.i.d.* normal independent variables with stationary convolution kernels. Note the undulating GAM smooth in blue, suggesting non-stationarity that is not in fact present in the underlying generative process.

flexibility with which the statistical model can discover and control for temporally diffuse processes, and should therefore both (1) shed light on theoretical questions that make direct predictions about effect timecourses and (2) improve the robustness of analyses of other kinds of psycholinguistic phenomena, which may be affected by temporal diffusion of word predictability in unanticipated ways.

2.1.5 Other Temporal Phenomena

The present concern about temporal diffusion in psycholinguistic data complements recent psycholinguistic interest in other kinds of temporal confounds, especially auto-correlation and non-stationarity (Baayen et al., 2017, 2018).[5] Baayen et al. (2018) demonstrate the utility of including a first-order auto-regressive term in a GAM model to control for auto-correlated error, while Baayen et al. (2017) relax the assumption of stationarity by augmenting GAM models with an independent variable $C \in [0, 1]$ representing the proportion

[5]For related findings at high temporal resolution, see Cho et al. (2018).

of the series completed at the current timestep. When a spline is fitted directly to C, the obtained curve can be interpreted as an estimate of fluctuation in the base response rate over time. And when a spline is fitted to the interaction between C and a predictor, the obtained curve can be interpreted as an estimate of fluctuation in the influence of the predictor over time. Baayen et al. (2017) show that directly modeling non-stationarity can have important impacts on both effect estimates and significance testing when applied to time series generated by human subjects.

While auto-regressive/non-stationary GAM models can capture temporal effects, the crucial point of divergence from this work is that they do not capture temporal diffusion. They allow the influence of an independent variable to fluctuate with time, but continue to assume independence of the response from preceding observations of the predictor(s). In order to handle temporal diffusion, auto-regressive/non-stationary GAM models must still make use of the problematic spillover technique discussed above.

CDR and non-stationary GAM models can therefore be seen to address distinct potential confounds in time series data: temporal diffusion of effects (CDR) vs. auto-correlation and non-stationarity (GAM). All three confounds can be addressed relatively straightforwardly by deploying CDR as a pre-process to GAM fitting, resulting in a two step analysis in which the data are first convolved with CDR and then analyzed using GAM (see §3.6 for further elaboration on this general idea). Alternatively, a deep neural generalization of CDR (Chapter 8) can jointly accommodate continuous-time temporal diffusion and non-linear, interactive effects. However, note that convolutional structure can in some cases explain apparent autocorrelation and non-stationarity. For example, the plot in Figure 2.1 shows a time series of synthetic responses generated by convolving *i.i.d.* normal independent variables with gamma-shaped convolution kernels, as described in §4.1. The overall base response rate in Figure 2.1 appears to fluctuate with time. This is supported by the undulating GAM spline (shown in blue), suggesting non-stationarity. Responses are also clearly auto-correlated, as shown by the higher frequency oscillations evident in the plot. However,

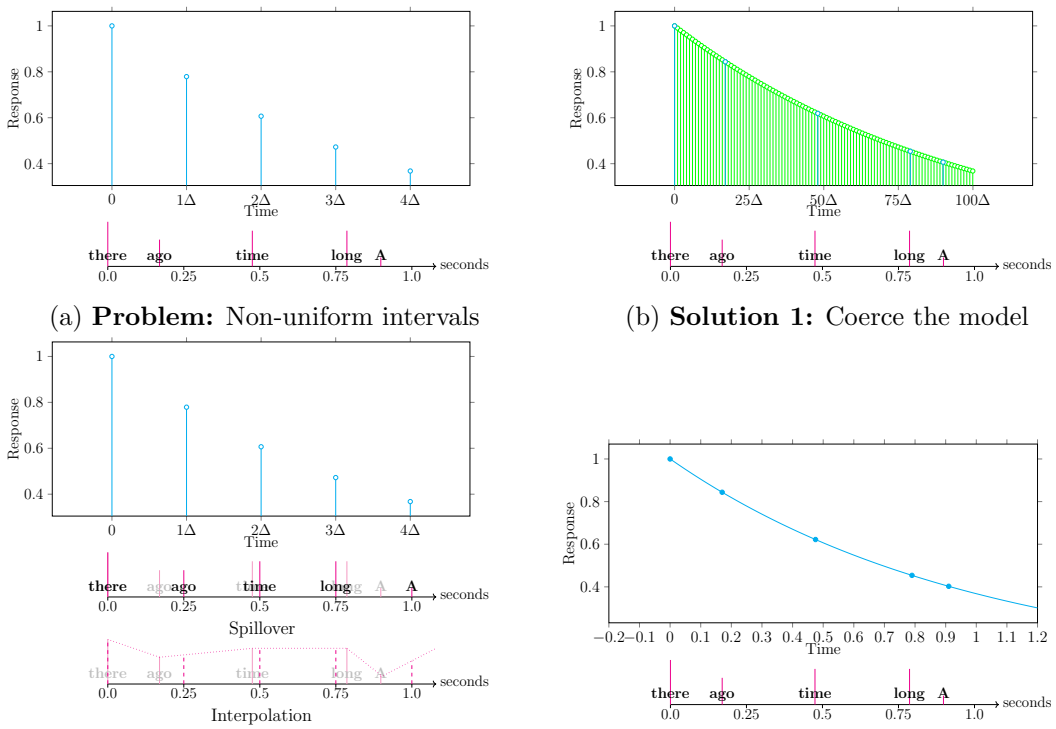
the data were in fact generated by a strictly stationary convolutional process and are *i.i.d.* normal conditional on the convolution. Apparent autocorrelation and non-stationarity are artifactual. Thus, one possible source of apparent auto-correlation and non-stationarity in time series data may be latent convolutional structure, and, in these cases, diffusion is the core temporal confound that must be brought under statistical control.

2.2 Existing Deconvolutional Models

In order to infer the structure of an IRF g in eq. 1.1 from data, it is first necessary to construct a solution space over which to perform inference. One way of doing so is through **discrete-time deconvolution**, a class of methods which recast deconvolution as a special case of linear regression by discretizing time into a finite number of equidistant steps and then estimating timestep-specific parameters. Continuous-time IRFs can be inferred from these discrete estimates if desired using various post-hoc smoothing techniques. One example of this general approach, known as *finite impulse response* (FIR) models in the signal processing literature (Neuvo et al., 1984; Saramaeki et al., 1993) and *distributed lag* (DL) models in the time series literature (Koyck, 1954; Griliches, 1967), consists of including regressors from previous timesteps. For simplicity, I will henceforth refer to such models as FIR. A fixed-effects FIR model of order O with K predictors is a linear model of $\mathbf{y} \in \mathbb{R}^N$ with design matrix $\mathbf{X} \in \mathbb{R}^{N \times K}$ and parameters $\mathbf{b} \in \mathbb{R}^{1+O \cdot K}$, i.e. an intercept, plus one coefficient for each of K predictors for each of O timesteps into the past:

$$\mathbf{y}_{\text{FIR}_O} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{1}^N & \mathbf{X} & \mathbf{0}^{1\times K} & \cdots & \mathbf{0}^{(O-1)\times K} \\ & & \mathbf{X}_{[1\cdots N-1,*]} & \cdots & \mathbf{X}_{[1\cdots N-O+1,*]} \end{bmatrix} \mathbf{b}, \sigma^2\right) \tag{2.2}$$

The sequence of O coefficients for a given predictor defines a discrete-time IRF, and the linear combination of predictors with $\mathbf{b}$ defines a temporal convolution operation (i.e. a weighted sum along the time dimension).



(a) **Problem:** Non-uniform intervals

(b) **Solution 1:** Coerce the model

(c) **Solution 2:** Coerce the data

(d) **Solution 3:** CDR

Figure 2.2: The problem of variably spaced events for discrete-time convolution in a hypothetical univariate FIR model on predictor *word length* (characters) with step duration Δ. Plots show hypothetical IRFs, with predictor values for each word shown in magenta in timelines below each plot. The example sentence is typeset in reverse because the impulse response describes the changing influence of words as they recede into the past. An order 5 FIR model cannot be used directly on a variably-spaced word sequence because there is no value of Δ that aligns the FIR coefficients with the stimulus events (2.2a). One solution (2.2b) is to use a high-resolution IRF of order sp, where s is the length of the history window in seconds and p is the inverse of the precision of the temporal measurement. Although this model can be directly applied to variably-spaced events, it overparameterized to the point of being unidentifiable. In this hypothetical training example where $s = 1$ and $p = 100$, only 5/100 parameters have data. Another solution is to coerce the data into a format that fits the assumptions of FIR (2.2c). Temporal variation can be deleted by "snapping" words to coefficients in one-to-one alignment under the assumption of a fixed but unknown value for Δ (Spillover). This technique is distortionary if the stimuli are variably spaced and their underlying contribution is a function of clock time rather than relative event index. Alternatively, the predictor can be continuously interpolated between events, and the interpolated signal is resampled at points (vertical dashed lines) that align with the discrete IRF coefficients (Interpolation). This technique is distortionary for event-based predictors that are not underlyingly continuous. CDR (2.2d) avoids both sparsity and distortion by replacing the discrete IRF with a parametric continuous function of time (in this example, $f(x;\beta) = e^{-\beta x}$). A continuous IRF can be queried exactly at any point, has a parametric complexity that is independent of the temporal span or resolution of the response kernel, can be applied directly and without distortion to variably-spaced time series, and is agnostic to temporal alignment between stimuli and response.

Another prominent example of discrete time deconvolutional approaches is vector autoregressive (VAR) modeling (Sims, 1980). VAR generalizes FIR to predict the next timepoint (row) of $\mathbf{X}$ rather than a distinguished response $\mathbf{y}$. VAR thus estimates parameters $\mathbf{B} \in \mathbb{R}^{(1+O \cdot K) \times K}$, and generates predictions $\mathbf{Y}^{N \times K}$ through linear transformation:

$$\mathbf{Y}_{\text{VAR}_O} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{1}^N & \mathbf{0}^{1\times K} & \cdots & \mathbf{0}^{O\times K} \\ & \mathbf{X}_{[1\cdots N-1,*]} & \cdots & \mathbf{X}_{[1\cdots N-O,*]} \end{bmatrix} \mathbf{B}, \sigma^2\right) \tag{2.3}$$

Thus, unlike FIR, which predicts a distinguished response variable via convolution of the history of predictor values, VAR predicts all K variables at the current timestep through summed linear transformations of the O preceding timesteps. VAR fits can be used to extract IRFs between any pair of variables in the model. Both of these techniques are widely used in the fMRI literature (Friston et al., 1994; Harrison et al., 2003), and both can be augmented with random effects (Beckmann et al., 2003; Gorrostieta et al., 2012). Since VAR estimates IRFs between all pairs of variables in the data, it is perhaps unnecessarily powerful for typical psycholinguistic studies that seek to model a distinguished response variable. The remainder of this discussion of discrete-time deconvolution therefore focuses on FIR.

As discussed in §1, in psycholinguistics, temporal diffusion is often addressed by adding "spillover" predictors encoding predictor values from preceding events. From eq. 2.2 above, it follows that this approach reduces to FIR modeling: a linear model containing O spillover positions of K predictors is an FIR model of order O on those K predictors, and the set of O coefficients for each predictor defines a discrete-time IRF. In practice, the term *FIR* tends to be used in signal processing settings where the temporal distance between samples is *fixed*, while the term *spillover* tend to be used in experimental settings where the temporal distance between samples is *ignored*. While this distinction matters for the interpretation of the discovered IRFs, since discrete-time IRFs only have a clock-time interpretation when the offset between timesteps is fixed and known, it is immaterial for the definition of the

statistical model itself. Even the GAM models with spillover used in e.g. Smith and Levy (2013) can be thought of as a variant of FIR with an IRF kernel whose shape depends on the values of the impulses at each timestep. Because of the identity (in the case of LME) or close relationship (in the case of GAM) of spillover models to FIR models, the remainder of this article will no longer distinguish the two, since the same objections apply regardless of terminological choice.

Since additional spillover positions contribute additional parameters, rich spillover controls can easily create such heavily parameterized models that realistically sized datasets cannot support them, especially when used in the context of mixed-effects or spline regressions that fit random effects or multi-dimensional smooths for each spillover position of each predictor. For example, linear mixed effects models with by-subject random slopes fitted using `lme4` (Bates et al., 2015) on the Dundee corpus (Kennedy et al., 2003) fail to converge with two or more spillover regressors.[6] These models only contained spillover variants of four predictors, and the problem of overparameterization can be even more severe in models that contain more control variables (e.g. Demberg and Keller, 2008, where LME models contained up to thirteen predictors). In addition to concerns about overparameterization, spillover regressors can also introduce spurious multicollinearity to the extent that predictors are autocorrelated, which can be problematic for model identification and interpretation (Kutner et al., 2003).

As a consequence, analysts are forced in practice to trade off the richness of the time-course model with other sources of complexity. For example, Smith and Levy (2013) use GAM models containing rich spillover structures (up to three timesteps) but relatively poor random effects (by-subject random intercept), while van Schijndel and Schuler (2015) use

[6]Models used log go-past durations as the dependent variable and included the predictors *word length* (in characters), *saccade length* (in words), *unigram log probability*, and *5-gram surprisal*, with probabilities computed by KenLM language models (Heafield et al., 2013) trained on the Gigaword 3 corpus (Graff and Cieri, 2003). Spillover positions from 0 (*in situ*) to n for each predictor were included for six models, one for each of $n \in \{0, 1, 2, 3, 4, 5\}$, along with a random intercept by word and random slopes by subject for each predictor (and each spillover position of each predictor). Outlier filtering was performed following van Schijndel and Schuler (2015), yielding a total of 193,309 data points.

LME models with rich random effects (by-subject and by-word random intercepts along with by-subject random slopes for every predictor) but no spillover regressors.

However, perhaps more fundamental for language research than the aforementioned computational problems is the inability of discrete-time deconvolutional models to represent variably spaced events. Figure 2.2a visually exemplifies this problem, and Figures 2.2b and 2.2c exemplify possible solutions to it within a discrete-time framework, each of which has undesirable properties. As shown in Figure 2.2a, an FIR model assumes a single fixed interval Δ between coefficients, and thus the discrete-time IRF cannot directly convolve the properties of variably-spaced words because no such interval exists. This problem is visualized by the lack of temporal alignment between the words in the example and the FIR coefficients, and it can be addressed by coercing the model to match the data or the data to match the model. To coerce the model to match the data, the interval Δ can be reduced to the level of precision of the temporal measurement (e.g. 1 millisecond) along with a compensatory increase in the number of FIR coefficients per unit time (Figure 2.2b), ensuring alignment to an FIR coefficient of all past events within some finite window. As visualized in the figure, such an approach exaggerates the problem of overparameterization and data sparsity to such a degree that I am unaware of any psycholinguistic studies that attempt to use this technique (how many events in a psycholinguistic experiment are spaced exactly 142ms apart?). To coerce the data to match the model, the stimuli can be (1) forced into one-to-one alignment with the FIR coefficients under the simplifying assumption that Δ is fixed but unknown or (2) interpolated and resampled at points that align with the FIR coefficients (Figure 2.2c). The forced alignment approach is equivalent to spillover, and it is distortionary for variably spaced events to the extent that the underlying contribution of those events is a function of clock time rather than relative event index. The interpolation approach has been used e.g. in fMRI modeling (Huth et al., 2016), and it is distortionary to the extent that the stimuli represent transient events rather than samples

from a continuously evolving feature space.[7]

CDR avoids both problems (overparameterization and distortion) by defining the IRF as a continuous-time parametric kernel (Figure 2.2d). Because a continuous IRF has infinite precision, it can be queried exactly at any point, thereby avoiding the trade-off faced by discrete-time models between parsimony on the one hand and the temporal span and resolution of the response kernel on the other. An additional advantage of continuous-time deconvolution is the ability to model asynchronously measured data (see Chapter 1). Because continuous-time deconvolution is parsimonious, faithful to the underlying temporal structure in the stimulus, and agnostic to temporal alignment between stimulus and response, it is more appropriate than FIR (spillover) approaches for analyzing many kinds of psycholinguistic data. Despite these conceptual advantages, continuous-time deconvolution is not currently used in psycholinguistics[8] and is little used in cognitive science more generally (aside from some previous neuroimaging studies that optimize the parameters of gamma-shaped hemodynamic response functions, e.g. Kruggel and von Cramon, 1999; Kruggel et al., 2000; Miezin et al., 2000; Lindquist and Wager, 2007; Lindquist et al., 2009).

In addition to the discrete-time frameworks discussed above, related continuous-time regression models have also been proposed. Prior work has defined continuous-time extensions of distributed lag models (Sims, 1971; Robinson, 1975, 1976; Bergstrom, 1984). However, these approaches rely on Fourier analysis of the discretized covariate vector in order to model the continuous IRF. Consequently, they impose two problematic restrictions: (1) the covariates must be underlyingly continuous, and (2) discrete samples from the covariates must be taken at a uniform time interval (Robinson, 1975). They are therefore even less applicable to non-uniform discrete time series than their discrete-time analogs, which do not impose continuity constraints, while also being subject to the same critiques of the

[7] As discussed in §3.1, interpolation of variably-spaced samples from predictors that *are* underlyingly continuous over time (e.g. ambient noise level) is appropriate and in fact necessary to avoid distortion in CDR's event-based convolution procedure.

[8] Aside from the studies described in this thesis that apply CDR to psycholinguistic data.

uniformity requirement.

Mathematically, the most closely related existing model to CDR is the the Hawkes process model, also known as a self-exciting counting process model (Hawkes, 1971). Hawkes process models are used to analyze stochastic point processes in which the occurrence of an event locally increases the instantaneous probability of other events occurring, and thus the intensity function of the process is *self-exciting* in continuous time. Formally, a Hawkes process generates the intensity $\lambda(t)$ given (possibly non-stationary) base intensity $\mu(t)$ and event times T via convolution with the triggering function $g(t)$ (analagous to an IRF):

$$\lambda(t) = \mu(t) + \sum_{\tau \in T, \tau < t} g(t - \tau) \tag{2.4}$$

Commonly, g is chosen to be the two parameter exponential function $g(t) = \alpha e^{\beta t}$, $\alpha, \beta > 0$, enforcing exponential decay of the self-excitation function with amplitude α and decay rate β,[9] and the intensity $\lambda(t)$ is taken to be the parameter of a Poisson distribution describing the instantaneous concentration of events given the history, or equivalently, the rate parameter of an exponential distribution describing the expected waiting time until the next event given the history (Cooper, 2005). Parameters μ, α, and β are usually estimated from data using non-linear numerical optimization (Ozaki, 1979). This framework has been generalized in many ways, including extension to multivariate event data (i.e. simultaneous modeling of multiple event streams; Embrechts et al., 2011), extension to *marked* processes that contain regressors in addition to timestamps (Lapham, 2014), and the use of recurrent neural network intensity functions (Mei and Eisner, 2017).

Although both CDR and Hawkes processes involve a continuous parametric convolution over the time dimension, a fundamental difference between them is that CDR seeks to model a designated response variable while Hawkes processes seek to model the future temporal

[9] Other kernel types, including power law kernels $h(t) = \frac{\alpha}{(t+\beta)^{\eta+1}}$ (Lapham, 2014), non-parametric basis kernels (Zhou et al., 2013), and self-regulating neural network kernels (Mei and Eisner, 2017), are also widely used.

realization of the sequence of events. To my knowledge, no existing formulation of Hawkes process models can be used to address the temporal diffusion problem targeted in this study.

Part II

Continuous-Time Deconvolutional Regression

Chapter 3

CDR Model Definition and Implementation

The mathematical definition of CDR is given in §3.1, with a worked step-through of the equations in §3.1.1. §3.2, §3.3, §3.4, and §3.5 exposit additional mathematical properties of the model.

3.1 CDR Model

The CDR model assumes the following quantities as input:

- $X \in \mathbb{N}$: Number of predictor observations
- $Y \in \mathbb{N}$: Number of response observations
- $K \in \mathbb{N}$: Number of predictors
- $R \in \mathbb{N}$: Number of impulse response parameters
- $J \in \mathbb{N}$: Number of unique time series[1]
- $\mathbf{X} \in \mathbb{R}^{X \times K}$: Design matrix of X predictor observations of K dimensions each
- $Z \in \mathbb{N}$: Number of random grouping factor levels[2]
- $\mathbf{Z} \in \{0, 1\}^{Y \times Z}$: Boolean matrix indicating random grouping factor levels associated with each response observation
- $\mathbf{y} \in \mathbb{R}^{Y}$: Vector of Y response observations

[1] $J \ll X, Y$ because each time series indexed by $\{1, \dots, J\}$ contains many predictor and response observations.

[2] The sum total of all levels of each random grouping factor in the model, e.g. the number of subjects plus the number of items.

	Name	Type	Description
Dimensions	X	$\mathbb{N}$	Number of predictor observations
	Y	$\mathbb{N}$	Number of response observations
	Z	$\mathbb{N}$	Number of random grouping factor levels
	K	$\mathbb{N}$	Number of predictors
	R	$\mathbb{N}$	Number of impulse response parameters
	J	$\mathbb{N}$	Number of unique time series
Data	$\mathbf{X}$	$\mathbb{R}^{X \times K}$	X predictor observations
	$\mathbf{y}$	$\mathbb{R}^{Y}$	Y response observations
	$\mathbf{Z}$	$\{0,1\}^{Y \times Z}$	Random effects indicator
	$\mathbf{t}$	$\mathbb{R}^{X}$	Timestamps of observations in $\mathbf{X}$
	$\mathbf{t}'$	$\mathbb{R}^{Y}$	Timestamps of observations in $\mathbf{y}$
	$\mathbf{c}$	$\{1,2,\ldots,J\}^{X}$	Time series IDs of observations in $\mathbf{X}$
	$\mathbf{d}$	$\{1,2,\ldots,J\}^{Y}$	Time series IDs of observations in $\mathbf{y}$
Parameters	μ	$\mathbb{R}$	Fixed intercept
	$\mathbf{m}$	$\mathbb{R}^{Z}$	Random intercepts
	$\mathbf{u}$	$\mathbb{R}^{K}$	Fixed coefficients
	$\mathbf{U}$	$\mathbb{R}^{Z \times K}$	Random coefficients
	$\mathbf{v}_k$	$\mathbb{R}^{R}$	Fixed IRF parameters for the k^{th} predictor
	$\mathbf{V}_k$	$\mathbb{R}^{Z \times R}$	Random IRF parameters for the k^{th} predictor
	σ^2	$\mathbb{R}_+$	Variance
Model	$g_k(t;\theta)$	$\mathbb{R}_+ \to \mathbb{R}$	IRF kernel for the k^{th} predictor, function of time t given parameters θ
	$\mathbf{m}'$	$\mathbb{R}^{Y}$	Fixed + random intercepts
	$\mathbf{U}'$	$\mathbb{R}^{Y \times K}$	Fixed + random coefficients
	$\mathbf{V}'_k$	$\mathbb{R}^{Y \times R}$	Fixed + random IRF parameters for predictor k
	$\mathbf{F}$	$\{0,1\}^{Y \times X}$	Convolution mask
	$\mathbf{G}_k$	$\mathbb{R}^{Y \times X}$	Convolution matrix for the k^{th} predictor
	$\mathbf{X}'$	$\mathbb{R}^{Y \times X}$	Convolved design matrix

Table 3.1: Summary of variables in CDR model definition

- $\mathbf{t} \in \mathbb{R}^X$: Vector of timestamps associated with each observation in $\mathbf{X}$
- $\mathbf{t}' \in \mathbb{R}^Y$: Vectors of timestamps associated with each observation in $\mathbf{y}$
- $\mathbf{c} \in \{1, 2, \dots, J\}^X$: Vector of time series IDs associated with each observation in $\mathbf{X}$
- $\mathbf{d} \in \{1, 2, \dots, J\}^Y$: Vectors of time series IDs associated with each observation in $\mathbf{y}$
- $g_k(t; \theta) \in \mathbb{R}_+ \to \mathbb{R}$ for $k \in \{1, 2, \dots, K\}$: Parametric IRF kernels specifying response at time t given parameters θ, one for each of K predictors

Following e.g. `lme4` (Bates et al., 2015), I use *random grouping factor* to refer to variables that capture categorical random variation in a model (e.g. *participant* or *item*) and *random grouping factor level* to refer to individual values of a random grouping factor (e.g. the value `participant A` of the random grouping factor *participant*). Furthermore, by *time series* I denote a single unique sequence of observations of predictors and responses. A single dataset may contain multiple time series. For example, a psycholinguistic experiment may contain data from several participants, each of whom read several texts. Each participant-text pair could be treated as a unique time series. Different time series are considered statistically independent and are indexed by unique time series IDs (represented in $\mathbf{c}$ and $\mathbf{d}$). Note that X (number of predictor observations) and Y (number of response observations) can differ in a CDR model because $\mathbf{X}$ will be forced into a conformable dimensionality with $\mathbf{y}$ via convolution over time (see $\mathbf{X}'$ below). This property permits CDR analysis of predictor/response streams with different acquisition times.

CDR seeks to estimate the following quantities, which mediate between $\mathbf{X}$ and $\mathbf{y}$:

- a scalar intercept $\mu \in \mathbb{R}$
- a vector $\mathbf{m} \in \mathbb{R}^Z$ of Z random intercepts
- a vector $\mathbf{u} \in \mathbb{R}^K$ of K fixed coefficients[3]

[3]Throughout this thesis I use the term *coefficients* to refer to what are often called *slopes* in linear models. This is to avoid falsely implying that the coefficients represent straight-line functions of the predictors, when in fact they are applied non-linearly to the predictors via the impulse response. Alternatively, the coefficients can be construed as slopes on the *convolved* predictors $\mathbf{X}'$, as shown in eq. 3.7.

- a matrix $\mathbf{U} \in \mathbb{R}^{Z \times K}$ of ZK random coefficients, i.e. random estimates for each of K predictors for each of Z random effects levels
- K vectors $\mathbf{v}_k \in \mathbb{R}^R$ of R fixed IRF kernel parameters for K fixed predictors
- K matrices $\mathbf{V}_k \in \mathbb{R}^{Z \times R}$ of ZR random IRF kernel parameters, i.e. random estimates for each of K predictors for each of R IRF parameters for each of Z random effects levels
- a scalar variance $\sigma^2 \in \mathbb{R}_+$ of the response

Random parameters $\mathbf{m}$, $\mathbf{U}$, and $\mathbf{V}_k$ are constrained to be zero-centered within each random grouping factor.[4]

A fixed-effects CDR model therefore contains $2 + K + KR$ parameters: one intercept, K coefficients (one for each predictor), KR IRF parameters (R parameters for each predictor), and one variance of the response. Mixed-effects CDR models can also include random variation in the intercept, coefficients, and/or IRF parameters. This yields at most $1 + (Z + 1)(1 + K + KR)$ estimates for a mixed effects model with Z total random grouping factor levels (for example, the number of subjects plus items). Sub-maximal numbers of estimates can arise by forcing random effects components to zero. For example, zeroing out $\mathbf{v}_k$ and $\mathbf{V}_k$ eliminates random coefficients and IRF parameters for the k^{th} predictor.

To support mixed modeling, the fixed and random effects must first be combined by adding fixed effects with their random offsets using the indicator matrix $\mathbf{Z}$, resulting in intercept vector $\mathbf{m'} \in \mathbb{R}^Y$, coefficient matrix $\mathbf{U'} \in \mathbb{R}^{Y \times K}$ and IRF parameter matrices

[4] In practice, I also assume normally distributed random effects, either implicitly (via L_2 regularization) or explicitly (via variational priors and posteriors).

$\mathbf{V}'_k \in \mathbb{R}^{Y \times R}$ for $k \in \{1, 2, \ldots, K\}$:

$$\mathbf{m}' \stackrel{\text{def}}{=} \mu + \mathbf{Z}\,\mathbf{m} \tag{3.1}$$

$$\mathbf{U}' \stackrel{\text{def}}{=} \mathbf{1}\mathbf{u}^\top + \mathbf{Z}\,\mathbf{U} \tag{3.2}$$

$$\mathbf{V}'_k \stackrel{\text{def}}{=} \mathbf{1}\mathbf{v}_k^\top + \mathbf{Z}\,\mathbf{V}_k \tag{3.3}$$

To support convolution, let $\mathbf{F} \in \{0, 1\}^{Y \times X}$ be a mask that admits only those observations in $\mathbf{X}$ that precede each $\mathbf{y}_{[y]}$ in the same time series, for $1 \leq x \leq X$, $1 \leq y \leq Y$:

$$\mathbf{F}_{[y,x]} \stackrel{\text{def}}{=} \begin{cases} 1 & (\mathbf{c}_{[x]} = \mathbf{d}_{[y]}) \text{ and } (\mathbf{t}_{[x]} \leq \mathbf{t}'_{[y]}) \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

K sparse convolution matrices $\mathbf{G}_k \in \mathbb{R}^{Y \times X}$ for $k \in \{1, 2, \ldots, K\}$ are defined as follows:

$$\mathbf{G}_k \stackrel{\text{def}}{=} g_k\left(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}'_k\right) \odot \mathbf{F} \tag{3.5}$$

The convolution that yields the design matrix of convolved predictors $\mathbf{X}' \in \mathbb{R}^{Y \times K}$ is then defined using a product of the convolution matrices and the design matrix:

$$\mathbf{X}'_{[*,k]} \stackrel{\text{def}}{=} \mathbf{G}_k\,\mathbf{X}_{[*,k]} \tag{3.6}$$

The full model mean is the sum of (1) the intercepts and (2) the sum-product of the convolved predictors with the coefficient parameters:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{m}' + (\mathbf{X}' \odot \mathbf{U}')\,\mathbf{1}, \sigma^2\right) \tag{3.7}$$

A summary table of the variable definitions above is provided in Table 3.1, and a step-through example of the CDR equations is provided in §3.1.1.

Note that this is an event-based implementation of convolution that is only exact when the predictors fully describe discrete impulse signals. Exact convolution of samples from continuous signals is generally not possible because the signal is generally not analytically integrable. For continuous signals, the CDR procedure above defines a Riemann sum approximation of the integral as long as (1) the predictor is sampled at a fixed frequency or (2) the predictor is interpolated at a fixed frequency between variably-spaced samples.

3.1.1 CDR Model: A Worked Example

Consider a model containing two predictors p_1 and p_2 and two random effects levels s_1 and s_2. Assume the following 6 rows for each of $\mathbf{X}$ (predictors), $\mathbf{t}$ (predictor timestamps), and $\mathbf{c}$ (predictor series IDs):

$$\mathbf{X} = \begin{matrix} p_1 & p_2 \end{matrix} \begin{bmatrix} 5 & 0 \\ -1 & 3 \\ 6 & 1 \\ 2 & 2 \\ 0 & 1 \\ -2 & -1 \end{bmatrix}, \ \mathbf{t} = \begin{bmatrix} 0 \\ 2 \\ 3 \\ 0 \\ 1 \\ 4 \end{bmatrix}, \ \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}$$

Assume the following four rows for each of $\mathbf{Z}$ (random effects indicator), $\mathbf{t}'$ (response timestamps), and $\mathbf{d}$ (response series IDs):

$$\mathbf{Z} = \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} s_1 & s_2 \end{matrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \; \mathbf{t}' = \begin{bmatrix} 1 \\ 4 \\ 2 \\ 3 \end{bmatrix}, \; \mathbf{d} = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix},$$

Assume a Gaussian IRF kernel with scalar location and scale parameters w_1, w_2:

$$g_1(x; w_1, w_2) = g_2(x; w_1, w_2) = e^{\frac{-(x-w_1)^2}{w_2}}$$

Assume the following model parameters μ, $\mathbf{m}$, $\mathbf{u}$, $\mathbf{U}$, $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{V}_1$, $\mathbf{V}_2$, and σ_2. Note that random effects $\mathbf{m}$, $\mathbf{U}$, $\mathbf{V}_1$, and $\mathbf{V}_2$ are zero-centered:

$$\mu = 1.2, \; \mathbf{m} = \begin{matrix} s_1 \\ s_2 \end{matrix} \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}, \; \mathbf{u} = \begin{matrix} p_1 \\ p_2 \end{matrix} \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}, \; \mathbf{U} = \begin{matrix} \\ s_1 \\ s_2 \end{matrix} \begin{matrix} p_1 & p_2 \\ \end{matrix} \begin{bmatrix} -0.3 & 0.7 \\ 0.3 & -0.7 \end{bmatrix}$$

$$\mathbf{v}_1 = \begin{matrix} \\ w_1 \\ w_2 \end{matrix} \begin{matrix} p_1 \\ \end{matrix} \begin{bmatrix} 0.8 \\ 1.7 \end{bmatrix}, \; \mathbf{v}_2 = \begin{matrix} \\ w_1 \\ w_2 \end{matrix} \begin{matrix} p_2 \\ \end{matrix} \begin{bmatrix} 1.1 \\ 0.5 \end{bmatrix}, \; \mathbf{V}_1 = \begin{matrix} \\ s_1 \\ s_2 \end{matrix} \begin{matrix} w_1 & w_2 \\ \end{matrix} \begin{bmatrix} -0.2 & 0.1 \\ 0.2 & -0.1 \end{bmatrix}, \; \mathbf{V}_2 = \begin{matrix} \\ s_1 \\ s_2 \end{matrix} \begin{matrix} w_1 & w_2 \\ \end{matrix} \begin{bmatrix} 0.3 & 0.4 \\ -0.3 & -0.4 \end{bmatrix}$$

$$\sigma^2 = 1.3$$

The CDR equations are used to generate estimates for the four elements of $\mathbf{y}$ using the inputs and parameters. The vector $\mathbf{m}' = \mu + \mathbf{Z}\,\mathbf{m}$ contains intercepts for each element of $\mathbf{y}$

and is computed as follows:

$$\mathbf{m}' = \overbrace{1.3}^{\mu} + \overbrace{\begin{matrix} s_1 & s_2 \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \end{matrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}}^{\mathbf{m}}$$

$$= \overbrace{1.3}^{\mu} + \overbrace{\begin{bmatrix} -0.2 \\ -0.2 \\ 0.2 \\ 0.2 \end{bmatrix}}^{\mathbf{Z}\,\mathbf{m}}$$

$$= \overbrace{\begin{bmatrix} 1.1 \\ 1.1 \\ 1.5 \\ 1.5 \end{bmatrix}}^{\mu+\mathbf{Z}\,\mathbf{m}}$$

The matrix $\mathbf{U}' = \mathbf{1}\mathbf{u}^\top + \mathbf{Z}\,\mathbf{U}$ contains coefficients for both predictors for each element of $\mathbf{y}$ and is computed as follows:

$$\mathbf{U}' = \overbrace{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix}0.1 & 0.5\end{bmatrix}}^{\mathbf{u}^\top} + \overbrace{\begin{bmatrix}1 & 0\\1 & 0\\0 & 1\\0 & 1\end{bmatrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix}p_1 & p_2\\-0.3 & 0.7\\0.3 & -0.7\end{bmatrix}}^{\mathbf{U}}$$

$$= \overbrace{\begin{bmatrix}0.1 & 0.5\\0.1 & 0.5\\0.1 & 0.5\\0.1 & 0.5\end{bmatrix}}^{\mathbf{1}\mathbf{u}^\top} + \overbrace{\begin{bmatrix}-0.3 & 0.7\\-0.3 & 0.7\\0.3 & -0.7\\0.3 & -0.7\end{bmatrix}}^{\mathbf{Z}\,\mathbf{U}}$$

$$= \overbrace{\begin{bmatrix}-0.2 & 1.2\\-0.2 & 1.2\\0.4 & -0.2\\0.4 & -0.2\end{bmatrix}}^{\mathbf{1}\mathbf{u}^\top + \mathbf{Z}\,\mathbf{U}}$$

The matrices $\mathbf{V}_1' = \mathbf{1}\mathbf{v}_1^\top + \mathbf{Z}\,\mathbf{V}_1$ and $\mathbf{V}_2' = \mathbf{1}\mathbf{v}_2^\top + \mathbf{Z}\,\mathbf{V}_2$ contain IRF parameters for responses to p_1 and p_2, respectively, for each of the four elements of $\mathbf{y}$. They are computed as follows:

$$\mathbf{V}_1 = \overbrace{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix}0.8 & 1.7\end{bmatrix}}^{\mathbf{v}_1^\top} + \overbrace{\begin{matrix}s_1 & s_2\end{matrix}}^{\mathbf{Z}} \begin{bmatrix}1 & 0\\1 & 0\\0 & 1\\0 & 1\end{bmatrix} \overbrace{\begin{matrix}p_1 & p_2\end{matrix}}^{\mathbf{V}_1} \begin{bmatrix}-0.2 & 0.1\\0.2 & -0.1\end{bmatrix}$$

$$= \overbrace{\begin{bmatrix}0.8 & 1.7\\0.8 & 1.7\\0.8 & 1.7\\0.8 & 1.7\end{bmatrix}}^{\mathbf{1}\mathbf{v}_1^\top} + \overbrace{\begin{bmatrix}-0.2 & 0.1\\-0.2 & 0.1\\0.2 & -0.1\\0.2 & -0.1\end{bmatrix}}^{\mathbf{Z}\,\mathbf{V}_1}$$

$$= \overbrace{\begin{bmatrix}0.6 & 1.8\\0.6 & 1.8\\1.0 & 1.6\\1.0 & 1.6\end{bmatrix}}^{\mathbf{1}\mathbf{v}_1^\top + \mathbf{Z}\,\mathbf{V}_1}$$

$$\mathbf{V}_2' = \overbrace{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix}1.1 & 0.5\end{bmatrix}}^{\mathbf{v}_2^\top} + \overbrace{\begin{matrix}s_1 & s_2\end{matrix}}^{\mathbf{Z}} \overbrace{\begin{bmatrix}1 & 0\\1 & 0\\0 & 1\\0 & 1\end{bmatrix}}^{} \overbrace{\begin{matrix}p_1 & p_2\end{matrix}}^{\mathbf{V}_2} \begin{bmatrix}0.3 & 0.4\\-0.3 & -0.4\end{bmatrix}$$

$$= \overbrace{\begin{bmatrix}1.1 & 0.5\\1.1 & 0.5\\1.1 & 0.5\\1.1 & 0.5\end{bmatrix}}^{\mathbf{1}\mathbf{v}_2^\top} + \overbrace{\begin{bmatrix}0.3 & 0.4\\0.3 & 0.4\\-0.3 & -0.4\\-0.3 & -0.4\end{bmatrix}}^{\mathbf{Z}\,\mathbf{V}_2}$$

$$= \overbrace{\begin{bmatrix}1.4 & 0.9\\1.4 & 0.9\\0.8 & 0.1\\0.8 & 0.1\end{bmatrix}}^{\mathbf{1}\mathbf{v}_2^\top + \mathbf{Z}\,\mathbf{V}_2}$$

The mask matrix $\mathbf{F}_{[y,x]} \overset{\text{def}}{=} \begin{cases} 1 & (\mathbf{c}_{[x]} = \mathbf{d}_{[y]}) \text{ and } (\mathbf{t}_{[x]} \leq \mathbf{t}'_{[y]}) \\ 0 & \text{otherwise} \end{cases}$ indicates predictor observations that precede each element of $\mathbf{y}$ in the same time series. Timestamp vectors $\mathbf{t}$, $\mathbf{t}'$ and

series ID vectors $\mathbf{c}$, $\mathbf{d}$ are shown on the top and left axes for expository purposes.

$$
\mathbf{F} = \begin{array}{cc} & \\ & \\ 1 & 1 \\ 4 & 1 \\ 2 & 2 \\ \underbrace{3}_{\mathbf{t}'} & \underbrace{2}_{\mathbf{d}} \end{array}
\begin{array}{c}
\begin{array}{cccccc} 0 & 2 & 3 & 0 & 1 & 4 \end{array} \} \mathbf{t} \\
\begin{array}{cccccc} 1 & 1 & 1 & 2 & 2 & 2 \end{array} \} \mathbf{c} \\
\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}
\end{array}
$$

To compute convolution matrices $\mathbf{G}_1$, $\mathbf{G}_2$, an array $\mathbf{t'1}^\top - \mathbf{1t}$ is first computed, containing distance in time of predictors from responses:

$$\mathbf{t'1}^\top - \mathbf{1t} = \overbrace{\begin{bmatrix}1\\4\\2\\3\end{bmatrix}}^{\mathbf{t'}} \overbrace{\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\end{bmatrix}}^{\mathbf{1}^\top} - \overbrace{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}}^{\mathbf{1}} \overbrace{\begin{bmatrix}0 & 2 & 3 & 0 & 1 & 4\end{bmatrix}}^{\mathbf{t}^\top}$$

$$= \overbrace{\begin{bmatrix}1 & 1 & 1 & 1 & 1 & 1\\4 & 4 & 4 & 4 & 4 & 4\\2 & 2 & 2 & 2 & 2 & 2\\3 & 3 & 3 & 3 & 3 & 3\end{bmatrix}}^{\mathbf{t'}\,\mathbf{1}^\top} - \overbrace{\begin{bmatrix}0 & 2 & 3 & 0 & 1 & 4\\0 & 2 & 3 & 0 & 1 & 4\\0 & 2 & 3 & 0 & 1 & 4\\0 & 2 & 3 & 0 & 1 & 4\end{bmatrix}}^{\mathbf{1t}^\top}$$

$$= \overbrace{\begin{bmatrix}1 & -1 & -2 & 1 & 0 & -3\\4 & 2 & 1 & 4 & 3 & 0\\2 & 0 & -1 & 2 & 1 & -2\\3 & 1 & 0 & 3 & 2 & -1\end{bmatrix}}^{\mathbf{t'}\,\mathbf{1}^\top - \mathbf{1t}^\top}$$

These distances are supplied as inputs to the impulse response functions g_1, g_2, and irrelevant cells (i.e. cells from the future or cells from other time series) are masked using $\mathbf{F}$. The resulting convolution matrices $\mathbf{G}_1 = g_1(\mathbf{t'}\,\mathbf{1}^\top - \mathbf{1t}^\top;\, \mathbf{V}'_1) \odot \mathbf{F}$, $\mathbf{G}_2 = g_2(\mathbf{t'}\,\mathbf{1}^\top - \mathbf{1t}^\top;\, \mathbf{V}'_2) \odot \mathbf{F}$ contain the convolution weights to apply to the elements of $\mathbf{X}$ in order to generate $\mathbf{y}$. They are computed as follows, where g_k is parameterized row-wise by $\mathbf{V}'_k$ and applied elementwise

to $\mathbf{t}'\,\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top$:

$$\mathbf{G}_1 = g_1\left(\overbrace{\begin{bmatrix} 1 & -1 & -2 & 1 & 0 & -3 \\ 4 & 2 & 1 & 4 & 3 & 0 \\ 2 & 0 & -1 & 2 & 1 & -2 \\ 3 & 1 & 0 & 3 & 2 & -1 \end{bmatrix}}^{\mathbf{t}'\,\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top}; \overbrace{\begin{bmatrix} 0.6 & 1.8 \\ 0.6 & 1.8 \\ 1.0 & 1.6 \\ 1.0 & 1.6 \end{bmatrix}}^{\mathbf{V}_1'}\right) \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$

$$= \overbrace{\begin{bmatrix} 0.91 & 0.24 & 0.02 & 0.91 & 0.82 & 0.00 \\ 0.00 & 0.34 & 0.91 & 0.00 & 0.04 & 0.00 \\ 0.54 & 0.54 & 0.08 & 0.54 & 1.00 & 0.00 \\ 0.08 & 1.00 & 0.54 & 0.08 & 0.54 & 0.08 \end{bmatrix}}^{g_1(\mathbf{t}'\,\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_1')} \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$

$$= \overbrace{\begin{bmatrix} 0.91 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.34 & 0.91 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.54 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.08 & 0.54 & 0.00 \end{bmatrix}}^{g_1(\mathbf{t}'\,\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_1') \odot \mathbf{F}}$$

$$\mathbf{G}_2 = g_2\left(\overbrace{\begin{bmatrix} 1 & -1 & -2 & 1 & 0 & -3 \\ 4 & 2 & 1 & 4 & 3 & 0 \\ 2 & 0 & -1 & 2 & 1 & -2 \\ 3 & 1 & 0 & 3 & 2 & -1 \end{bmatrix}}^{\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top}; \overbrace{\begin{bmatrix} 1.4 & 0.9 \\ 1.4 & 0.9 \\ 0.8 & 0.1 \\ 0.8 & 0.1 \end{bmatrix}}^{\mathbf{V}_2'}\right) \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$

$$= \overbrace{\begin{bmatrix} 0.84 & 0.00 & 0.00 & 0.84 & 0.11 & 0.00 \\ 0.00 & 0.67 & 0.84 & 0.00 & 0.06 & 0.11 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.67 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}}^{g_2(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_2')} \odot \overbrace{\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}}^{\mathbf{F}}$$

$$= \overbrace{\begin{bmatrix} 0.84 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.67 & 0.84 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}}^{g_2(\mathbf{t}'\mathbf{1}^\top - \mathbf{1}\mathbf{t}^\top; \mathbf{V}_2') \odot \mathbf{F}}$$

The two columns of the convolved predictor matrix $\mathbf{X}' = \mathbf{G}_2\,\mathbf{X}_{[*,2]}$ are computed by pre-multiplying each column with its corresponding convolution matrix:

$$\mathbf{X}'_{[*,1]} = \overbrace{\begin{bmatrix} 0.91 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.34 & 0.91 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.54 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.08 & 0.54 & 0.00 \end{bmatrix}}^{\mathbf{G}_1} \overbrace{\begin{bmatrix} 5 \\ -1 \\ 6 \\ 2 \\ 0 \\ -2 \end{bmatrix}}^{\mathbf{X}_{[*,1]}} \begin{matrix} \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \\ p_1 \end{matrix}$$

$$= \overbrace{\begin{bmatrix} 4.55 \\ 5.15 \\ 1.08 \\ 0.16 \end{bmatrix}}^{\mathbf{G}_1\,\mathbf{X}_{[*,1]}}$$

$$\mathbf{X}'_{[*,2]} = \overbrace{\begin{bmatrix} 0.84 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.67 & 0.84 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.67 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix}}^{\mathbf{G}_2} \overbrace{\begin{bmatrix} 0 \\ 3 \\ 1 \\ 2 \\ 1 \\ -1 \end{bmatrix}}^{\mathbf{X}_{[*,2]}}$$

$$p_1$$

$$= \overbrace{\begin{bmatrix} 0.00 \\ 2.85 \\ 0.67 \\ 0.00 \end{bmatrix}}^{\mathbf{G}_2\,\mathbf{X}_{[*,2]}}$$

$$\mathbf{X}' = \begin{bmatrix} 4.55 & 0.00 \\ 5.15 & 2.85 \\ 1.08 & 0.67 \\ 0.16 & 0.00 \end{bmatrix}$$

The expected response $\hat{\mathbf{y}} = \mathbf{m}' + (\mathbf{X}' \odot \mathbf{U}')\mathbf{1}$ is computed by rescaling $\mathbf{X}'$ by the coefficient matrix $\mathbf{U}'$, summing across predictors, and shifting by the intercept $\mathbf{m}'$, as shown:

$$\hat{\mathbf{y}} = \overbrace{\begin{bmatrix} 1.1 \\ 1.1 \\ 1.5 \\ 1.5 \end{bmatrix}}^{\mathbf{m}'} + \left(\overbrace{\begin{bmatrix} 4.55 & 0.00 \\ 5.15 & 2.85 \\ 1.08 & 0.67 \\ 0.16 & 0.00 \end{bmatrix}}^{\mathbf{X}'} \odot \overbrace{\begin{bmatrix} -0.2 & 1.2 \\ -0.2 & 1.2 \\ 0.4 & -0.2 \\ 0.4 & -0.2 \end{bmatrix}}^{\mathbf{U}'} \right) \overbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}^{\mathbf{1}}$$

$$= \overbrace{\begin{bmatrix} 1.1 \\ 1.1 \\ 1.5 \\ 1.5 \end{bmatrix}}^{\mathbf{m}'} + \overbrace{\begin{bmatrix} -0.91 & 0.00 \\ -1.03 & 3.42 \\ 0.43 & -0.13 \\ 0.06 & 0.00 \end{bmatrix}}^{\mathbf{X}' \odot \mathbf{U}'} \overbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}^{\mathbf{1}}$$

$$= \overbrace{\begin{bmatrix} 1.1 \\ 1.1 \\ 1.5 \\ 1.5 \end{bmatrix}}^{\mathbf{m}'} + \overbrace{\begin{bmatrix} -0.91 \\ 2.39 \\ 0.30 \\ 0.06 \end{bmatrix}}^{(\mathbf{X}' \odot \mathbf{U}')\mathbf{1}}$$

$$= \overbrace{\begin{bmatrix} 0.19 \\ 3.49 \\ 1.80 \\ 1.56 \end{bmatrix}}^{\mathbf{m}' + (\mathbf{X}' \odot \mathbf{U}')\mathbf{1}}$$

3.2 Effect Estimates in CDR

Many scientific applications of linear modeling are interested in testing a null hypothesis about the scalar-valued *effect estimate* obtained for a predictor (e.g. that it is equal to 0). Because CDR models estimate continuous functions of time rather than scalars (as in linear regression), the estimated IRF must be distilled into a scalar in order to yield a comparable notion. Here, CDR effect estimates are defined by integrating the IRFs, since the integral describes the total expected influence on the response from observing a unit impulse of each predictor. In particular, the unscaled and scaled fixed effect estimates $\mathbf{g}$, $\mathbf{g}' \in \mathbb{R}^K$ are defined as follows, where scaling is performed using the coefficient vector $\mathbf{u}$ (§3.1):

$$\mathbf{g}_{[k]} \stackrel{\text{def}}{=} \int_0^\infty g_k\left(t; \mathbf{v}_k\right) dt \tag{3.8}$$

$$\mathbf{g}' \stackrel{\text{def}}{=} \mathbf{g} \odot \mathbf{u} \tag{3.9}$$

$$\tag{3.10}$$

In mixed-effects CDR models with random impulse response parameters, the IRF shape — and therefore the integral of the IRF — can vary between levels of the random grouping factor. As a result, zero-centering the random coefficients $\mathbf{U}$ within each grouping factor is insufficient to guarantee zero-centered random effect estimates. For example, in a 2-level mixed univariate CDR model with fixed effect sizes $\mathbf{g}' = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, random coefficients $\mathbf{U} = \begin{bmatrix} -1 & 1 \end{bmatrix}^\top$, and unscaled random effect estimates (IRF integrals) $\mathbf{H} = \begin{bmatrix} 1 & 10 \end{bmatrix}^\top$, $\mathbf{U}$ has mean 0, but the random effect estimate vector $\mathbf{H}' = \mathbf{H} \odot \mathbf{U} - \mathbf{1}\mathbf{g}'^\top = \begin{bmatrix} -2 & 9 \end{bmatrix}^\top$ has mean 3.5, yielding a biased population level effect estimate.

To overcome this, IRFs g_k are constrained to have a unit integral over the positive real line:

$$1 = \int_0^\infty g_k(t;\theta)\,dt; \;\; k \in \{1, 2, \ldots, K\}, \;\; \theta \in \mathbb{R}^R \tag{3.11}$$

Under this constraint, zero-centered coefficients are guaranteed to yield zero-centered effect estimates, and the population-level (fixed) effect estimates are unbiased.[5]

An important implementational consideration for finite training data is that the model will not have empirical support over the positive infinite real line, and thus the infinite integral involves some degree of extrapolation. To ensure that effect estimates have strong empirical support, in practice the integral in eq. 3.9 is upper bounded at the 75^{th} percentile of temporal offsets seen in training. This seemed to be a reasonable default that concentrates the effect estimate on empirically well-attested regions of the support of the IRF. However, particular research questions may motivate the use of other kinds of bounds (e.g. if the research domain imposes a principled constraint on the duration of interest for the IRF).

3.3 The Deconvolutional Intercept

In linear regression, the *intercept* is a bias term implemented by fitting a coefficient to a vector of ones, one for each data point. The intercept term estimates the base response of the system when the other predictors are equal to 0. Because CDR contains a linear model on the convolved predictors (eq. 3.7), it is just as important to include an intercept term in CDR models as in linear ones. However, in a deconvolutional setting, it is also possible that the response is partially described by the *timing* of stimuli alone, independently of their properties. This possibility can be brought under control by additionally convolving the intercept with an estimated impulse response. Analagously to a linear intercept term, this

[5] In models without random IRF parameters, the IRF integrals are identical across levels of the random grouping factor, and effect estimates are thus unbiased with or without normalization. Normalizing can still have numerical advantages since it factors effect size and shape, so we apply normalization to all models reported here, regardless of whether they contain random IRF parameters.

convolved intercept estimates the base response of the system when the other predictors are equal to 0, but unlike the linear intercept, the convolved intercept is sensitive to stimulus timing. The estimate for the deconvolutional intercept is therefore the expected change in the response over time from observing an event, regardless of the properties of that event. I refer to this deconvolutional intercept as *rate* (see also e.g. Brennan et al., 2016) and consider it to be an essential control to include in CDR models. Without it, variance in the response due to event timing must be captured by other components in the model, which is potentially problematic for interpretation and hypothesis testing.

Depending on the problem definition, *rate* effects may have a theoretical interpretation. For example, as argued in §4.2, a negative *rate* estimate in reading data can be seen as an inertia effect, since the negative *rate* contributions of preceding words compound to suppress reading times on the current word as a function of their recency. In other words, fast reading in the recent past will engender fast reading now, a possibility which *rate* estimates allow the model to account for.

Note that *rate* can only be estimated in a continuous-time deconvolutional setting because variation over time in the rate of events is necessary to identify it. In an FIR model, the *rate* predictor is equal to 1 at every timestep, rendering it identical to the intercept. The ability to detect *rate* effects in variably-spaced time series is a major advantage of CDR.

3.4 Scale and Shift in CDR Models

Linear models are invariant to transformations that rescale and/or shift the design matrix, in that for a linear model f with intercept p, slopes $\mathbf{q}$, predictors $\mathbf{X}$, scale vector $\mathbf{r}$, and

shift vector $\mathbf{s}$, the following identity holds:

$$
\begin{aligned}
f\left(\mathbf{X}\,\mathrm{diag}(\mathbf{r}) + \mathbf{1}\mathbf{s}^\top; p, \mathbf{q}\right) &= p + \left(\mathbf{X}\,\mathrm{diag}(\mathbf{r}) + \mathbf{1}\mathbf{s}^\top\right)\mathbf{q} \\
&= p + \mathbf{s}^\top\mathbf{q} + \mathbf{X}\,\mathrm{diag}(\mathbf{r})\mathbf{q} \\
&= f\left(\mathbf{X}; p + \mathbf{s}^\top\mathbf{q}, \mathrm{diag}(\mathbf{r})\mathbf{q}\right)
\end{aligned} \tag{3.12}
$$

In other words, a linear model of shifted/rescaled $\mathbf{X}$ is equivalent to a shifted/rescaled linear model of $\mathbf{X}$. This result entails that (non-interacted) predictors can be shifted and rescaled (e.g. standardized) prior to fitting without altering the solution, which can help with numerical optimization of complex linear models.

In a non-linear optimization setting such as that required for CDR, normalization can be helpful for accelerating convergence (Ioffe and Szegedy, 2015; Salimans and Kingma, 2016; Ba et al., 2016), and this section therefore explores the impact of scale and shift of the inputs to CDR models. Invariance under rescaling follows trivially from eq. 3.6. It also follows from eq. 3.6 that CDR is *not* invariant under additive shift $\mathbf{s}$:

$$
\mathbf{G}_k\left(\mathbf{X}_{[*,k]} + \mathbf{1}\mathbf{s}_{[k]}\right) = \mathbf{G}_k\,\mathbf{X}_{[*,k]} + \mathbf{G}_k\,\mathbf{1}\mathbf{s}_{[k]} \tag{3.13}
$$

As shown, the shift scalar $\mathbf{s}_{[k]}$ is also convolved with $\mathbf{G}_k$, resulting in an additional term $\mathbf{G}_k\,\mathbf{1}\mathbf{s}_{[k]} \in \mathbb{R}^N$ which cannot be absorbed by the intercept because its value differs for each data point (unlike $\mathbf{s}^\top\mathbf{q}$ in eq. 3.12, whose value is identical for all elements of $\mathbf{X}$). However, note that by convolving a matrix of ones, $\mathbf{G}_k\,\mathbf{1}\mathbf{s}_{[k]}$ implicitly defines a deconvolutional intercept term with IRF g_k and scale $\mathbf{u}_{[k]}\mathbf{s}_{[k]}$. In other words, shifting K predictors introduces K deconvolutional intercepts, each with IRF shape and scale tied to the shape and scale estimates of the predictors themselves. Since these deconvolutional intercepts are summed together in eq. 3.7 and convolution obeys the distributive property, they together define a

new impulse response g_0 for the deconvolutional intercept:

$$g_0(t) = \sum_{k=1}^{K} \mathbf{u}_{[k]} \mathbf{s}_{[k]} g_k(t) \tag{3.14}$$

Since $g_0(t) = 0$ can only be guaranteed if $\mathbf{s} = \mathbf{0}$, the model is not invariant under shift. However, in models with an explicit *rate* predictor as recommended in §3.3, g_0 simply modulates the IRF estimate for *rate*. Under the conventional assumption that the intercept (*rate* predictor) is the first column of the design matrix $\mathbf{X}$, the implicit kernel g_1' for the deconvolutional intercept in models with shift can be computed as:

$$g_1'(t) = g_0(t) + g_1(t) \tag{3.15}$$

The deconvolutional intercept thus does "absorb" shift in a limited sense. It must nonetheless be kept in mind that unless identity is enforced between $g_{1,\dots,K}$, the estimated shape of g_1' will consist of a sum of response kernels and can therefore fall outside the solution space defined by the parametric IRF kernel assigned to *rate*.

3.5 Multicollinearity

The formulation in eq. 3.7 is simply a linear model on the convolved design matrix $\mathbf{X}'$. Therefore, the primary difference between linear and CDR models is that CDR additionally infers the parameters that generate $\mathbf{X}'$ jointly with the model intercept and coefficients.

Since CDR depends internally on linear combination to generate its outputs, it is vulnerable to confounds from multicollinearity (correlated predictors) in much the same way that linear models are. In linear models, multicollinearity increases uncertainty about how to allocate covariation between predictors and response, since the predictors themselves covary. In the extreme case of perfect multicollinearity (i.e. one or more predictors are an

exact linear combination of one or more other predictors), the model has no solution (Neter et al., 1989).

Multicollinearity in CDR works in much the same way, with the added complexity that CDR models also have a temporal dimension which may allow the fitting procedure to discover real characteristics of the global impulse response structure while struggling proportionally to the degree of multicollinearity to decompose that structure into predictor-specific IRFs. To understand this, note that the expected response t seconds after stimulus presentation is a weighted sum of the IRFs at t, with weights provided by the predictor values of the stimulus. When multicollinearity is low, the expected overall response can vary widely from one stimulus to another, since the IRFs are reweighted at each stimulus by roughly orthogonal predictor values. This variation in expected overall response provides clues to the system as to the magnitude, direction, and temporal shape of the individual response to each predictor. As multicollinearity increases, the expected overall response increasingly converges to a single shape which is shared across all stimuli (albeit scaled by the stimulus magnitude). In this setting, the model should still be able to correctly recover the global response characteristics, but may decompose it into predictor-specific responses that increasingly deviate from the true data generating model. In the extreme case of perfect multicollinearity, the expected response is identical for each stimulus, and the model will construct IRFs whose summation approximates the true global response profile but whose attribution of IRF components to predictors is arbitrary.

Empirical results (§4.1.4) indicate that CDR models are quite robust to multicollinearity. Nonetheless, models fitted to highly collinear data should be interpreted with caution, and perfectly collinear data should be avoided altogether. As in linear models, multicollinearity can be avoided by orthogonalizing predictors in advance (e.g. via principal components analysis). Empirical evaluation of orthogonalization procedures in the CDR setting is left to future work, and no orthogonalization is used in any of the analyses presented here.

3.6 Hypothesis Testing

The ultimate scientific purpose of most statistical models is to test a claim about nature. Hypothesis testing is challenging in a CDR context for two reasons. First, CDR currently does not have analytical estimators for standard errors of the model parameters. As a result, exact null hypothesis significance tests cannot be performed on the basis of a single model. Single-model tests are nonetheless also problematic in a linear regression context when predictors are collinear (Neter et al., 1989), a pervasive issue in psycholinguistics that has motivated a shift toward ablative tests based on model comparison (Frank and Bod, 2011). Second, reliance on non-linear stochastic optimization introduces estimation noise through the possibility of imperfect convergence to an optimum or convergence to a non-global optimum. As a result, training likelihood cannot be guaranteed to be maximized, which can result in degenerate outcomes for in-sample ablative tests (e.g. the ablated model can have better likelihood than the full model).

For this reason, two tests that are commonly used for linear models may be unreliable for CDR. First, tests based on credible intervals may be unreliable because the credible intervals produced by (variational) Bayesian CDR reflect the local neighborhood of the discovered solution (optimum), which may not account for the existence of more distant optima.[6] Credible intervals tests in CDR are therefore anticonservative. Indeed, the analyses reported below show that CDR-estimated credible intervals tend to be very tight. Second, likelihood ratio testing (LRT) may be unreliable because the test statistic is a function of the maximum likelihood estimates, and CDR likelihood cannot be guaranteed to be maximized. Instead, this thesis considers two types of hypothesis test for CDR models: (1) a **direct test** by bootstrap comparison of model fit to out-of-sample data, and (2) a **2-step test** in which CDR is used first to estimate a data-driven convolution $\mathbf{X}'$ of the design matrix $\mathbf{X}$ and then existing statistical models (e.g. OLS, LME, GAM) are fitted to $\mathbf{X}'$ and used to perform the

[6] Maximum likelihood CDR models do not estimate uncertainty, and therefore CDR does not support (frequentist) confidence intervals.

test.

To perform a direct test, training and evaluation sets must be created, either by running two separate experiments or by partitioning the data from a single experiment.[7] Two CDR models are fitted to the training set, one with a fixed effect for the variable to be tested (full model), and one without one (ablated model). Out-of-sample error vectors are then generated by predicting from each model on the evaluation set, and an aggregate test statistic (e.g. absolute difference in mean squared error) is computed over the two vectors. To perform the test (a paired permutation test), an empirical distribution is created for the test statistic: for n iterations, the by-item errors from each model are randomly swapped pairwise to generate two new error vectors, and a new test statistic over the resampled errors is computed and stored. The test rejects the null hypothesis at level α if the observed test statistic is greater than $(1 - \alpha) \times 100\%$ of the resampled test statistics.

To perform a 2-step test, a single CDR model containing all fixed effects of interest is fitted to the data, and the predictors are convolved using the estimated IRFs. Standard statistical models (e.g. OLS, LME, GAM) are then fitted to the convolved predictors and used to perform any of the tests that they support (e.g. LRT). Note that to perform an ablative test like LRT in a 2-step setting, the ablation is only applied at the second step (e.g. the LME stage). If ablation is also applied at the CDR stage, then the predictors in the full and ablated models *are not necessarily the same*, invalidating the test.

Both tests potentially suffer from non-convexity, since they are both conditional on possibly sub-optimal IRF estimates. Nonetheless, I offer the following arguments in defense of using CDR for hypothesis testing. First, the synthetic experiments reported in §4.1 show a strong tendency for CDR to closely recover the true data-generating model, even under adverse training conditions like variably spaced events, multicollinearity, and ill-fitting IRF

[7] When partitioning and/or filtering outliers prior to CDR fitting, it is important to keep in mind that partitioning and outlier filtering should only be performed on the *response* vector. Partitioning/filtering the design matrix is equivalent to assuming that the removed events did not take place, which can distort the IRF estimates. The CDR software library described in §3.7 provides utilities for data partitioning and filtering, which automatically apply only to the response data.

kernels. There is thus empirical reason to believe that convergence to a bad optimum is not a serious problem in practice. Second, the assumption that the models fall within a tolerance of the global optimum (i.e. are "good enough") also underlies ablative tests in popular linear regression libraries like `lme4`, which use numerical optimization and define a tolerance-based stopping criterion. Third, in pursuit of understanding complex non-linear phenomena like human language comprehension, CDR may permit discovery of previously unknown patterns precisely by relaxing the strict linearity and independence assumptions that support linear models' convenient statistical and mathematical properties. Optimality guarantees are of little value if the underlying generative process lies far outside the model's solution space.

The direct test potentially suffers more than the 2-step test from the issue of non-convexity, since in the 2-step test the coefficients benefit from convergence guarantees at the second step (e.g. LME) and the IRFs do not vary with ablation. However, the 2-step test potentially suffers more than the direct test from multicollinearity, since it cannot adjust IRF shapes in the ablated model that might have been influenced by multicollinear predictors in the full model. And although the 2-step procedure alone can guarantee maximum training likelihood conditional on the fitted IRF, this may not be of critical importance because the direct test does not implicitly require that training likelihood is maximized, since it is based on out-of-sample error rather than asymptotic distributional guarantees (cf. LRT). Indeed, because of the possibility of overfitting, much research in statistics and machine learning has been dedicated to the study of regularization techniques and stopping criteria that avoid minimizing training error in pursuit of minimizing generalization error (Yao et al., 2007; Raskutti et al., 2014; Srivastava et al., 2014), and out-of-sample bootstrap model comparison is one of the most widely used statistical tests for non-convex model comparison in machine learning (Demšar, 2006). Refocusing on out-of-sample rather than in-sample performance has the added benefit of building external model validity directly into the statistical test, which is potentially timely in light of growing concern over the replication

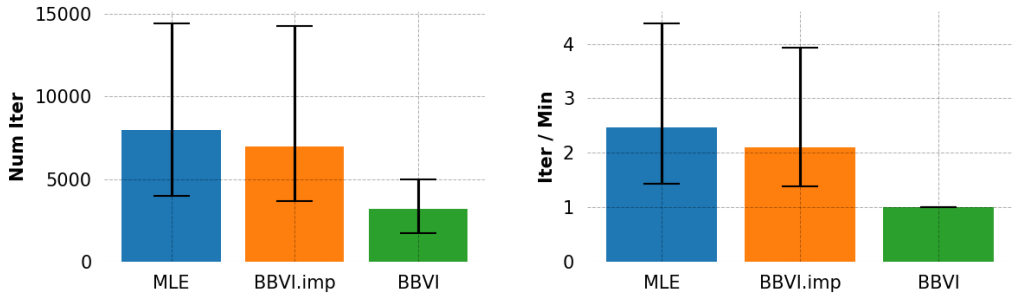
	Synthetic			Experimental		
Inference	Median	25%	75%	Median	25%	75%
MLE	1.000	1.000	1.001	1.002	0.997	1.011
BBVI.imp	1.000	1.000	1.000	1.002	0.997	1.014
BBVI	1.010	1.002	1.076	1.003	0.973	1.027

Table 3.2: Distribution of fixed effects estimates of LME models fitted to CDR-convolved synthetic and human subjects (Experimental) data (median, 25^{th} percentile, and 75^{th} percentile). Estimates concentrate near 1, indicating that CDR-estimated coefficients are generally close to the global optimum given the IRF.

crisis in psychological science (Pashler and Wagenmakers, 2012; Makel and Plucker, 2014; Simons, 2014; Open Science Collaboration, 2015; Gilmore et al., 2017; Yarkoni and Westfall, 2017). For this reason, I argue that the direct test is a defensible method for evaluating scientific hypotheses using CDR.

One additional possibility is a hybrid of these two approaches where 2-step fitting is used over the training set and then the second step fits are used to perform an out-of-sample non-parametric test. While this approach may be slightly more conservative than the direct test, it may have little practical utility because CDR is very good at finding optimal coefficients. Table 3.2 shows aggregated fixed-effects estimates obtained by running LME models over the CDR-convolved data $\mathbf{X}'$ from the both the synthetic and the human experimental datasets analyzed in this article. If the CDR-estimated coefficients are globally optimal given the impulse responses, then the LME estimates for all fixed effects should be 1. As shown, the LME-estimated fixed effects indeed cluster tightly around 1 for all three inference types explored here. This outcome suggests that CDR offers little room for improvement on its coefficient estimates, supporting the adequacy of the direct test for out-of-sample comparison.

In settings where the lack of optimality guarantee is unacceptable, CDR can still be a useful tool for data exploration, since it can estimate and visualize likely response profiles for predictors of interest (on exploratory data). Those estimates can then be used to construct effective and parsimonious FIR models. For example, suppose researchers have obtained a



(a) Number of iterations

(b) Ratio of iterations to minimum in same experimental condition

Figure 3.1: **Median training time by inference type.** Error bars show empirical 1st and 3rd quartiles.

CDR-estimated IRF for predictor A which decays to near 0 in 300ms, and that the mean interval between events in their experiment is 250ms. This finding suggests they might be able to capture most of the temporally diffuse response to A simply by inserting one additional spillover regressor for A into their linear model. This kind of information would otherwise be difficult to obtain without first fitting models with multiple different spillover configurations of the same predictors, a computationally-intensive procedure. CDR can also be used to overcome the inability of discrete-time models to estimate generalized effects of stimulus timing (§3.3) by estimating a convolved *rate* predictor which can be added as an effect to standard regression models. Results in §4.2 indicate that this may be particularly important for some psycholinguistic response variables, especially response times in self-paced reading.

3.7 Implementation

The experiments described in this article apply an open-source Python implementation of the CDR model described in §3.1 (`https://github.com/coryshain/cdr`), built using the Tensorflow (Abadi et al., 2015) and Edward (Tran et al., 2016) machine learning libraries.

Eq. 3.7 is implemented as a Tensorflow computation graph and optimized using either MLE or variational Bayesian inference.

3.7.1 Initialization

CDR fits are initially centered at the *null* model, i.e. a model in which there is no relationship between the predictors and response. In such a model, the intercept is the population mean, the variance is the population variance, all coefficients are 0 (predictors have no influence on the response), all random effects are 0 (there is no random deviation from the population means), and the IRF shapes are inconsequential (there is no response to the predictors). Thus, μ is initialized at the mean of the response, σ^2 is initialized at the variance of the response, and $\mathbf{m}$, $\mathbf{u}$, $\mathbf{U}$, and $\mathbf{V}_k$ are initialized at 0. Appropriate initializations for the fixed IRF parameters $\mathbf{v}_k$ are domain-specific, although kernels supported by this implementation come with overridable defaults as laid out in the documentation. The kernel initializations used in this study are described in §3.7.4.

3.7.2 Convergence

Because these CDR models use stochastic gradient optimization, it is necessary to define a convergence criterion by which the model parameters can be deemed (locally) optimal. Intuitively, the model has converged when it has ceased to improve with training. Diagnosing this condition automatically and model-independently is challenging because (1) the absolute rate of change in the loss over time depends on the scale of the data and the definition of the model and (2) the magnitude of the change in loss per iteration can be both noisy and non-decreasing due to stochastic optimization over a non-convex surface.

The present implementation addresses these challenges by retaining a history of the losses over a finite number of timesteps n and declaring convergence when the loss is uncorrelated with training time at a predetermined significance level α. Basing the convergence

criterion on correlation eliminates any influence of scale on either the loss or the representation of training time, instead grounding convergence in the strength of the linear relationship between these two quantities. To reduce the influence of noise and high-frequency autocorrelation on the test statistic, the implementation also permits a stride length m such that losses are only pushed to the history every m iterations, with intermediate values aggregated through a moving average.

In particular, given a vector of losses by iteration $\mathbf{l} \in \mathbb{R}^{\lceil n/m \rceil}$, and a vector $\mathbf{t} \in \mathbf{Z}^{\lceil n/m \rceil}$ of corresponding iteration numbers, let correlation of loss with training time be a test statistic:

$$\rho_t \stackrel{\text{def}}{=} \text{corr}(\mathbf{l}, \mathbf{t}) \tag{3.16}$$

Given a significance level α, the null hypothesis $H_{0t} \stackrel{\text{def}}{=} \rho_t = 0$ can be tested by computing probability p_{ρ_t} using a Student's t distribution with $\lceil n/m \rceil - 2$ degrees of freedom and checking that $p_{\rho_t} < \alpha$. When this test fails to reject H_{0t}, the losses are uncorrelated with training time (ρ_t is insignificantly different from 0 at level α).

This correlation-based hypothesis test defines binary criterion s:

$$s \stackrel{\text{def}}{=} \begin{cases} 1 & p_{\rho_t} > \alpha \\ 0 & \text{otherwise} \end{cases} \tag{3.17}$$

In other words, $s = 1$ if and only if H_{0t} is retained. To avoid premature convergence when by chance ρ_t happens to have small magnitude, the procedure additionally stores the history of values of s from each of the past $\lceil n/m \rceil$ strides in vector $\mathbf{s} \in \{0, 1\}^{\lceil n/m \rceil}$ and computes a proportion p_s of successful convergence checks:

$$p_s \stackrel{\text{def}}{=} \frac{m||\mathbf{s}||_1}{n} \tag{3.18}$$

Convergence is declared when $p_s > \alpha$ (i.e. when at least $\alpha \times 100\%$ of the previous $\lceil n/m \rceil$

convergence checks were positive). The overall stringency of this criterion increases with both n/m (because the test is higher-powered) and α (because the tests reject at a higher threshold and a larger proportion p_s is required). Experiments reported in this study use the following convergence hyperparameters: $n = 500$, $m = 1$, and $\alpha = 0.5$. In other words, convergence is declared when H_{0t} is retained at $\alpha = 0.5$ for at least 50% of the previous 500 training iterations.

3.7.3 Addressing Non-Normally Distributed Error: Log-Normal and Sinh-Arcsinh Transforms

Like ordinary linear regression, the CDR model defined in §3.1 assumes a Gaussian error distribution. However, it is often necessary to analyze data that violate this assumption. A common solution to non-normal errors is to apply a normalizing transform, such as a log transform or a power transform (Box and Cox, 1964). Transforms can complicate model interpretation by changing the linking function. For example, log transforming the response creates a log-linear rather than linear model on the convolved data, and model estimates thus describe a multiplicative rather than additive change in the response as a function of the predictors.

This study therefore also considers an alternative approach, made possible by the use of stochastic gradient optimization directly on the likelihood surface. Instead of defining the error distribution as Gaussian, it can be defined as a sinh-arcsinh transform on the Gaussian distribution.[8] The sinh-arcsinh transformed Gaussian is a generalization of the Gaussian distribution that additionally contains skewness and tailweight parameters $\epsilon \in \mathbb{R}$ and $\delta \in \mathbb{R}_+$

[8] The sinh-arcsinh transform on the standard normal distribution with skewness ϵ and tailweight δ yields probability density $f_{\epsilon,\delta}$:

$$f_{\epsilon,\delta}(x) \stackrel{\text{def}}{=} \left\{2\pi(1+x^2)\right\}^{-1/2} \delta C_{\epsilon,\delta}(x) \exp\left\{-S^2_{\epsilon,\delta}(x)/2\right\} \tag{3.19}$$

$$S_{\epsilon,\delta}(x) \stackrel{\text{def}}{=} \sinh\left\{\delta \sinh^{-1}(x) - \epsilon\right\} \tag{3.20}$$

$$C_{\epsilon,\delta}(x) \stackrel{\text{def}}{=} \left\{1 + S^2_{\epsilon,\delta}(x)\right\}^{1/2x} \tag{3.21}$$

In practice, location and scale can also be parameterized. For additional details, see Jones and Pewsey (2009).

(Jones and Pewsey, 2009). When $\epsilon = 0$ and $\delta = 1$, the distribution is Gaussian. When $\epsilon < 0$, the distribution has negative skew, and when $\epsilon > 0$, the distribution has positive skew. Tail thickness increases with δ. Both ϵ and δ are estimated from data, along with all other model parameters. The advantage of using sinh-arcsinh error over normalizing transforms is that it can flexibly adapt to asymmetrically distributed data without transforming it, thus preserving the original scale of the response as well as the additive interpretation of model estimates while also relaxing normality assumptions. Normalizing transforms and sinh-arcsinh error distributions are explored in §4 for the reading and fMRI experiments, where results show that sinh-arcsinh improves goodness of fit over Gaussian error across all model designs, supporting its adoption for CDR modeling. However, note that sinh-arcsinh error is not appropriate for settings in which estimates will ultimately be used in ways that assume normally-distributed error. For example, researchers may wish to evaluate CDR models with respect to squared error or percent variance explained. Such evaluations assume a Gaussian likelihood and are therefore not appropriate for asymmetric error distributions like sinh-arcsinh.

3.7.4 Experimental Procedure: General Model Parameters

In all experiments reported in this article, CDR models are fitted using the Nadam optimizer (Dozat, 2016)[9] with a constant learning rate of 0.001 and minibatches of size 1024. Nadam is a gradient-based optimizer that performs well for many non-convex optimization problems, and 0.001 is a widely-used default learning rate (see Tensorflow documentation: `https://www.tensorflow.org/`). Minibatch sizes in powers of two improve efficiency on standard compute architectures, and it was found during development that a size of 2^{10} roughly optimized iteration speed on available hardware. These parameters were not otherwise systematically tuned.

[9]The Adam optimizer (Kingma and Ba, 2014) with Nesterov momentum (Nesterov, 1983)

In order to compare estimation methods, all models are fitted using maximum likelihood (MLE), black-box variational inference with independent improper uniform priors and independent normal posteriors (BBVI-improper), and black-box variational inference with independent normal priors and posteriors (BBVI).[10] BBVI-improper and BBVI use the same black box estimation procedure. The only difference between them is that BBVI-improper lacks any penalties for divergence from a prior.

The analyses reported below in this article reveal dramatically faster convergence in BBVI inference mode than in MLE or BBVI-improper modes. This asymmetry is demonstrated by Figure 3.1, which shows (1) the mean number of training iterations (complete passes through the training data) across all experimental conditions and (2) the mean ratio of training iterations to the minimum number of iterations used by any model within the same experimental condition. As shown, BBVI requires on average less than half as many training iterations as MLE or BBVI-improper to reach convergence (Figure 3.1a) and nearly always requires fewer iterations than MLE or BBVI-improper in any given model configuration (Figure 3.1b). It is possible that the priors may discourage the model from following tiny gradients, thereby accelerating convergence. Because of this computational advantage, BBVI is likely to be the most useful estimation method, and it is therefore used in all model comparisons reported below (both for null hypothesis significance testing and for comparison of CDR to baselines), even when other estimation techniques achieve better error.

In the BBVI setting, reasonable prior variances for intercepts, coefficients, and error depend on the scale of the response. For this reason, this CDR implementation uses the variance of the response in the training set as the prior variance for these parameters. Reasonable prior variances for the impulse response parameters are independent of the scale of the response. For simplicity, these experiments use a variance of 1. In order

[10] Although the software implementation includes experimental support for multivariate normal priors and variational posteriors, this results in a quadratic increase in the number of parameters, and it did not to provide a clear performance benefit.

to improve both convergence and general applicability of the hyperparameters used in this study to other kinds of data, the response variable is implicitly standardized (z-transformed) prior to fitting. This transform is inverted in order to compute predictions and likelihoods. As a result, all BBVI priors used in this study have unit variance and mean equal to the initialization value used in MLE inference (described below for IRF parameters and in §3.7.1 for all other parameters).

While it is in principle sensible in a BBVI setting to use the prior as the initial value for the variational posterior, in practice it turns out that doing so can lead to training divergence in complex models due to early initial sampling of poor solutions from an excessively wide distribution. The variational posterior is therefore initialized with a tighter standard deviation (one one-hundredth of the standard deviation of the prior) in all BBVI-improper and BBVI models. Note that this is merely an initialization technique — the model can adjust the width of the variational posterior throughout training as required by the data.

It is standard practice in mixed-effects modeling to penalize the random effects (Bates et al., 2015). All psycholinguistic analyses with random effects reported below follow Bates et al. (2015) by penalizing all random effects (random intercepts, coefficients, and impulse response parameters) using L2 regularization with regularization level $\lambda = 1.0$. In the BBVI setting, a similar kind of constraint is implemented by imposing a tighter prior on the random effects than on the fixed effects (standard deviation of 0.1 and 1.0, respectively).

The BBVI priors used in this study were not tuned in any way, and different priors may be motivated for different datasets based on foreknowledge of the experimental domain. The size of the hyperparameter space explored in this study is so large that additional systematic exploration of the influence of different prior settings is computationally prohibitive and left to future work. That said, the present results suggest that these choices of priors do not strongly constrain the solution space, since BBVI inference finds qualitatively similar responses to the BBVI-improper and MLE inferences, which have no priors (see Chapter 4).

Models reported here use some combination of *exponential*, *normal*, *shifted gamma*, and

pseudo non-parametric linear combination of Gaussians (LCG) impulse response kernels. The *exponential*, *normal*, and *shifted gamma* kernels are the probability density functions associated with each type of probability distribution, with the addition of normalization terms to ensure that IRFs integrate to 1 over the positive real line (§3.2). For these parametric kernels, the normalization term is the survival function of the distribution (complement of the cumulative density function) at $x = 0$.

The probability density functions of the exponential, normal, and shifted gamma distributions are respectively:

$$f_{\text{Exp}}(x;\beta) = -\beta e^{-\beta x} \tag{3.22}$$

$$f_{\text{Normal}}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{\sigma^2}} \tag{3.23}$$

$$f_{\text{ShiftedGamma}}(x;\alpha,\beta,\delta) = \frac{\beta^\alpha (x-\delta)^{\alpha-1} e^{-\beta(x-\delta)}}{\Gamma(\alpha)} \tag{3.24}$$

The exponential distribution integrates to 1 over the positive real line and requires no further normalization. For the *normal* and *shifted gamma* kernels, normalization requires the corresponding survival functions:

$$S_{\text{Normal}}(x;\mu,\sigma^2) = 1 - \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \tag{3.25}$$

$$S_{\text{ShiftedGamma}}(x;\alpha,\beta,\delta) = 1 - \frac{1}{\Gamma(\alpha)}\gamma\left(\alpha,\beta\left(x-\delta\right)\right) \tag{3.26}$$

where

$$\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} dt \tag{3.27}$$

$$\gamma(s,x) = \int_0^x t^{s-1} e^{-t} dt \tag{3.28}$$

Thus, the *exponential*, *normal*, and *shifted gamma* IRF kernels are defined respectively as:

$$\text{Exp}(x;\beta) = f_{\text{Exp}}(x;\beta) \tag{3.29}$$

$$\text{Normal}(x;\mu,\sigma^2) = \frac{f_{\text{Normal}}(x;\mu,\sigma^2)}{S_{\text{Normal}}(0;\mu,\sigma^2)} \tag{3.30}$$

$$\text{ShiftedGamma}(x;\alpha,\beta,\delta) = \frac{f_{\text{ShiftedGamma}}(x;\alpha,\beta,\delta)}{S_{\text{ShiftedGamma}}(0;\alpha,\beta,\delta)} \tag{3.31}$$

These kernels encode increasingly flexible assumptions about the response shape. In particular, the *exponential* kernel assumes that the influence of a predictor is strongest immediately (at $t = 0$) and decreases monotonically over time. The only question is how quickly, which is determined by the rate parameter β. The *normal* kernel relaxes this monotonicity assumption. Like the *exponential* kernel, it can fit monotonically decreasing IRFs (by finding location $\mu \leq 0$), but it can also fit late-peaking (rising then falling) IRFs (by finding $\mu > 0$), allowing the peak response to occur at some delay from the stimulus (e.g. Smith and Levy, 2013, who found a peak surprisal response on the following word in a self-paced reading experiment). In the latter case, the *normal* kernel assumes a symmetric rise/fall pattern. The *shifted gamma* kernel additionally relaxes the symmetry assumption. It can find approximately symmetric late-peaking IRFs, but it can also find IRFs that e.g. rise quickly and then decay slowly. The *shifted gamma* kernel approximately subsumes the solution space of the *normal* kernel, which approximately subsumes the solution space of the *exponential* kernel.

As mentioned above, this study also considers a quasi-non-parametric comparison kernel consisting of a linear combination of Gaussians (LCG; Goshtasby and O'Neill, 1994; Gimel'farb et al., 2004). The LCG kernel contains n component *normal* kernels, each with location, scale, and amplitude parameters. Having a more flexible control permits evaluation of the extent of bias introduced by parametric kernels when applied to real world data

in which the ground-truth IRF is unknown (§4.2, §4.3). This LCG kernel is defined as:

$$\mathrm{LCG}(x; \mu_{1,\ldots,n}, \sigma^2_{1,\ldots,n}, \beta_{1,\ldots,n}) = \frac{\sum_{i=1}^{n} f_{\mathrm{Normal}}(x; \mu_i, \sigma_i^2) \cdot \beta_i}{\sum_{i=1}^{n} S_{\mathrm{Normal}}(x; \mu_i, \sigma_i^2) \cdot \beta_i} \tag{3.32}$$

The LCG kernel is highly flexible (see §4.1, §4.2, and §4.3), yet computationally efficient, since it has an analytical normalization constant (the sum of the component survival functions). More expressive kernels are possible, such as spline functions or kernel smoothing, but these do not have analytical integrals and therefore require numerical integration at each optimization step in order to properly normalize them. In initial experiments, this was found to greatly slow training without clear improvement to final fit, and therefore LCG is used for the non-parametric comparison. In these experiments, all LCG kernels have 10 components.

Beyond any constraints imposed by the definitions of the density functions above, the *shifted gamma* kernel additionally requires that $\alpha > 1$ and $\delta < 0$.[11] Constraints on bounded parameters are enforced using the softplus bijection:

$$\mathrm{softplus}(x) = \log(e^x + 1) \tag{3.33}$$

All models use the same default initializations for these kernels. For *exponential* kernels, $\beta = 1$. For *shifted gamma* kernels, $\alpha = 2$, $\beta = 1$, and $\delta = -1$. For *normal* kernels, $\mu = 0$ and $\sigma^2 = 1$. For the LCG kernels, $\mu_i = 0$, $\sigma_i = i$ and $\alpha_i = 1$ if $i = 0$, $\alpha_i = 0$ otherwise for $i \in \{1, \ldots, 10\}$. This initializes the kernel with a single non-zero component, and with incrementally wider initial components to allow the model to find later or earlier peaks. Initializing all components with $\beta = 0$ leads to numerical degeneracies because of the requirement that the LCG function integrate to 1.

[11] $\alpha > 1$ helps deconfound the shape and shift parameters by ensuring that the response underlyingly has a rising-falling profile, in which case strictly falling responses can only be found by shifting the peak to the left of 0. $\delta < 0$ ensures that the instantaneous response (response at $x = 0$) is well defined.

Prediction from the network uses an exponential moving average of parameter iterates with a decay rate of 0.999. BBVI models are evaluated using *maximum a posteriori* estimates obtained by setting all parameters to their posterior means. This procedure is motivated by the law of large numbers: because all parameters have independent normal distributions in the variational posterior, samples from that posterior converge in probability to the posterior mean. For computational reasons, predictor histories are truncated at 256 timesteps (words) into the past.

Chapter 4

CDR Synthetic, Reading, and fMRI Experiments

4.1 Synthetic Experiments

Before applying CDR to human-generated time series, we first empirically validate it through simulations using synthetic data with known ground-truth impulse responses, since this permits direct comparison of the CDR estimates to the true data-generating model. In the process, we systematically explore the sensitivity of CDR estimates to several potential sources of influence that are likely to arise in practice: noise in the response variable, non-uniform time intervals between events, multicollinearity, and misspecification of the impulse response kernel. As shown below, CDR recovers the data-generating model in a wide range of settings and is robust to adverse training conditions like multicollinearity and IRF misspecification.

4.1.1 Experimental Design

Several design details are common to all simulations reported here. All datasets contain twenty randomly generated predictors. In all simulations except the multicollinearity manipulations, these twenty predictors are sampled from independent standard normal distributions. In all datasets, ground-truth coefficients for each predictor are sampled from a uniform distribution $\mathcal{U}(-10, 10)$, and a ground-truth impulse response is created for each predictor by randomly sampling parameters for a given impulse response kernel. The response is then generated by convolving each predictor with its assigned impulse response, sampling the convolved signal at predetermined query points (timestamps), scaling the

sampled signal by the ground-truth coefficients, and summing the scaled sample across predictors in order to generate a response vector. In all simulations except the noise manipulations, Gaussian noise with standard deviation 10 is added to the generated response vector. In all simulations except the time manipulations, time intervals between stimulus events and response samples are asynchronous and sampled from an exponential distribution with mean 100ms. For simplicity, all synthetic datasets consist of a single timeseries containing 10,000 response samples. For all simulations, we provide (1) qualitative assessments of IRF identification by visually comparing true and estimated responses and (2) quantitative assessments of IRF identification by computing root mean squared deviation (RMSD) of the estimated response from the true response over 1,000 timepoints spaced equidistantly on the 95^{th} percentile of temporal offsets seen in training.

4.1.2 Simulation A: Noise

Simulation A explores the sensitivity of CDR estimates to noise in the response variable. A single set of 20 independent predictors is sampled as described above, and a single set of *shifted gamma* impulse responses is sampled from the following distributions: $\alpha \sim \mathcal{U}(1, 6)$, $\beta \sim \mathcal{U}(0, 5)$, and $\delta \sim \mathcal{U}(-1, 0)$. Gaussian noise with standard deviation 0 (noise free), 1, 10, and 100 is then injected into the convolved response, and CDR models with *shifted gamma* IRF kernels are fitted separately to each level of noise. The true simulated response has a signal power (mean squared value) of 1907, and thus the signal to noise ratios of the synthetic datasets are respectively ∞, 1907, 19.07, and 0.1907.

Qualitative results are presented in Figure 4.1. As shown, estimates closely match the true model in all conditions, indicating that CDR estimates are robust to noise in the response data. The BBVI-improper and BBVI inferences differ substantially in their quantification of uncertainty. BBVI-improper is so confident in its estimates, even under high noise, that the 95% credible intervals are not visible in the plots. BBVI, by contrast, has

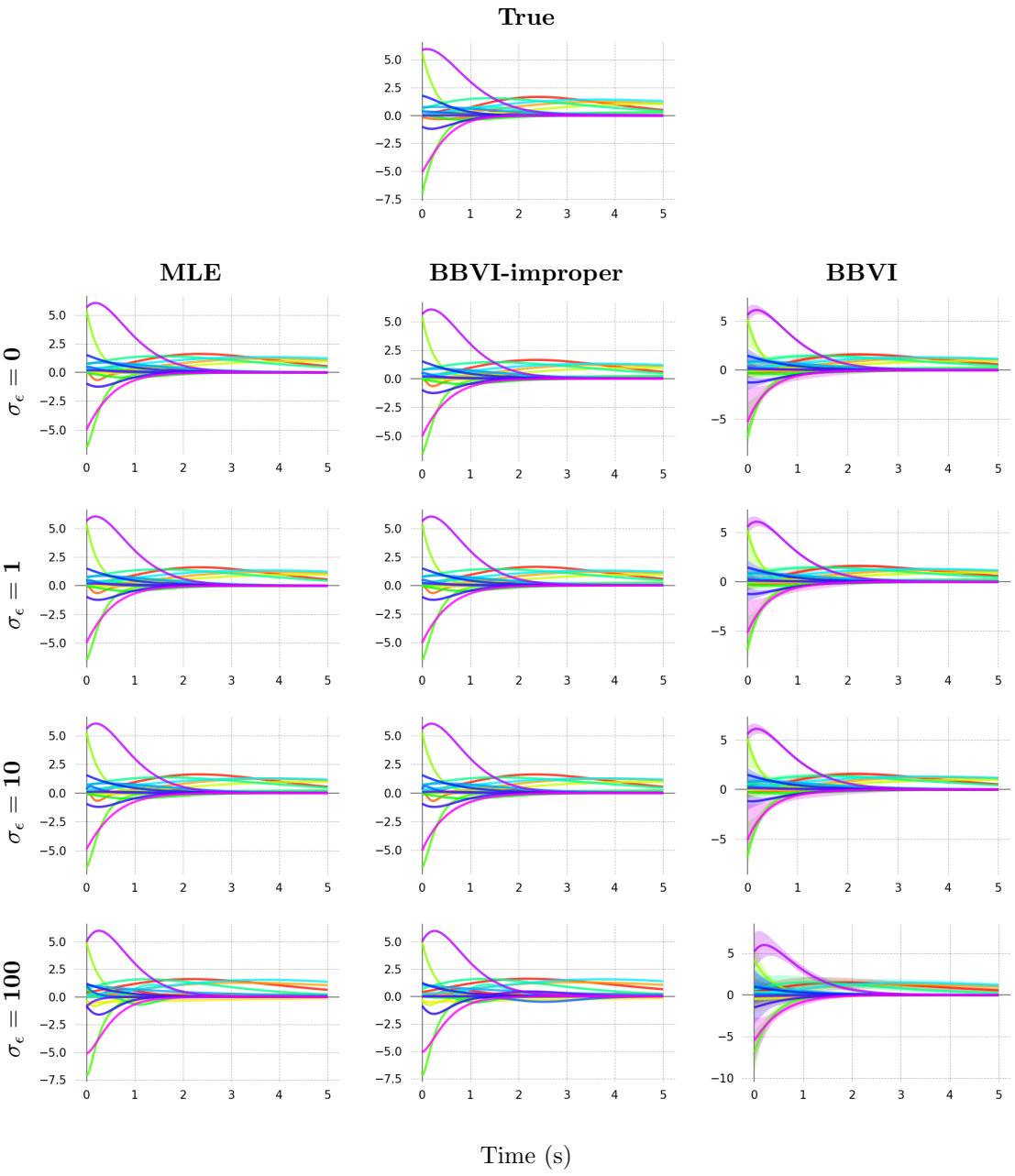


Figure 4.1: **Simulation A:** *Noise.* True synthetic model vs. CDR-estimated models with increasingly large standard deviation σ_ϵ of the Gaussian noise distribution. BBVI and BBVI-improper are plotted with 95% credible intervals.

wider credible intervals in general, with the widest intervals in the highest noise setting. The use of proper priors therefore appears to support more plausible estimates of uncertainty.

Quantitative results are presented in Figure 4.4a. As shown, IRF identification is mostly insensitive to gradations of low-level noise but predictably degrades at the highest level of noise. True IRF recovery levels are highly similar in all model types (MLE, BBVI-improper, BBVI).

4.1.3 Simulation B: Time

Simulation B explores the sensitivity of CDR estimates to different kinds of time intervals between predictor and response observations. We consider three manipulations: (1) fixed vs. variable spacing, (2) long vs. short intervals, and (3) synchronous vs. asynchronous measures of predictors and response. Six conditions are constructed to evaluate these influences, representing roughly increasing levels of complexity in temporal structure:

- **Fixed synchronous short (FSS):** Predictors and response are aligned and placed at fixed 100ms intervals.
- **Fixed synchronous long (FSL):** Predictors and response are aligned and placed at fixed 500ms intervals.
- **Random synchronous short (RSS):** Predictors and response are temporally aligned and intervals are sampled from an exponential distribution with mean 100ms.
- **Random synchronous long (RSL):** Predictors and response are temporally aligned and intervals are sampled from an exponential distribution with mean 500ms.
- **Random asynchronous short (RAS):** Predictors and response are not temporally aligned: intervals are sampled independently for predictors on the one hand and response on the other from an exponential distribution with mean 100ms.

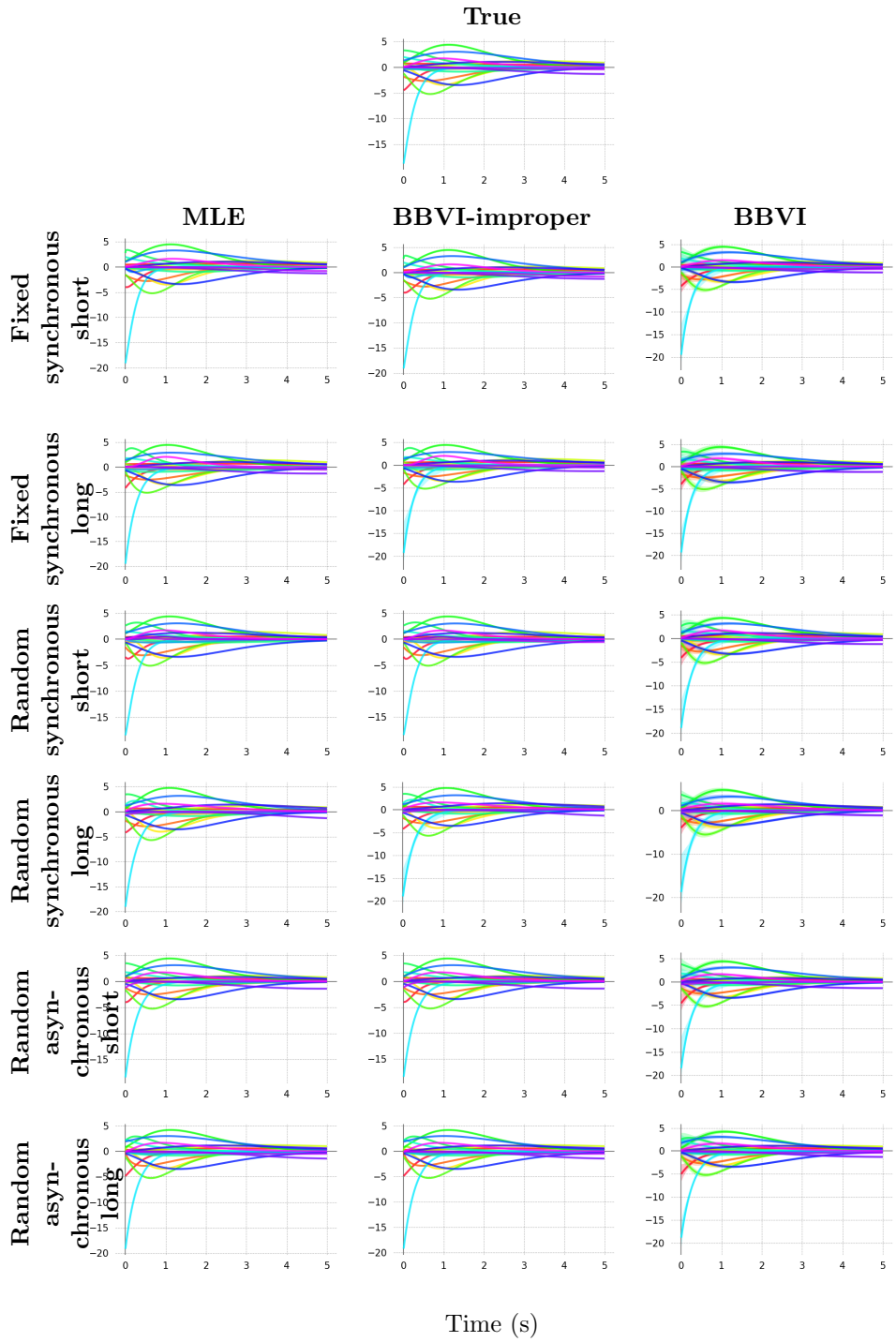


Figure 4.2: **Simulation B:** *Time.* True synthetic model vs. CDR-estimated models with varying types of time interval.

- **Random asynchronous long (RAL):** Predictors and response are not temporally aligned: intervals are sampled independently for predictors on the one hand and response on the other from an exponential distribution with mean 500ms.

The data generating model is constructed using the same procedure as in Simulation A.

Qualitative results are presented in Figure 4.2. As in Simulation A, estimated models closely match the ground truth in all conditions, indicating that CDR estimates are robust to variation in the temporal spacing and synchrony of events. Quantitative results in Figure 4.4b bear out this impression, showing that IRF recovery is no worse for random than fixed intervals or for asynchronous than synchronous events. Figure 4.4b shows a clear advantage for long intervals over short intervals in this simulation. Given the finite-length history window (256 events) used in these experiments, interval length instantiates a trade-off between resolution (improved by shorter intervals) and coverage (improved by longer intervals). Thus, in this simulation, the improved coverage of the time dimension afforded by longer intervals overcomes any cost from loss of temporal resolution. This trade-off is especially pronounced in the random synchronous short condition using MLE and BBVI-improper estimation modes, where IRF recovery performance is dramatically worse (Figure 4.4b). As shown in Figure 4.2, this is largely driven by the failure of these models to discover a late-peaking negative IRF (negative response in purple toward the right edge of the x axis). This failure does not occur in the longer interval condition, indicating that it is indeed driven by lack of temporal coverage. These results highlight the importance of using a history window that is long enough to capture the underlying temporal dynamics.

4.1.4 Simulation C: Multicollinearity

Simulation C explores the sensitivity of CDR estimates to multicollinearity in the predictors. To manipulate multicolinearity in the predictors, predictor streams were drawn from multivariate normal distributions in which the variance-covariance matrix had a diagonal of 1 and all off-diagonal elements were set to the desired level of correlation. For example,

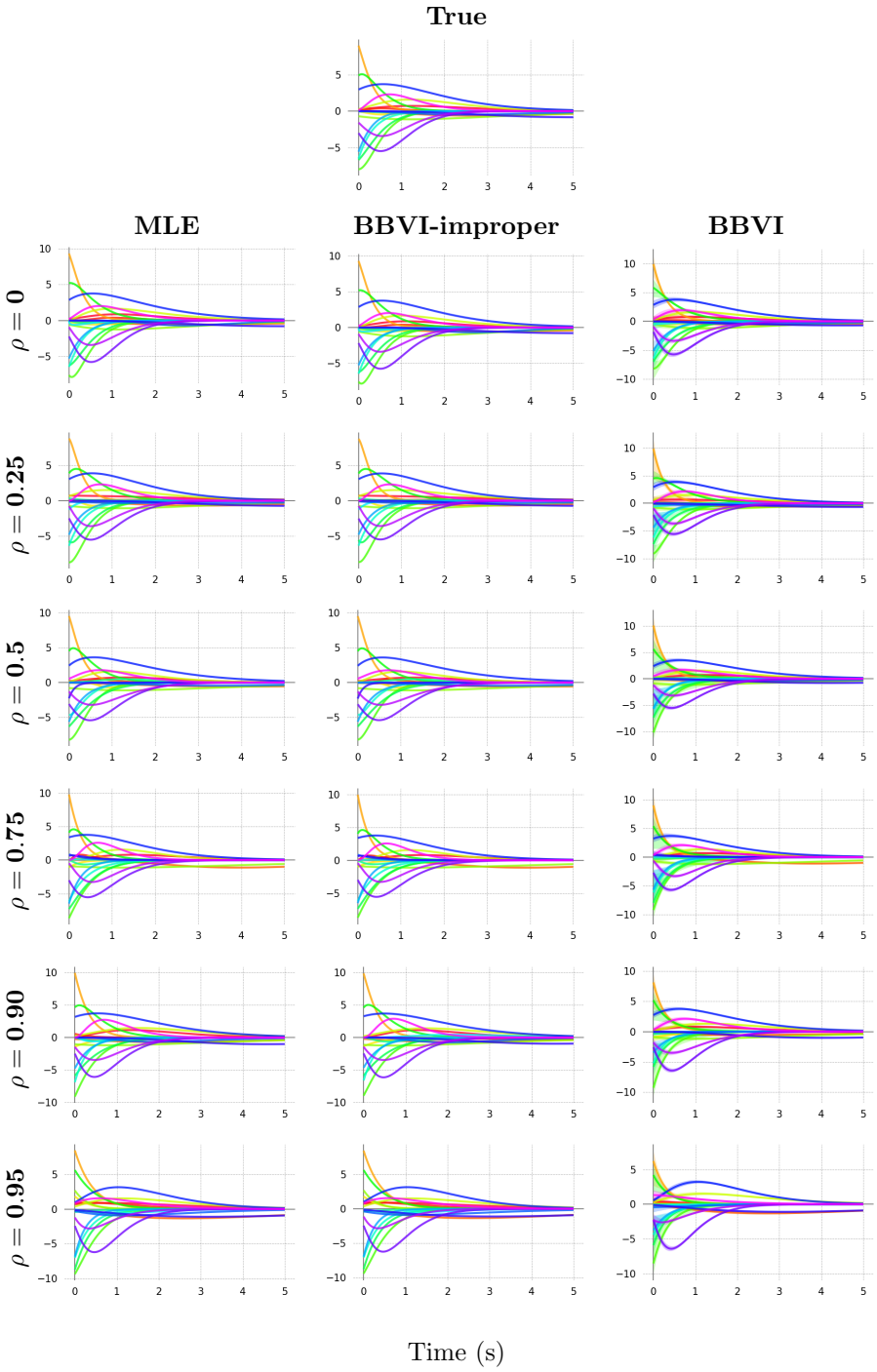
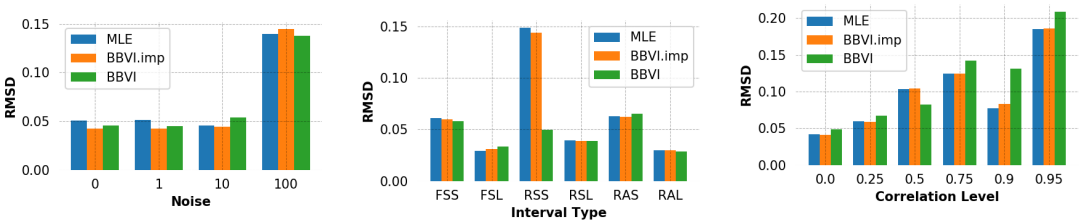


Figure 4.3: **Simulation C:** *Multicollinearity.* True synthetic model vs. CDR-estimated models with increasingly multicollinear predictors.



(a) **Simulation A**. *Noise* (b) **Simulation B:** *Time* (c) **Simulation C:** *Multicollinearity*

Figure 4.4: Root mean squared deviation (RMSD) of estimated models from ground truth in Simulations A, B, and C.

predictors with correlation level $\rho = 0.5$ were drawn using the following variance-covariance matrix:

$$\begin{bmatrix} 1 & 0.5 & 0.5 & \dots & 0.5 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 & \dots & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & \dots & 0.5 & 0.5 & 0.5 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & 0.5 & \dots & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & \dots & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & \dots & 0.5 & 0.5 & 1 \end{bmatrix}$$

Six sets of predictors were generated in this way, one for each of $\rho = 0$ (uncorrelated predictors), $\rho = 0.25$, $\rho = 0.5$, $\rho = 0.75$, $\rho = 0.9$, and $\rho = 0.95$. The data-generating model was constructed using the same procedure as in Simulations A and B.

Qualitative results are presented in Figure 4.3. As shown, CDR recovers much of the underlying model structure across conditions, even in a highly adverse setting where every predictor is correlated with every other predictor at a level of 0.95. Quantitative results (Figure 4.4c) show a roughly linear increase in RMSD with multicollinearity, but with reasonably good IRF recovery under highly multicollinear data; even at $\rho = 0.9$, RMSDs are only about twice those when $\rho = 0$.

	Parameters		
Kernel	SimD	NatStor	Dundee
Exponential	42	1970	166
Normal	62	2686	232
Gaussian	82	3402	298
LCG	662	21845	2080

Table 4.1: Number of parameters by kernel family and corpus (Simulation D, Natural Stories, and Dundee, respectively). Differences between corpora are driven by the random effects, since Natural Stories contains many more participants (181) than Dundee (10), while Simulation D contains no random effects. Note that BBVI and BBVI-improper double these figures by additionally fitting variances for each parameter in the variational posterior.

4.1.5 Simulation D: IRF Misspecification

Simulation D explores the ability of CDR to find reasonable IRF estimates in the presence of mismatch between the true and modeled IRF kernels. To this end, three data-generating models are constructed with different underlying response shapes: *exponential* (E), *normal* (N), and *shifted gamma* (G). CDR models of each kernel type are fitted to each of these datasets, such that two out of three modeled kernels for each dataset are not matched to the underlying model (e.g. fitting *exponential* IRFs to the output of a *normal* data-generating model). The target outcome under kernel mismatch is to discover the best available estimate given the solution space defined by the modeled kernel. This analysis also explores the use of non-parametric LCG kernels (described in Section 3.7.4), since their solution space approximately subsumes all ground truth models used in this simulation. As shown in Table 4.1, the LCG kernel is much more heavily parameterized than the others.

Note that these different kernels have asymmetrical patterns of compatibility. For example, the solution space of the *shifted gamma* kernel contains the *exponential* kernel, since the exponential distribution is a special case of the shifted gamma distribution (i.e. when $\alpha = 1$ and $\delta = 0$).[1] The converse does not hold: *shifted gamma* contains late-peaking responses that fall outside the strictly monotonic solution space of the *exponential* kernel.

[1]Because these values lie at the parameter bounds for the *shifted gamma* kernel used here, the model cannot exactly reach them. However, it can come arbitrarily close within 32-bit floating point precision.

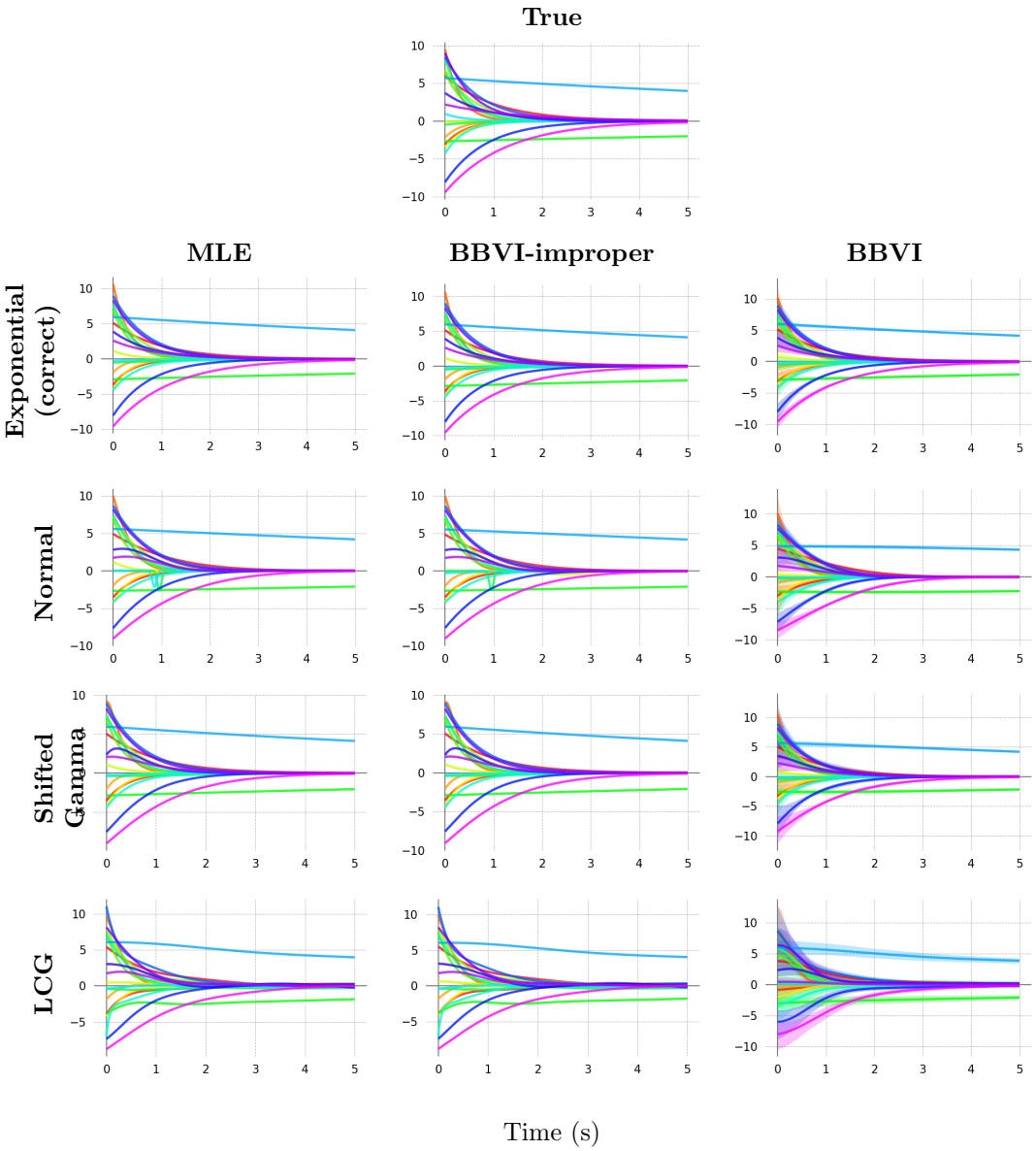


Figure 4.5: **Simulation D:** *Misspecification (exponential ground truth).* True exponential model vs. CDR-estimated models using various IRF kernels.

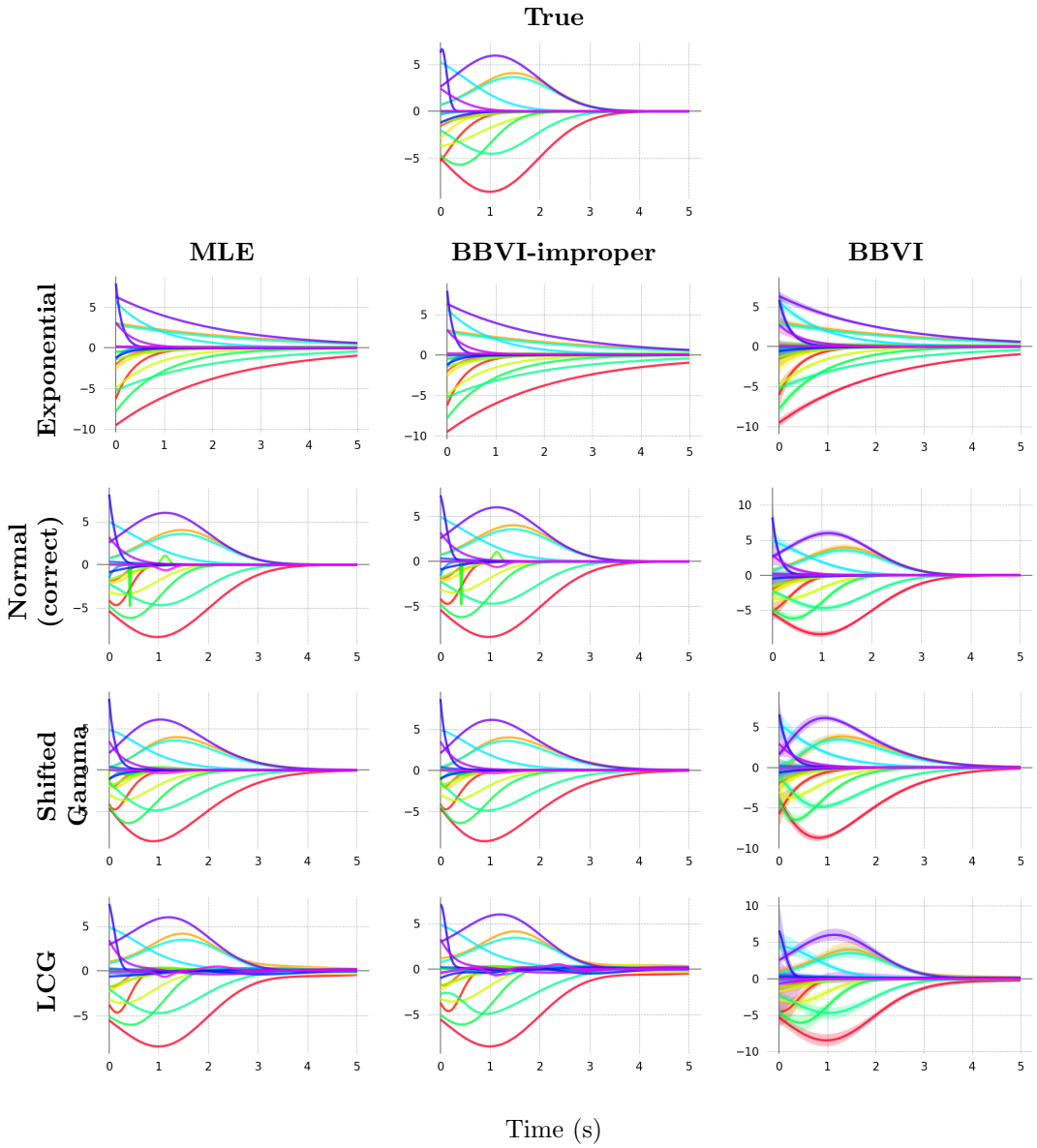


Figure 4.6: **Simulation D:** *Misspecification with normal ground truth.* True normal model vs. CDR-estimated models using various IRF kernels.

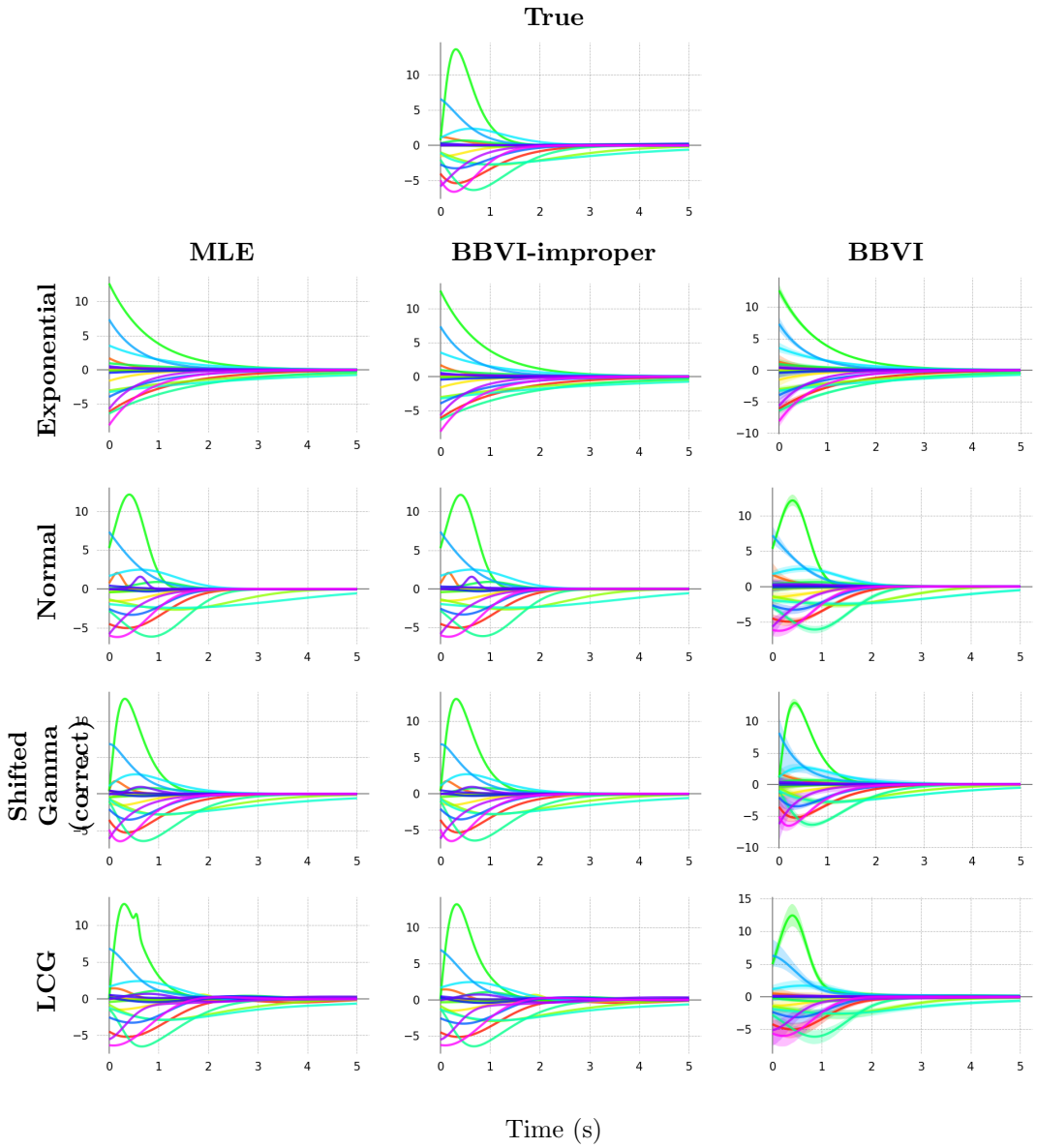


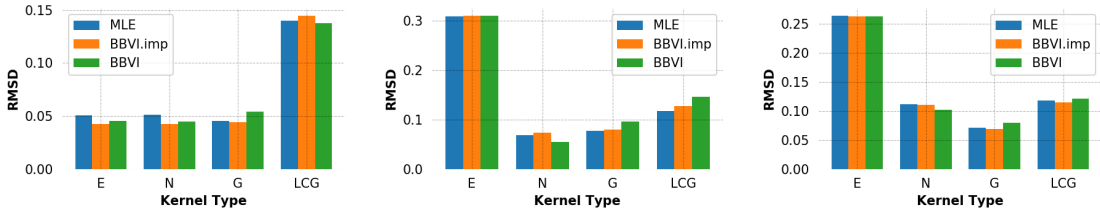
Figure 4.7: **Simulation D:** *Misspecification (shifted gamma ground truth).* True shifted gamma model vs. CDR-estimated models using various IRF kernels.

The *normal* kernel is more flexible than the *exponential* kernel and may therefore be able to better approximate *shifted gamma* responses, but it is additionally constrained by symmetry about the mean.

The predictors, coefficients, and *shifted gamma* data-generating model are constructed using the same procedure as in Simulations A and C. The impulse responses for the *exponential* data-generating model are constructed by sampling $\beta \sim \mathcal{U}(0, 5)$. The impulse responses for the *normal* data-generating model are constructed by sampling $\mu \sim \mathcal{U}(-2, 2)$, $\sigma^2 \sim \mathcal{U}(0, 2)$.

Qualitative results are presented in Figures 4.5, 4.6, and 4.7. As in previous simulations, CDR estimates closely match the ground truth in all conditions, indicating that CDR successfully identifies the model within the constraints imposed by its kernel, even under mismatch between the true and modeled kernels. This finding is reassuring, since it suggests that the particular parameterization used may not be of great importance, as long as the kernel adequately covers the space of plausible solutions. As expected, models with *exponential* kernels struggle to identify underlying *normal* or *shifted gamma* responses with a late-peaking profile, since late peaks fall outside the solution space of the *exponential* kernel. Within those constraints, the *exponential* model estimates are reasonable. Nonetheless, in an application to real data, the *exponential* kernel is likely a poor choice if late-peaking responses are plausible. In all conditions, the LCG IRF also reliably identifies the true model. Unsurprisingly, LCG estimates exhibit noisier temporal dynamics than their lower-dimensional counterparts, which are smooth by definition. This could potentially be addressed by regularization similar to the "roughness penalty" used in existing spline regression techniques (Wood, 2006). This possibility is left to future research.

Quantitative results are presented in Figure 4.8, which largely confirms patterns suggested by the qualitative evaluation. First, as expected, the matched kernel (e.g. *normal* fitted to *normal*) generally has the lowest RMSD across conditions, but RMSDs from *normal* and *shifted gamma* kernels are also quite competitive under mismatch. The *exponential*



(a) Exponential (b) Normal (c) Shifted gamma

Figure 4.8: **Simulation D:** *Misspecification.* Root mean squared deviation (RMSD) of estimated exponential (E), normal (N), shifted gamma (G), and non-parametric summed Gaussian (NPG) models from ground truths.

kernel suffers predictably under mismatch, because late-peaking responses are far outside its solution space (discussed above). LCG kernels show consistently larger RMSD than the matched model, likely due to roughness in the high-dimensional estimates.

4.2 Reading Experiments

This section applies CDR to discover effect timecourses in human reading behavior from two large naturalistic datasets. As argued in §2.1, analyzing naturalistic reading is a key application area of interest, both because (1) diffusion of effects may be especially pronounced in the naturalistic reading paradigm and (2) existing tools like FIR/spillover models have a number of shortcomings when applied to this domain (see §2.2). Our analyses closely follow those reported in Shain (2019) but focus primarily on validation of CDR rather than on an empirical claim about human sentence processing. To this end, much like in the simulation studies reported in §4.1, this study applies multiple estimation techniques and impulse response kernels to each corpus.

The primary interest of CDR as an explanatory model is that it yields detailed estimates of diffuse temporal structure that existing discrete-time regression techniques cannot provide. However, unlike the simulations above, the data-generating model for human reading responses is unknown, and its temporal structure is currently poorly understood. Thus,

direct comparison to the ground truth is not available for model validation on human data; indeed, a potential benefit of CDR modeling is the ability to shed light on this important question. For this benefit to be realized, it is first necessary to establish that CDR provides a "good" model of human responses according to some standard; otherwise, its estimates should not be trusted. Such a standard can be constructed using the predictive performance of established statistical methods. If CDR models generalize less well to unseen data than standard models, then their estimates of temporal structure should be treated with skepticism. However, if CDR predictions perform competitively with those of standard models, then this indicates that the model has tapped into generalizable properties of the response. This study compares the predictive performance of CDR to that of linear mixed effects (LME) and generalized additive (GAM) models, both of which have been used extensively in psycholinguistics. Nonetheless, the primary advantage of CDR lies in its ability to estimate continuous diffusion over time, with performance comparisons serving as a sanity check. Results show that CDR predictive performance is competitive with that of all baselines in each dataset and superior to all baselines overall, supporting the reliability of its estimates of temporal structure. In addition, CDR performance is stable across a range of estimation methods and IRF kernels.

4.2.1 Data

This reading study uses the Natural Stories (Futrell et al., 2018) and Dundee (Kennedy et al., 2003) datasets. Natural Stories is a self-paced reading (SPR) corpus consisting of context-rich narratives read by 181 subjects. Stimuli are designed to resemble naturally-occurring texts while increasing the representation of rare words and syntactic constructions. Subjects paged through the stories on a computer screen, pressing a button to reveal the next word. The amount of time spent on each word was recorded as the response variable. The corpus contains a total 1,013,290 events (where one event is a single subject viewing a

single word token).[2]

Dundee is an eye-tracking corpus containing newspaper editorials read by 10 subjects. The corpus contains a total of 259,957 events (where one event is a single subject fixating a single word token for the first time from the left).[3]

In both reading experiments, data are partitioned into training (50%), exploratory (25%) and test (25%) sets. The partitioning strategy attempts to respect the non-independence of words within the same sentence, using modular arithmetic to cycle sentence IDs e into different bins of the partition with a different phase for each subject u: $\text{partition}(e, u) = (e+u) \bmod 4$, assigning outputs 0 and 1 to the training set, 2 to the exploratory set, and 3 to the test set. Outlier filtering is also performed, largely following the procedures described in Shain and Schuler (2018).[4] Because CDR's convolution operation is only correct if applied to all preceding events within the history window, partitioning and filtering are applied only to the response, retaining all events in the predictor matrix.[5]

[2]This figure differs from the published count in (Futrell et al., 2018) because events are not filtered from the stimulus sequence, since they are needed for accurate deconvolution.

[3]A limitation of the Dundee corpus is the number of participants (10). Although each of these participants is quite densely sampled and should therefore be able to be reliably modeled, the small number of participants may limit the degree to which results based on Dundee can be expected to generalize to the population as whole. Nonetheless, the purpose of the present study is not to test hypothesized effects in human sentence processing, but rather to evaluate the empirical properties of a new modeling approach (CDR). Dundee is one of the most extensively analyzed naturalistic eye-tracking corpora in psycholinguistics (e.g. Demberg and Keller, 2008; Frank and Bod, 2011; Fossum and Levy, 2012; Smith and Levy, 2013; van Schijndel and Schuler, 2015; Goodkind and Bicknell, 2018, *inter alia*) and therefore serves as a "standard" dataset for initial evaluation of CDR for eye-tracking. CDR analysis of eye-tracking corpora with larger numbers of participants (e.g. Cop et al., 2017) is left to future work.

[4] For Natural Stories, following Shain et al. (2016), items were excluded if they have fixations shorter than 100ms or longer than 3000ms, if they start or end a sentence, or if subjects missed 4 or more subsequent comprehension questions. Analyses additionally removed any subjects with fewer than 100 datapoints after application of the other filters, both because such subjects are likely uncooperative (missing excessive comprehension questions or paging too rapidly through the text) and because their data are likely insufficient to support estimation of random effects. For Dundee, following van Schijndel and Schuler (2015), unfixated items were excluded as well as (1) items following saccades longer than 4 words and (2) starts and ends of sentences, screens, documents, and lines. In addition, following common practice in psycholinguistics, items were removed if their duration included a blink (e.g. Schotter et al., 2018). Most of these outlier filters are designed to minimize the influence of boundary effects like implicit prosody (Breen, 2014). Differences across corpora in exclusion criteria are driven by a combination of (1) differences in precedent established by studies that use these corpora (see citations), (2) differences in modality, since e.g. unfixated items and long saccades are only relevant to eye-tracking, and (3) differences in source data, since e.g. only Dundee provides information about screen, document, and line boundaries.

[5]In these experiments, an entire document is treated as a time series, with the result that words can continue to influence the response across sentence boundaries.

4.2.2 Experimental Setup

The purpose of Experiment A is to evaluate CDR as a statistical model of human reading. To this end, all analyses share the following design features.

Response Variable

In all experiments, the response variable of interest is reading latency, under the eye-mind assumption (Just and Carpenter, 1980) that longer latencies index greater processing difficulty. The definition of latency varies by experimental modality. For self-paced reading data (Natural Stories), reading latency is defined as reaction time — the interval between button presses.

For eye-tracking data (Dundee), a number of latency measures are possible (Rayner, 1998), and this study considers three such measures. *Scan path duration* is defined as the time elapsed between entering a word region (from either direction) and entering a different word region (in either direction). Under the assumption that word features (e.g. frequency and surprisal, defined below) do not accumulate in influence from consecutive saccades to the same word, the duration of all consecutive saccades to the same word region are summed into a single measure. *First pass duration* is defined as the time elapsed between entering a word region from the left and entering a different word region to its left or right. *Go-past duration* is defined as the time elapsed between entering a word region from the left and entering a word region to its right (including all intervening regressive fixations). These measures differ in their treatment of the temporal order of fixations relative to the spatial order of words on a page, which, due to regressive (backward) eye movements, are not strictly aligned in free reading. Scan path durations follow the temporal sequence of eye movements rather than the spatial sequence of words, whereas first pass and go-past durations either ignore (first pass) or aggregate (go-past) the durations of regressive eye

movements, thus bringing the fixation order and word order into alignment.[6]

In all cases, latency is measured in milliseconds. Because of the non-normal distribution of reading times in psycholinguistic experiments (e.g. Frank et al., 2013), analyses also explore log-transformed variants of reading time durations (Smith and Levy, 2011, following e.g.), as well as the use of a non-normal error distribution (sinh-arcsinh) for estimating IRFs from reading data.

Predictor Variables

Models use the following predictors: *sentence position* (index of word in sentence), *document position* (index of word in document), incoming *saccade length* (in words, eye-tracking only), *previous was fixated* (indicator for whether the preceding word was fixated, eye-tracking only), *word length* (in characters), *unigram surprisal*,[7] and *5-gram surprisal*.[8] *Unigram surprisal* and *5-gram surprisal* are computed by the KenLM toolkit (Heafield et al., 2013) trained on Gigaword 3 (Graff et al., 2007). Examples of studies using some or all of these predictors include Demberg and Keller (2008); Frank and Bod (2011); Smith and Levy (2013) and Baayen et al. (2018). Models also include a deconvolutional intercept, referred to as *rate*, which is designed to capture any generalized response to stimuli, independently of their properties (§3.3). *Rate* is excluded from all non-CDR baseline models reported below because it is identical to the intercept, and thus these baselines are unable to identify *rate* effects.

[6] The question of how to define events and measures in the reading record (scan path, first pass, go-past, regression probability, etc.) is orthogonal to the question of how to analyze the data (LME, GAM, CDR): as shown in the Dundee results presented here, comparable discrete-time LME/GAM models with spillover can also be constructed for each response definition, including scan paths.

[7] Unigrams are represented on a surprisal scale (negative log probability) simply to facilitate comparison with 5-gram effects, but I recognize that they are a degenerate (memory-less) model of surprisal and are usually included in psycholinguistic models to capture lexical retrieval rather than prediction effects (Staub, 2015).

[8] All surprisals used in this study are fully lexicalized in that the support of their underlying probability models is (a subset of) the vocabulary of English, rather than syntactic abstractions like parts of speech.

Models include a rich random effects structure to capture variation between individuals, with by-subject random intercepts, slopes, and IRF parameters. By-word random intercepts, though common in psycholinguistic studies (Demberg and Keller, 2008), are avoided because (1) they can absorb context-independent effects like *word length* and *unigram log probability* and (2) early experiments suggest that by-word intercepts lead to overfitting (based on exploratory set performance) in both CDR and baseline models. All predictors are rescaled by their standard deviations prior to fitting.[9]

Prior work suggests the following *a priori* expectations about the effect estimates for these predictors. *Saccade length*, *word length*, and *5-gram surprisal* are expected to increase processing difficulty (Demberg and Keller, 2008). According to several preceding studies, *unigram surprisal* should also positively modulate processing difficulty (see Staub, 2015, for review), although this pattern has recently been called into question (Chapter 5). *Previous was fixated* has been shown to have a facilitation effect (van Schijndel and Schuler, 2015). *Sentence position* and *document position* are designed to capture trends in the response over different timescales (sentences and documents). Previous work indicates that reading times decrease over the course of experiments (Baayen et al., 2017), suggesting an expected negative effect of *document position*. Previous estimates for *sentence position* have been small-magnitude and negative (Demberg and Keller, 2008; van Schijndel and Schuler, 2015). *Rate* effects in reading data have not been carefully studied, in part for lack of CDR (though see Shain and Schuler, 2018).

The predictors *saccade length*, *word length*, *unigram surprisal*, and *5-gram surprisal* are all motor, perceptual, or linguistic variables to which the sentence processing system has been shown to respond upon word fixation (Demberg and Keller, 2008) and to which the response might not be perfectly instantaneous. To the extent that temporally diffuse responses to any of these predictors exist, it is desirable that the model be able to capture them. By contrast, *document position* and *sentence position* merely index progress through

[9]Except *rate*, which has no variance and therefore cannot be scaled by its standard deviation of 0.

documents and sentences respectively. They are not perceptual or linguistic properties of the experiment, and it is unclear how any diffuse impulse response attributed to them would be interpreted. Following prior work (Demberg and Keller, 2008; Baayen et al., 2018), their presence in the model is motivated by the possibility of trends in the response. For this reason, parametric IRFs are fitted to all predictors except *document position* and *sentence position*, which are assigned a (parameter-free) Dirac delta IRF (i.e. a linear coefficient). *Document position* and *sentence position* are thus shown in plots as stick functions at $t = 0$. Because the functional family of the underlying response is unknown, these analyses explore the impact of the modeled response kernel by fitting *exponential* (E), *normal* (N), *shifted gamma* (G), and pseudo non-parametric linear combination of Gaussians (LCG) kernels with default initializations (§3.7.1), as in Simulation D (§4.1).

In scan path analyses of Dundee (which contain regressive fixations), it is plausible that the variables of interest exert different influences in the scan path record depending on whether they belong to a word that is being fixated for the first time vs. a word that is being re-fixated or fixated as part of a regression. This is especially true of e.g. surprisal — presumably a surprising word is less surprising after it has already been observed. While there are many conceivable ways of accounting for the possibility of such interactions in the model design, in the interests of parsimony this study uses a simple approach of splitting each variable into two predictors, one corresponding to fixations that are part of a regressive eye movement, and one corresponding to fixations that are not. These two variants thus partition the variable among the fixations. For example, the single vector of surprisal values by fixation is split into two vectors, one containing only those values that are associated with regressive fixations, and one containing only those values that are associated with non-regressive fixations, with zeros elsewhere. In all results, regressive estimates are distinguished with "(+reg)". In addition, scan path models include an indicator variable for *notregression*, to account for any generalized difference in response profile for regressive vs. non-regressive fixations.

Model Comparison

To establish a standard of comparison for evaluating predictive performance, baseline LME and GAM models are also fitted to the same data. Because the purpose of CDR is scientific modeling rather than engineering, the primary results of interest are the IRFs themselves and the insights they provide into human sentence processing. Therefore, the baseline models are used to construct a standard of reliability in predictive performance for each dataset, and comparison to them is intended to validate the CDR estimates. If CDR performs comparably to or better than baseline models in terms of generalization error, then this serves as evidence that the detailed estimates of temporal dynamics provided by CDR reliably characterize the response variable of interest. Both baseline types (LME and GAM) are fitted with and without three preceding spillover positions for each predictor (baselines with spillover are designated throughout this paper with the suffix -S), since a fourth-order FIR filter (spillover 0 through 3) is among the longest filters attested in previous naturalistic reading experiments (e.g. Smith and Levy, 2013).

GAM models are used as a reference because of their established usage in psycholinguistic data analysis. However, note that CDR and GAM are designed to address different limitations of linear models. CDR addresses the possible existence of continuous temporal diffusion of effects in non-uniform time series, while GAM does not. GAM addresses the possible existence of arbitrary smooth non-linear functional relationships between predictors and response, while CDR does not (though see Chapter 8 for a deep neural extension of CDR that does). The relative performance of CDR vs. GAM may therefore vary by dataset according to the relative importance of temporal diffusion vs. non-linear effects in describing the underlying response function. Extension of CDR to directly estimate non-linear response functions is left to future work, though see §2.1 for elaboration on a proposal to combine CDR and GAM models in a two-step regression framework.

In summary, the principal advantage of CDR for scientific modeling is the fact that it produces high-resolution estimates of temporal diffusion that cannot be obtained using

established techniques like LME or GAM. The fact that CDR additionally outperforms those other techniques in terms of overall generalization error (see Table 4.6 below) primarily supports the reliability of the model's estimates.

4.2.3 Results

Figure 4.9 presents CDR-estimated IRFs for Natural Stories, and Figures 4.11, 4.13, and 4.15 respectively present scan path, first pass, and go-past IRF estimates in Dundee. For ease of comparison, x- and y-axis dimensions are shared within each response definition per dataset. However, some estimates (Natural Stories and Dundee go-past) are so extreme that including them would compromise visual clarity. In these cases, estimates are clipped in the main plots, and plots of the full responses with adjusted y-axes are shown under the heading "Clipped plots".

Models show similar estimates of temporal dynamics across response definitions, error definitions, and IRF kernels, and generally conform to prior expectations about effect sizes and direction, as discussed below, including across different duration definitions in Dundee.[10] Furthermore, broad patterns emerge across datasets. *Rate* obtains a large-magnitude, negative, and slowly decaying IRF across datasets and kernel families. This is consistent with the existence of an *inertia* effect, such that quicker reading in the recent past

[10]There are some key exceptions to this pattern. First, some models of go-past duration in Dundee find very large early saccade length effects, in some cases orders of magnitude larger than any other effect in the model. This outcome is relatively uncommon. Exceptions are primarily found in Gaussian error models of untransformed Dundee go-past durations, although two such examples also occur in Gaussian error models of log-transformed go-past durations (see "Clipped plots" in Figure 4.15). The source of these apparent outlier estimates using go-past durations is left to future research. Second, the Gaussian error BBVI-estimated LCG model of Dundee scan path durations is a clear outlier compared to the rest of the Dundee scan path models: it finds very large credible intervals for some estimates, and its predictive performance is very poor, both in-sample and out-of-sample (Table 4.3). This result suggests divergent training and motivates caution when using heavily parameterized CDR models, at least under BBVI estimation, where initial random sampling of parameter estimates can yield poor-quality samples and large gradients. If this particular model were the basis of a scientific investigation, such evidence of training divergence would motivate the use of different hyperparameters (e.g. a lower learning rate). However, it is reassuring that this outcome is quite rare in practice, occurring in only one of the hundreds of models fitted for this study.

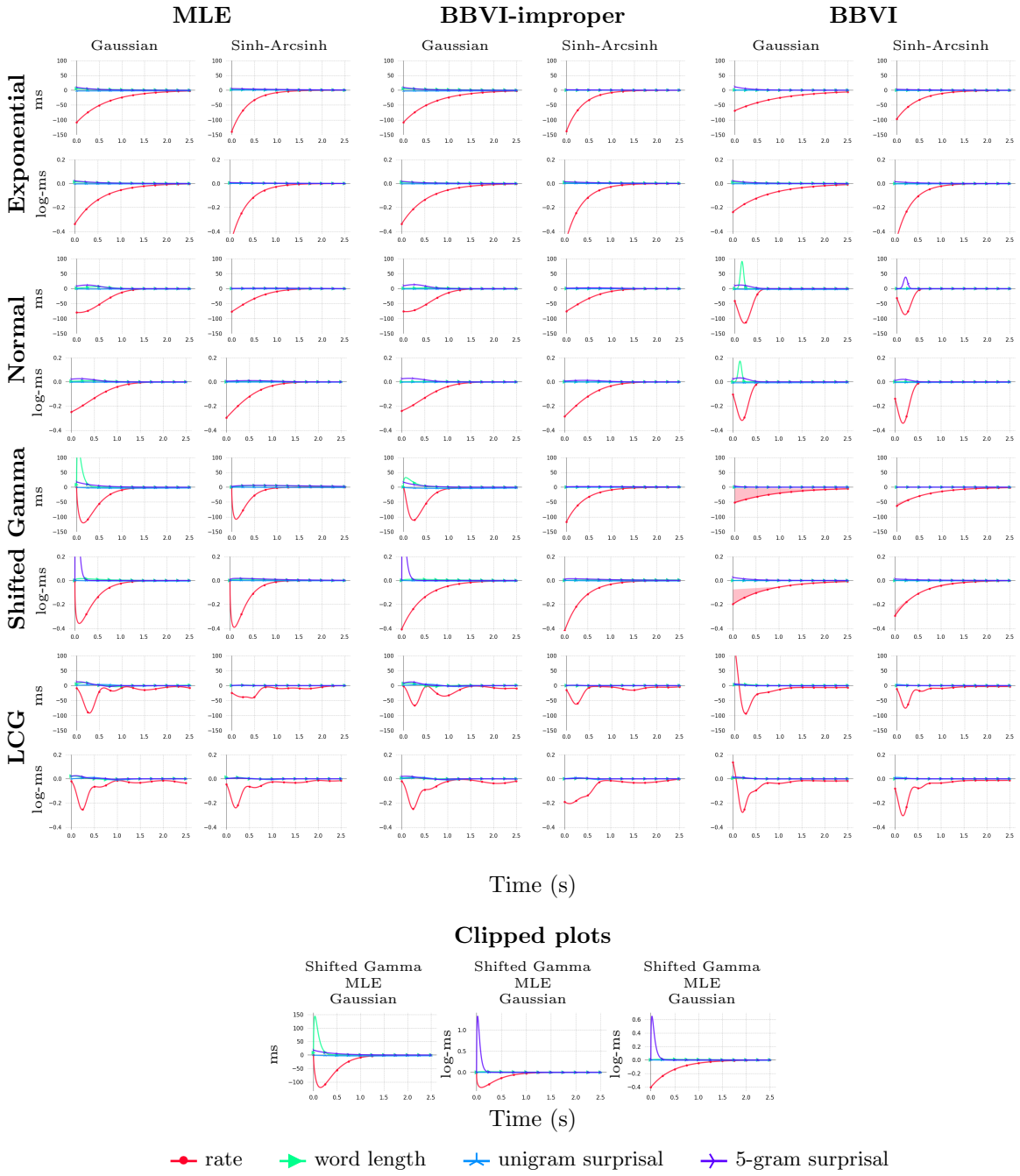


Figure 4.9: **Natural Stories:** IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels.

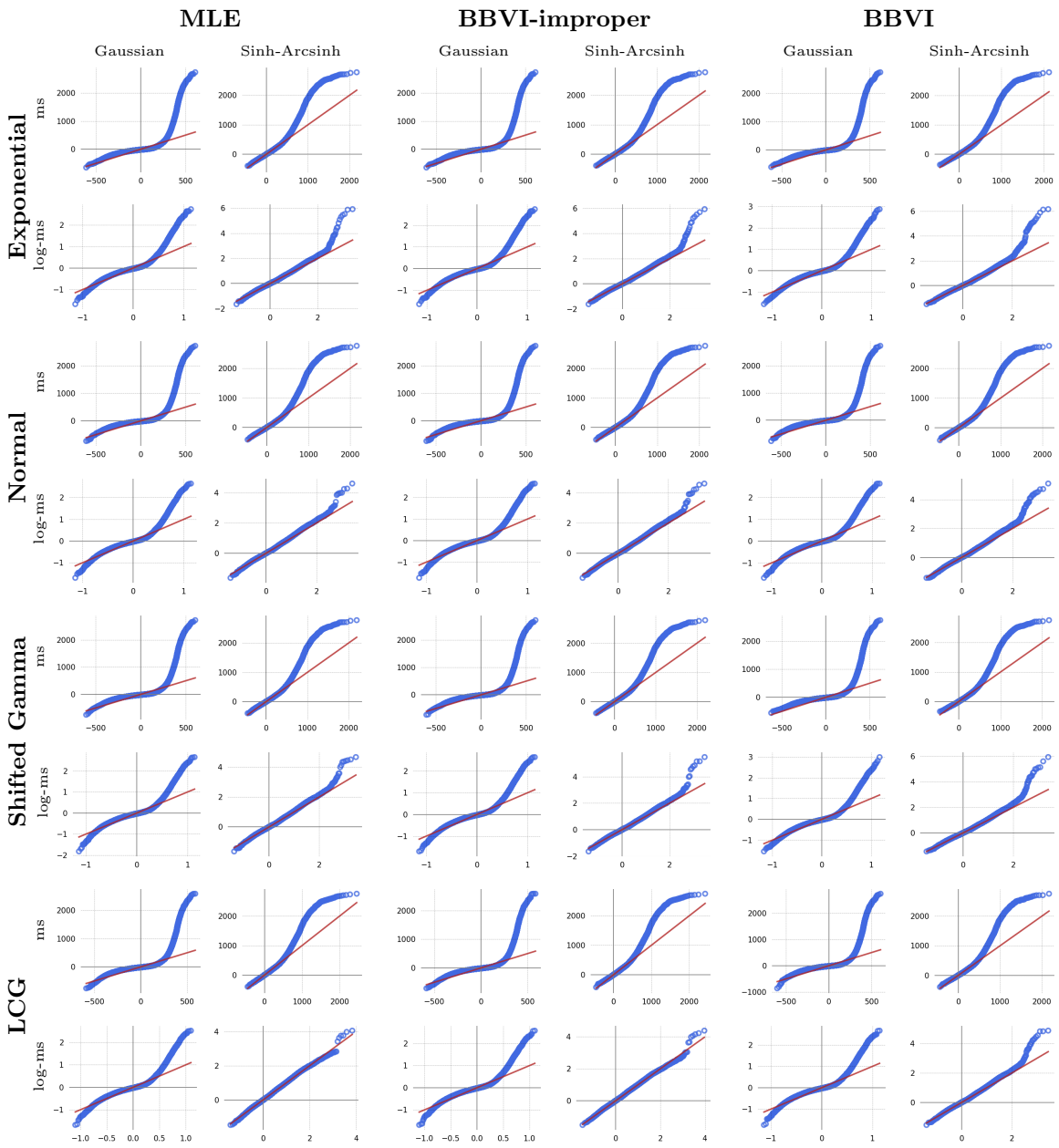


Figure 4.10: **Natural Stories:** Quantile-quantile plots of fitted (x-axis) to true (y-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red.

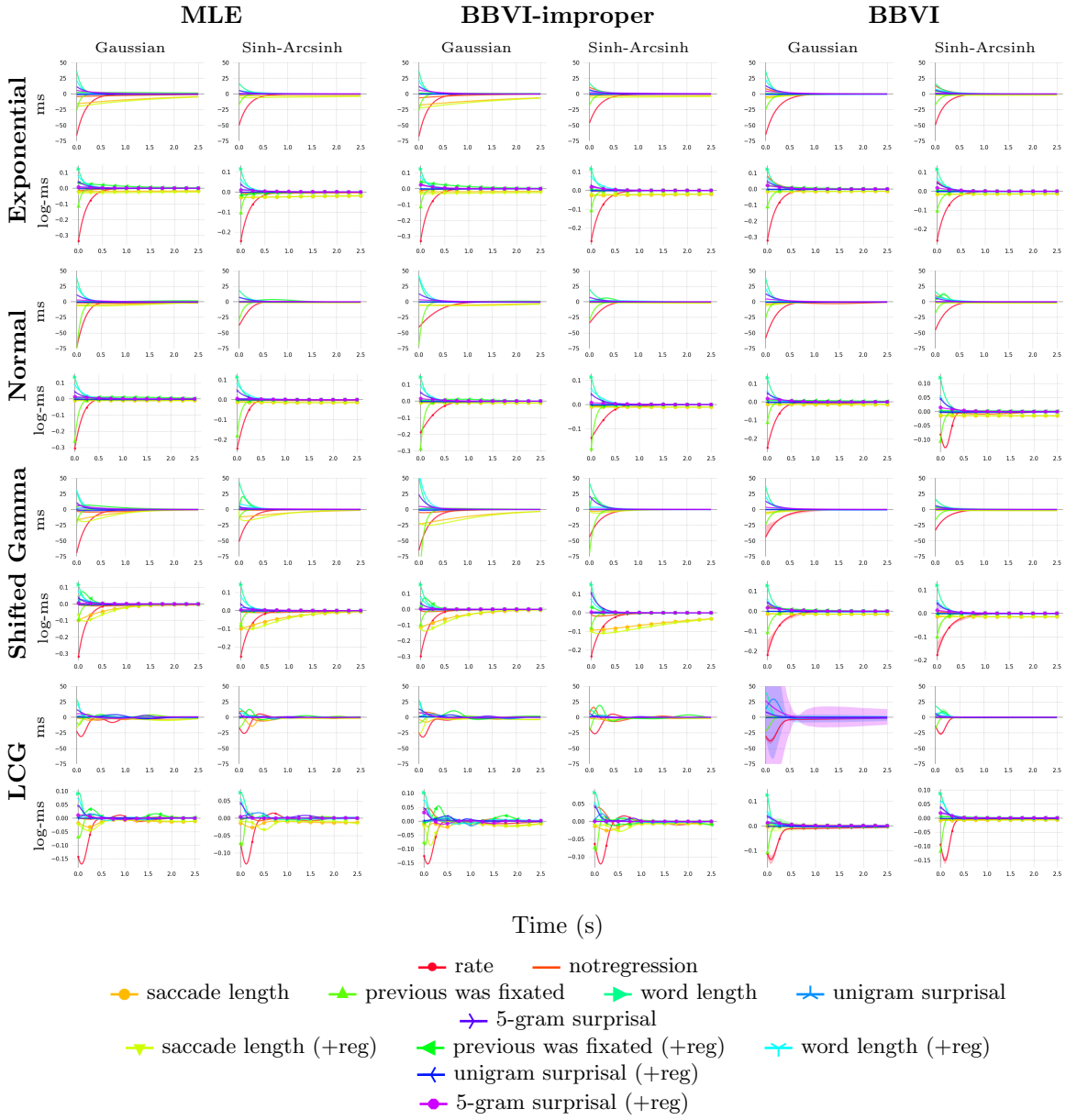


Figure 4.11: **Dundee (scan path):** IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. IRFs are distinguished by whether they correspond to response estimates within regressive eye movements (+reg) or not (–reg).

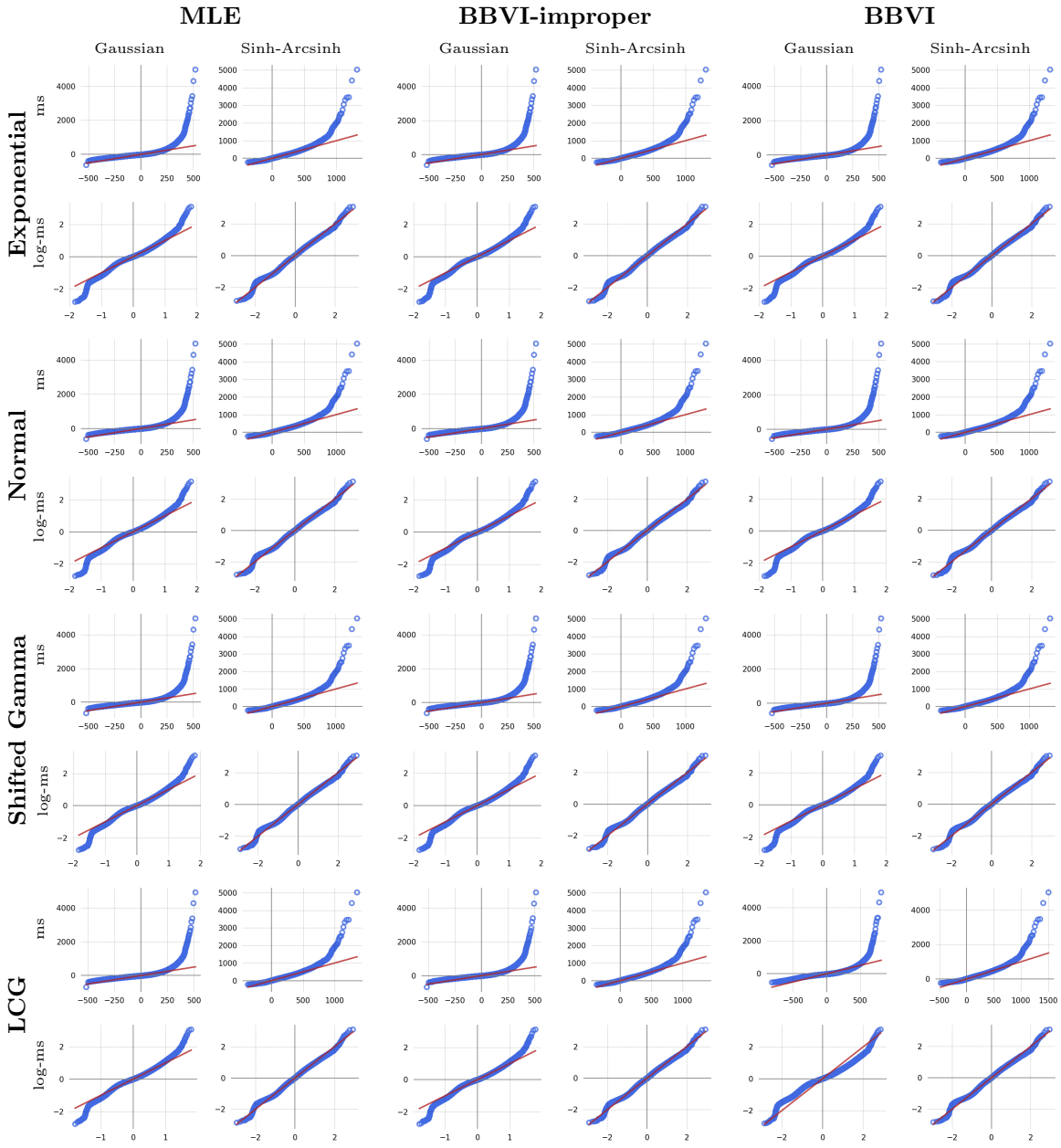


Figure 4.12: **Dundee (scan path):** Quantile-quantile plots of fitted (x-axis) to true (y-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red.

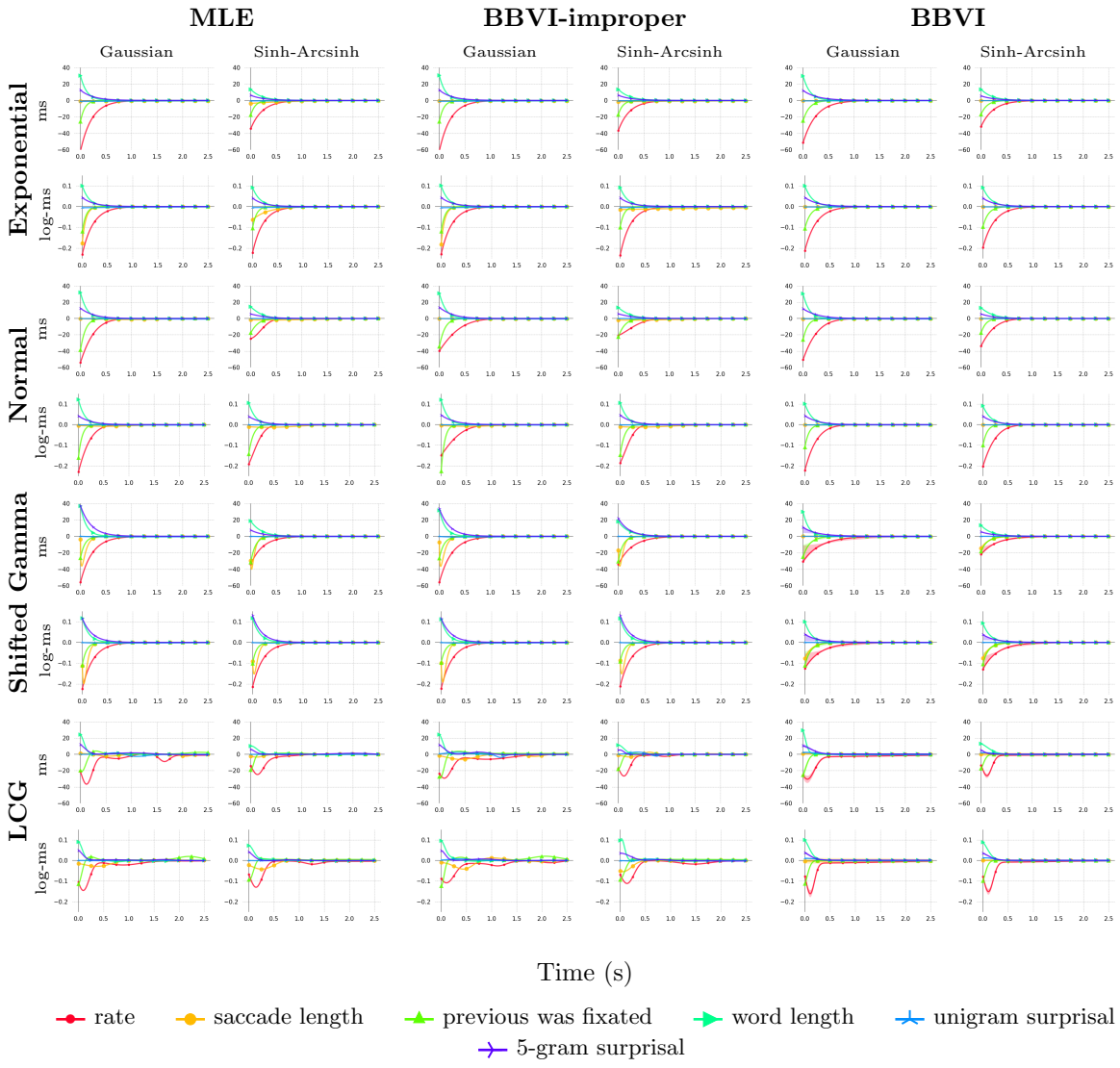


Figure 4.13: **Dundee (first past):** IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels.

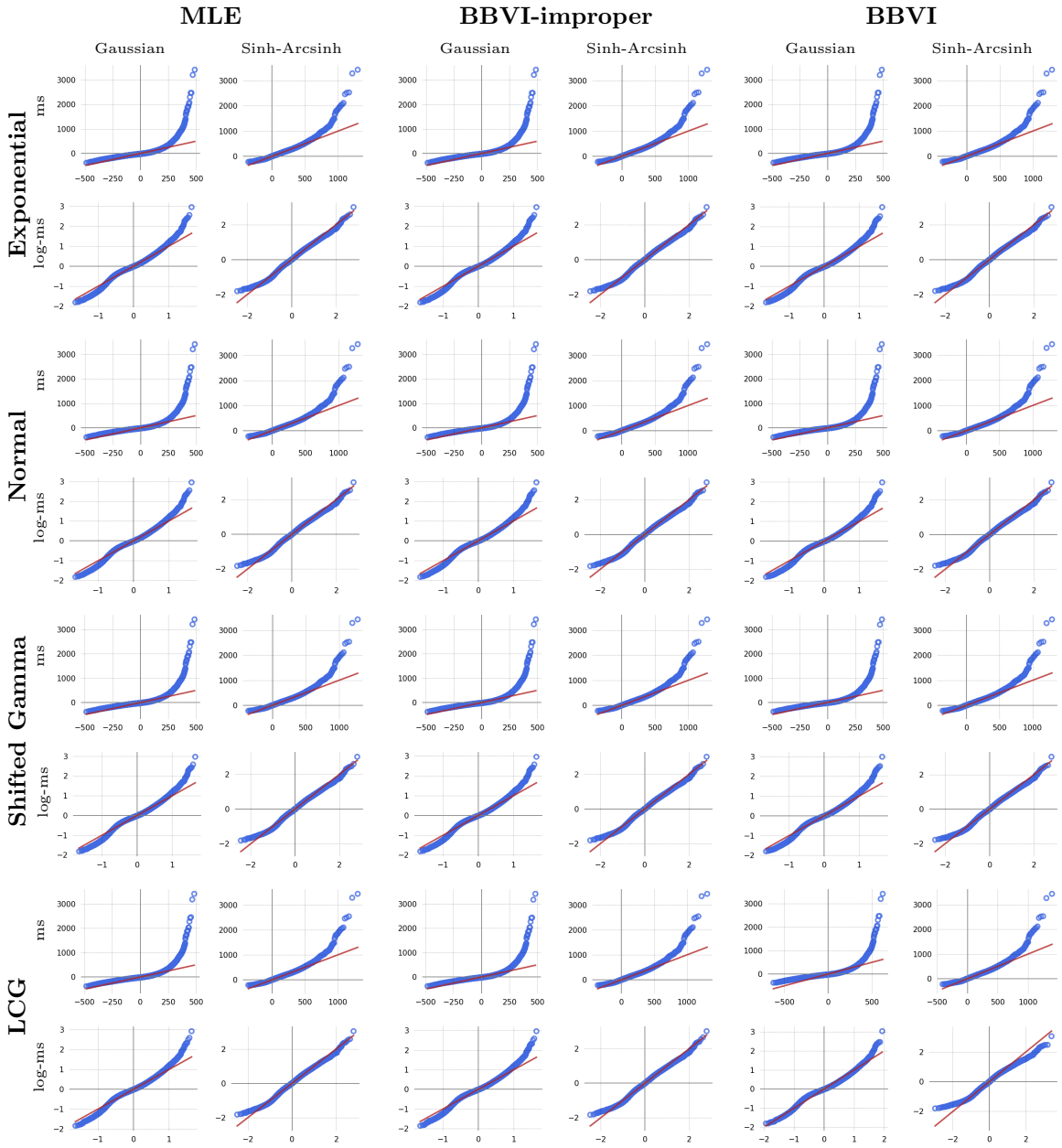


Figure 4.14: **Dundee (first pass):** Quantile-quantile plots of fitted (x-axis) to true (y-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red.

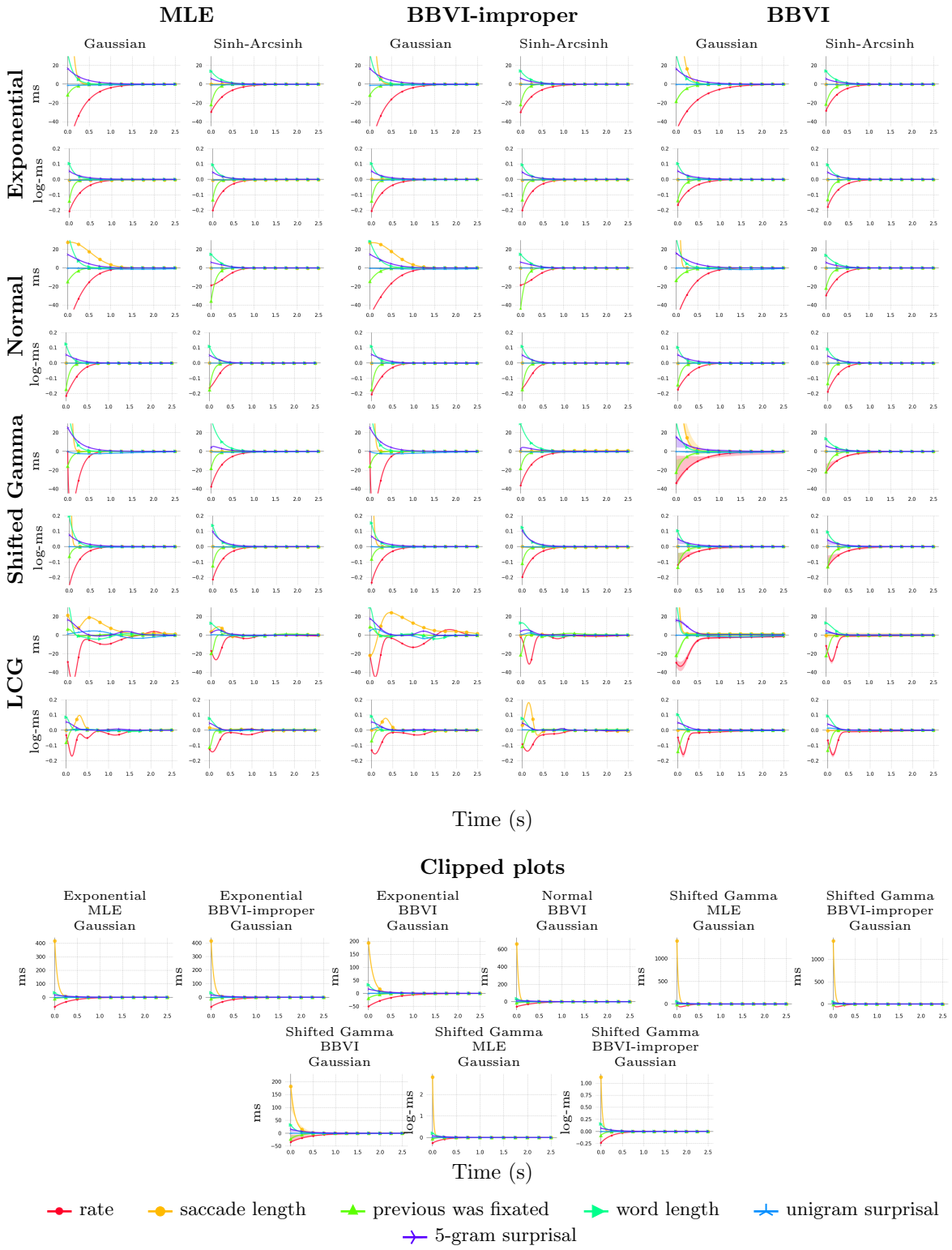


Figure 4.15: **Dundee (go-past):** IRF estimates using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels.

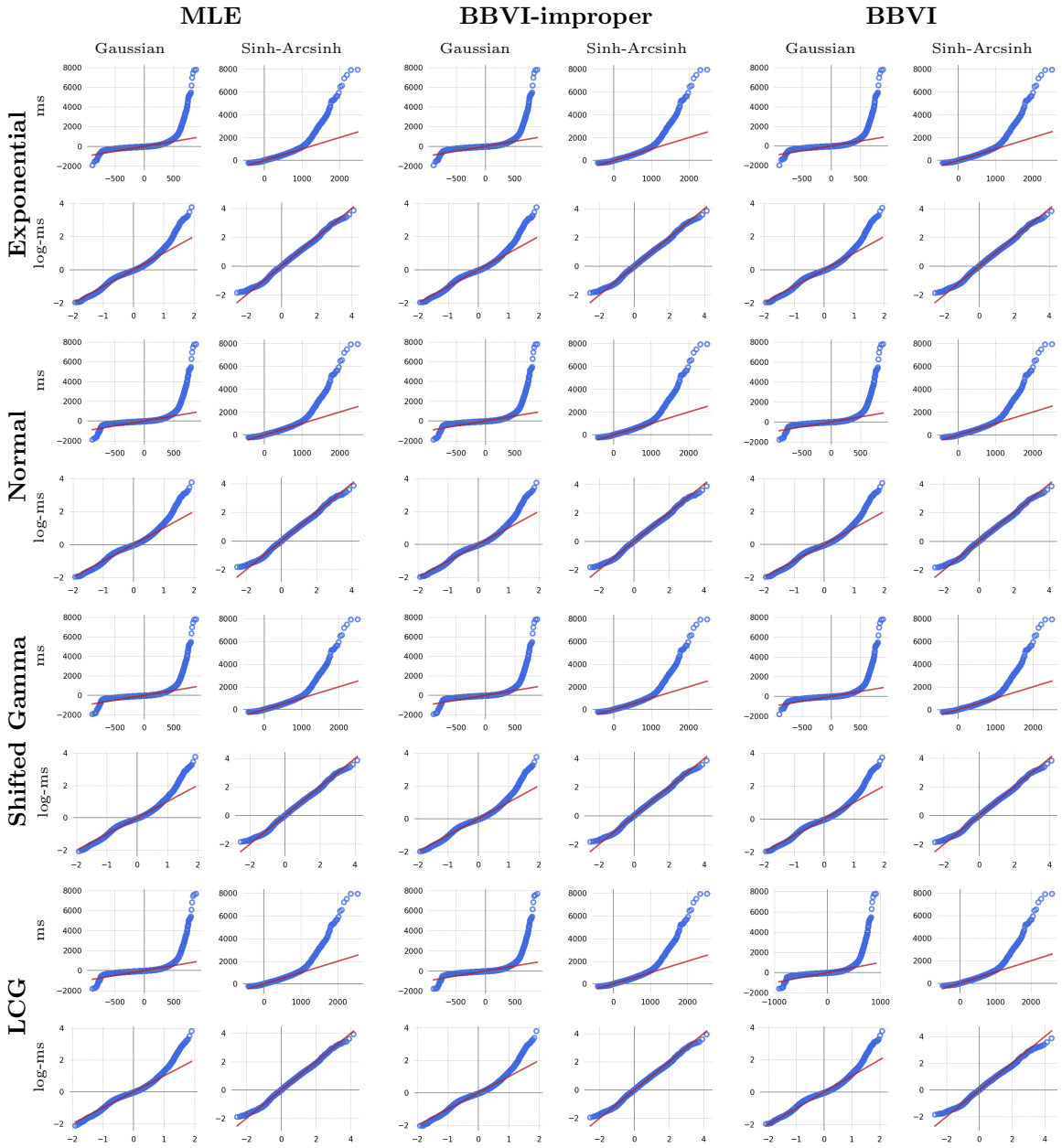


Figure 4.16: **Dundee (go-past):** Quantile-quantile plots of fitted (x-axis) to true (y-axis) error distributions on the training set, using linear vs. logarithmic responses, Gaussian vs. sinh-arcsinh error distributions, and various impulse response kernels. The theoretical best-fit line is plotted in red.

tends to engender quicker reading at the current word as well. This effect is especially pronounced in Natural Stories, where the *rate* estimate is many times larger in magnitude than that of any other predictor, suggesting that self-paced reading may be particularly susceptible to influences from inertia (i.e. habituation to repeated button presses). Nevertheless, other predictors are also estimated to influence reading latencies over and above *rate*. *Word length* and *5-gram surprisal* are given positive estimates, consistent with prior expectations of processing costs associated with each of these variables. Estimates for *unigram surprisal* in both corpora are generally negative or null, consistent with results reported in Chapter 5.

Models also reveal asymmetries. IRFs in Dundee decay more quickly than those in Natural Stories, suggesting a less pronounced influence of temporal diffusion in the eye-tracking modality compared to the self-paced reading modality. This suggestion is further supported by a much stronger improvement in predictive performance from using CDR for Natural Stories than Dundee (see Tables 4.2 and 4.3). Within the Dundee estimates, IRFs for *word length* generally decay more quickly than those for *unigram surprisal* or *5-gram surprisal*, which is consistent with the hypothesis that higher-level predictive coding and/or lexical retrieval processes entail more computation and therefore engender a slower response than a low-level occulomotor variable like *word length* (Shain and Schuler, 2018). This pattern does not seem to obtain in Natural Stories, suggesting an influence of modality on comprehension patterns. Dundee models also generally find a large-magnitude, negative, rapidly decaying IRF for *previous was fixated*, suggesting a strong but brief facilitation effect for single-word saccades, perhaps due to parafovial processing from the preceding word. Dundee scan path results also indicate that there are indeed substantially different estimates for variables depending the [±regression] dimension: all effects are attenuated under a regressive eye-movement.

As in the synthetic experiments reported in §4.1, the LCG estimates show wiggly dynamics but generally recapitulate the overall shape trends that emerge using parametric kernels. LCG models also generally do not achieve much better generalization error than

parametric models, although improvements to training fit are sometimes quite large (Tables 4.2 and 4.3). These findings suggest that the simpler parameteric kernels (1) are not too constraining to find near-optimal response shapes and (2) are less prone to overfitting than the pseudo-non-parametric kernels.

The models generally find exponential-like kernel shapes across datasets and kernel families. This suggests that the effects of these variables on reading behavior are mostly monotonic — the properties of a word exert the biggest influence on fixations to that word, with diminishing influence on fixations to subsequent words. While this has often been assumed in reading research (e.g. Rayner, 1998), here it is an emergent finding. Two key exceptions occur in the Natural Stories estimates, where *word length* is given a large late-peaking estimate under the *normal* kernel and the *rate* response is initially strongly positive and quickly dips strongly negative under the *LCG* kernel. The late-peaking *word length* possibly merits further investigation, although several considerations cast suspicion on it. First, *word length* is generally considered to be a low-level effect of visual word recognition and is thus not expected to have a late effect *a priori*. Second, many implementationally similar CDR models do not recover the same response profile for word length. Third, the model that found this particular effect shape for *word length* does not achieve systematically better generalization performance over those that did not (Table 4.2). That particular response component is thus plausibly an artifact, although follow-up study may be warranted. The initial positivity in the LCG estimate for *rate* is likely explained by two facts: (1) the *rate* estimate at time 0 is confounded with the model intercept, and (2) fixations shorter than 100ms (approximately the length of the positivity) are filtered out by standard preprocessing in Natural Stories. There is thus little training signal for the IRF shape over the interval 0-100ms, and the more flexible LCG kernel is capable of finding spurious estimates there.

Goodness of fit results are visualized as quantile-quantile plots in Figures 4.10, 4.12, 4.14, and 4.16. As shown, both the normalizing (log) transform and the asymmetric sinh-

arcsinh error distribution improve goodness of fit across model types, with the best fit consistently occurring in sinh-arcsinh models of log-transformed fixation durations. The true error distribution for raw fixation durations tends to have a heavier right tail than that of the fitted distribution, even under sinh-arcsinh. This suggests that sinh-arcsinh on its own may not be sufficiently expressive to account for very skewed data. However, sinh-arcsinh consistently eliminates the heavy *left* tail visible in the Gaussian models, and it tends to better account for the right tail, thus improving fit across the board compared to Gaussian models in matched experimental conditions. These results indicate that sinh-arcsinh error is beneficial across designs, even if it cannot completely eliminate poor fit on its own.

In summary, the IRFs recovered by CDR accord with prior expectations about the reading response and are largely consistent across modalities and kernel definitions, while shedding additional light on fine-grained details of temporal structure.

To validate the CDR estimates, comparisons to baseline LME and GAM models for all response definitions in Natural Stories and Dundee are presented in Tables 4.2 – 4.5. CDR models outperform (i.e. achieve better exploratory and/or test set error than) all baselines on Natural Stories, and they generally outperform the LME-based models on Dundee. GAM without spillover outperforms CDR on Dundee without log transformation and mostly underperforms CDR with log transformation, although the errors are similar in magnitude throughout. GAM with spillover systematically outperforms CDR in Dundee, though again the errors are relatively close, especially under log transformation. The limited performance of CDR relative to GAM with spillover is likely due to some combination of (1) the relatively constrained degree of temporal diffusion in Dundee, as revealed by the rapidly decaying response estimates in Figures 4.11, 4.13, and 4.15, and (2) the requirement of CDR to generate responses through linear combination of the convolved predictors, while GAM estimates non-linear relationships between predictors and response, and (3) the greater concision of CDR models, which contain substantially fewer parameters than GAM models

	Natural Stories (ms)			Natural Stories (log-ms)		
Model	Train	Expl	Test	Train	Expl	Test
LME	20179	20624	20369	0.0803	0.0818	0.0815
LME-S	19980†	20471†	20230†	0.0789†	0.0807†	0.0803†
GAM	20070	20501	20255	0.0798	0.0814	0.0810
GAM-S	*19873*	*20349*	*20109*	*0.0784*	*0.0802*	*0.0799*
CDR-E-MLE	17766	18172	—	0.0630	0.0643	—
CDR-E-BBVI.imp	17765	18168	—	0.0630	0.0643	—
CDR-E-BBVI	18106	18361	—	0.0644	0.0651	—
CDR-N-MLE	17487	18060	—	0.0622	0.0641	—
CDR-N-BBVI.imp	17500	18058	—	0.0622	0.0640	—
CDR-N-BBVI	17686	18182	—	0.0630	0.0645	—
CDR-G-MLE	17510	18053	—	0.0620	0.0656	—
CDR-G-BBVI.imp	17551	18045	—	0.0623	0.0643	—
CDR-G-BBVI	18118	18373	**18212**	0.0646	0.0652	**0.0654**
CDR-LCG-MLE	16222	18785	—	0.0569	0.0657	—
CDR-LCG-BBVI.imp	***16153***	18770	—	***0.0567***	0.0659	—
CDR-LCG-BBVI	17437	***17805***	—	0.0613	***0.0627***	—

Table 4.2: **Natural Stories.** CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

with spillover (GAM fits multidimensional smooth functions for each spillover position of each predictor, in addition to the random effects model). Indeed, Chapter 8 shows that relaxing linearity constraints on CDR allows generalization performance to surpass that of GAM-S. Nonetheless, the ensemble of exploratory set comparisons support the reliability of CDR estimates, since they yield similar or improved generalization performance across the board.

CDR is statistically evaluated against the baseline models in two ways: (1) by comparing test-set performance against each baseline and (2) by comparing overall *success rates* of CDR models against each baseline on the exploratory set. For (1), to avoid multiple comparisons, CDR-G-BBVI is selected as the representative CDR model across reading datasets, its generalization performance on the test set is compared to that of all baseline

Model	Dundee (ms)			Dundee (log-ms)		
	Train	Expl	Test	Train	Expl	Test
LME	14645†	15215†	15230†	0.1790†	0.1807†	0.1801†
LME-S	—	—	—	—	—	—
GAM	14476	15055	15064	0.1779	0.1795	0.1791
GAM-S	*14340*	***14942***	***14973***	*0.1757*	***0.1776***	***0.1773***
CDR-E-MLE	14510	*15055*	—	0.1769	0.1784	—
CDR-E-BBVI.imp	14508	*15055*	—	0.1769	0.1784	—
CDR-E-BBVI	14573	15107	—	0.1775	0.1788	—
CDR-N-MLE	14493	*15055*	—	0.1765	0.1783	—
CDR-N-BBVI.imp	14501	15069	—	0.1764	*0.1782*	—
CDR-N-BBVI	14545	15081	—	0.1775	0.1788	—
CDR-G-MLE	14501	15063	—	0.1767	0.1785	—
CDR-G-BBVI.imp	14508	15058	—	0.1766	0.1784	—
CDR-G-BBVI	14574	15104	15171	0.1776	0.1789	0.1789
CDR-LCG-MLE	14109	15204	—	***0.1711***	0.1810	—
CDR-LCG-BBVI.imp	***14083***	15283	—	***0.1711***	0.1806	—
CDR-LCG-BBVI	18295	18675	—	0.1779	0.1792	—

Table 4.3: **Dundee (scan path).** CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures. LME-S performance metrics could not be obtained for scan paths because of long runtimes required for training.

Model	Dundee (ms)			Dundee (log-ms)		
	Train	Expl	Test	Train	Expl	Test
LME	13152†	14204†	14026†	0.1516	0.1542	0.1531
LME-S	13112†	14162†	14024†	0.1507†	0.1532†	0.1526†
GAM	13007	14065	13871	0.1510	0.1536	0.1525
GAM-S	*12882*	***13948***	***13771***	*0.1491*	***0.1518***	***0.1508***
CDR-E-MLE	13040	14077	—	0.1500	0.1529	—
CDR-E-BBVI.imp	13040	14079	—	0.1500	0.1530	—
CDR-E-BBVI	13071	14103	—	0.1505	0.1529	—
CDR-N-MLE	13030	14163	—	0.1498	0.1542	—
CDR-N-BBVI.imp	13037	*14068*	—	0.1498	0.1543	—
CDR-N-BBVI	13063	14077	—	0.1504	*0.1526*	—
CDR-G-MLE	13034	14069	—	0.1499	0.1534	—
CDR-G-BBVI.imp	13037	14072	—	0.1499	0.1534	—
CDR-G-BBVI	13073	14106	13960	0.1505	0.1539	0.1520
CDR-LCG-MLE	***12769***	14189	—	***0.1465***	0.1551	—
CDR-LCG-BBVI.imp	12788	14109	—	0.1466	0.1547	—
CDR-LCG-BBVI	13069	14092	—	0.1501	*0.1526*	—

Table 4.4: **Dundee (first pass).** CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

models.[11] Comparisons use a paired permutation test (Demšar, 2006) with 10,000 resampling iterations, pooling error vectors from all tasks (both linear and log responses from both Natural Stories and Dundee) into a single test.[12] For (2), the empirical probability of an arbitrarily chosen CDR model outperforming each baseline model on the exploratory set (a *success*) is tested against a null hypothesis of chance probability (0.5), using a binomial test. This test assesses the general robustness of CDR compared to the baselines, aggregating over CDR hyperparameters by testing whether the probability of improvement from using an arbitrarily chosen CDR model type differs from chance. Results are given in Table 4.6. As shown, CDR-G-BBVI significantly outperforms all baselines in terms of test-

[11]The choice of BBVI for this analysis was motivated in §3.7.4. Of the BBVI models, comparisons focus on *shifted gamma* because it is the most flexible of the parametric kernels and therefore likely to be of interest to future applications of CDR to reading times. Differences in generalization performance between kernels tend to be small.

[12]To ensure comparability across corpora with different error variances, per-datum errors are first scaled by their standard deviations within each corpus. Standard deviations are computed over the joint set of error values in each pair of CDR and baseline models.

Model	Dundee (ms)			Dundee (log-ms)		
	Train	Expl	Test	Train	Expl	Test
LME	44184	39523	42948	0.2073†	0.2072†	0.2070†
LME-S	44097†	39476†	43014†	0.2057†	0.2057†	0.2058†
GAM	43976	***39289***	42704	0.2063	0.2061	0.2060
GAM-S	*43476*	39483	*42180*	*0.2034*	***0.2036***	***0.2035***
CDR-E-MLE	43094	41322	—	0.2049	0.2046	—
CDR-E-BBVI.imp	43094	41338	—	0.2049	0.2046	—
CDR-E-BBVI	43178	41926	—	0.2054	0.2051	—
CDR-N-MLE	42935	41449	—	0.2045	0.2043	—
CDR-N-BBVI.imp	42944	41351	—	0.2048	0.2042	—
CDR-N-BBVI	42975	40776	—	0.2053	0.2047	—
CDR-G-MLE	***42864***	39844	—	0.2039	*0.2036*	—
CDR-G-BBVI.imp	***42864***	39876	—	0.2041	0.2037	—
CDR-G-BBVI	43222	40970	**40018**	0.2054	0.2051	0.2052
CDR-LCG-MLE	42329	41798	—	***0.1998***	0.2061	—
CDR-LCG-BBVI.imp	42336	41678	—	0.2002	0.2052	—
CDR-LCG-BBVI	42995	39540	—	0.2047	0.2044	—

Table 4.5: **Dundee (go-past).** CDR vs. baselines, mean-squared error. CDR results shown using (E)xponential, (N)ormal, Shifted (G)amma, and non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

	Permutation test	Binomial test	
Baseline	p	Success rate	p
LME	**1.0e-4*** **	0.90	**2.2e-16*** **
LME-S	**1.0e-4*** **	0.87	**2.2e-16*** **
GAM	**1.0e-4*** **	0.67	**2.2e-6*** **
GAM-S	**1.0e-4*** **	0.36	0.36

Table 4.6: **Reading data model comparison**. Permutation tests of improvement on test set from CDR-G-BBVI over baselines (pooled across all tasks), along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated). **Note:** Dundee scan path durations are excluded from the LME-S comparison because runtime limits prevented training from terminating.

set error. In addition, the CDR success rate over all baselines but GAM-S is significantly greater than chance, indicating that an arbitrarily chosen CDR configuration is likely to outperform these baselines, even without careful tuning on an exploratory set.[13]

4.2.4 Discussion

This application of CDR to the study of human reading latencies shows consistent estimates across response kernels that largely align with prior expectations but additionally provide high-resolution insights into underlying temporal dynamics that are difficult to obtain using standard statistical models. Results additionally show a significant overall improvement in generalization error from CDR over baseline models, supporting the trustworthiness of estimated response functions.

4.3 fMRI Experiments

In this analysis, CDR is used to infer the shape of the hemodynamic response function (HRF) from fMRI measures of brain responses to variably-spaced naturalistic stimuli. There is already an extensive literature on HRF discovery using discrete-time deconvolutional methods (Josephs et al., 1997; Friston et al., 1998a; Miezin et al., 2000; Gitelman et al., 2003; Lindquist et al., 2009; Pedregosa et al., 2014, *inter alia*). These approaches rely on stimulus designs in which events are regularly spaced and aligned with the fMRI scan times. For fMRI researchers seeking the benefits to ecological validity afforded by the naturalistic experimental paradigm (Hasson et al., 2010; Hasson and Honey, 2012; Hasson et al., 2018; Campbell and Tyler, 2018), this requirement poses a problem, since many

[13]The failure to outperform GAM-S in the binomial test may to be due in part to the increased weight implicitly placed on Dundee, where CDR performs less well relative to GAM-S, by considering multiple response definitions. Binomial tests using first pass or go-past durations only are significant even over GAM-S, though the success rates are lower than against the other baselines. This again suggests that the relative importance of controlling for temporal diffusion (CDR) vs. non-linear effects (GAM) may be less great in the Dundee corpus, where diffusion appears to be relatively constrained. Nonetheless, a major advantage of the CDR approach is its ability to estimate the extent of diffusion in general, and thus to reveal circumstances in which diffusion plays a more or less pronounced role.

naturalistic stimuli (including language) do not consist of regularly spaced events occurring at integer multiples of the fMRI scanner's acquisition rate. Naturalistic fMRI experiments are therefore an important target application of continuous-time deconvolution, since CDR imposes no such requirement on the stimulus design. Here, CDR models are trained on a naturalistic fMRI dataset and compared to existing methods for modeling fMRI measures of neural responses to naturalistic language stimuli.

4.3.1 Data

fMRI data collection and preprocessing are described in detail in Chapter 6,[14] with only high level details presented here. Data were collected from 78 participants (30 males) exposed to auditory presentation of texts from the Natural Stories corpus (Futrell et al., 2018) read by one of two speakers (1 male, 1 female). Left-hemisphere fronto-temporal language regions are functionally localized on a participant-specific basis using a separate localizer task (Fedorenko et al., 2010; Braze et al., 2011; Vagharchakian et al., 2012; Blank et al., 2016; Scott et al., 2017, *inter alia*). The response variable consists of average blood oxygen level dependent (BOLD) contrast imaging signal within the voxels of six functionally defined regions of interest (fROIs) constituting the left-hemisphere fronto-temporal language network: inferior frontal gyrus (IFG) and its orbital part (IFGorb), middle frontal gyrus (MFG), anterior temporal cortex (ATL), posterior temporal cortex (PTL), and angular gyrus (AngG).

Similarly to Experiment A, training (50%), exploratory (25%), and test (25%) sets are created using modular arithmetic. The fMRI study uses a slower partitioning cycle (15 TRs or 30s) compared to the reading study in §4.2, which is motivated by a desire to reduce correlation between the elements of the partition in light of strong prior evidence that the BOLD signal is highly auto-correlated. In particular, TR numbers e are cycled into different bins of the partition with a different phase for each subject u: $\text{partition}(e, u) = \lfloor \frac{e+u}{15} \rfloor \bmod 4$,

[14] Data are available at `https://osf.io/eyp8q/`.

again assigning outputs 0 and 1 to the training set, 2 to the exploratory set, and 3 to the evaluation (test) set.

4.3.2 Experimental Settings

Predictor and Response Variables

The dependent variable is average BOLD response within each fROI, with all fROIs combined into a single model. Unlike reading latencies, BOLD measurements are not strictly positive and generally not heavily skewed. Therefore, in contrast to the reading experiments above, no normalizing (e.g. logarithmic) transformations are explored on the fMRI response, although analyses still explore the impact of Gaussian vs. sinh-arcsinh error. In addition to *rate*, *unigram surprisal*, and *5-gram surprisal*, all implemented identically to the reading experiments in §4.2, models also include a predictor for *sound power* (e.g. Brennan et al., 2016), estimated as frame-by-frame root mean squared energy (RMSE) of the audio stimuli computed using the Librosa software library (McFee et al., 2015). Because *sound power* is a continuous rather than event-based predictor, it is implemented by taking regular RMSE samples every 100ms. *Sound power* thus uses RMSE sample times as timestamps rather than word onsets, and CDR's event-based deconvolutional procedure thus implicitly uses a Riemann sum approximation of the continuous *sound power* convolution integral. To implement the assumption of a fixed-shape hemodynamic response in a given cortical region, the parameters of the IRF kernel are tied across all predictors within each region, while giving each predictor its own coefficient in order to estimate different response amplitudes. Models also include a linear predictor for repetition time number (*TR number*, the sample's index within the current story), designed to capture any linear trends in the overall response. Models contain by-fROI random intercepts, slopes, and HRF parameters for each of these predictors, along with by-subject random intercepts.[15]

[15] As discussed in Chapter 6, this dataset does not appear to support identification of richer by-subject random effects models, which generalize poorly to the out-of-sample sets.

Hemodynamic Response Kernels

Several IRF kernels based on the double-gamma hemodynamic response function (Boynton et al., 1996) are considered:

$$f(x;\alpha_1,\beta_1,c,\alpha_2,\beta_2) = \frac{\beta_1^{\alpha_1} x^{\alpha_1-1} e^{-\beta_1 x}}{\Gamma(\alpha_1)} - c\frac{\beta_2^{\alpha_2} x^{\alpha_2-1} e^{-\beta_2 x}}{\Gamma(\alpha_2)} \tag{4.1}$$

The normalization constant for this kernel is simply $\frac{1}{1-c}$, since it consists of a sum of two scaled gamma probability densities whose integrals over the positive reals are 1 and $-c$, respectively. A 5-parameter HRF5 kernel is therefore defined as:

$$\text{HRF5}(x;\alpha_1,\beta_1,c,\alpha_2,\beta_2) \stackrel{\text{def}}{=} f(x;\alpha_1,\beta_1,c,\alpha_2,\beta_2) \;/\; (1-c) \tag{4.2}$$

Because the double-gamma HRF is fairly heavily parameterized (5 parameters), this study also explores the impact of reparameterizations that reduce the flexibility of the kernel through parameter tying, constraining the kernel toward the canonical HRF, where $\alpha_1 = 6$, $\beta_1 = 1$, $c = \frac{1}{6}$, $\alpha_2 = 16$, and $\beta_2 = 1$ (Lindquist et al., 2009). Thus, in addition to the 5-parameter kernel shown in eq. 4.2 (HRF5), this study also considers the following kernel variants:

- A 4-parameter variant (HRF4) with tied rate parameter β:

$$\text{HRF4}(x;\alpha_1,\beta,c,\alpha_2) \stackrel{\text{def}}{=} \left(\frac{\beta^{\alpha_1} x^{\alpha_1-1} e^{-\beta x}}{\Gamma(\alpha_1)} - c\frac{\beta^{\alpha_2} x^{\alpha_2-1} e^{-\beta x}}{\Gamma(\alpha_2)}\right) \;/\; (1-c) \tag{4.3}$$

- A 3-parameter variant (HRF3) which additionally ties the shape parameters α_1 and α_2 to have a constant offset of 10, as used by SPM's canonical HRF:

$$\text{HRF3}(x;\alpha,\beta,c) \stackrel{\text{def}}{=} \left(\frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} - c\frac{\beta^{\alpha+10} x^{\alpha+9} e^{-\beta x}}{\Gamma(\alpha+10)}\right) \;/\; (1-c) \tag{4.4}$$

Kernel	Parameters
HRF1	128
HRF2	135
HRF3	142
HRF4	149
HRF5	156
LCG	331

Table 4.7: Number of parameters by kernel in the fMRI experiment. Note that BBVI and BBVI-improper double these figures by additionally fitting variances for each parameter in the variational posterior, and that sinh-arcsinh models additionally include parameters for the skewness and tailweight of the response.

- A 2-parameter variant (HRF2) which additionally fixes the undershoot constant c at SPM's default value of $\frac{1}{6}$:

$$\text{HRF2}(x; \alpha, \beta) \stackrel{\text{def}}{=} \left(\frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} - \frac{\beta^{\alpha+10} x^{\alpha+9} e^{-\beta x}}{6\Gamma(\alpha+10)} \right) \Big/ \left(\frac{5}{6} \right) \tag{4.5}$$

- A 1-parameter variant which fixes the shape parameter α at SPM's default value of 6, implementing a "stretchable" canonical HRF:

$$\text{HRF1}(x; \beta) \stackrel{\text{def}}{=} \left(\frac{\beta^6 x^5 e^{-\beta x}}{\Gamma(6)} - \frac{\beta^{16} x^{15} e^{-\beta x}}{6\Gamma(16)} \right) \Big/ \left(\frac{5}{6} \right) \tag{4.6}$$

In all of these kernels, parameters are initialized at the SPM defaults for the canonical HRF presented above.

As in reading and synthetic experiments, this study also consider LCG kernels. As discussed above, the LCG models are more heavily parameterized than those with parametric kernels (Table 4.7, see Table 4.1 for reading models).

Model Comparison

To validate the CDR estimates of the HRF, CDR fits are compared to those produced by four existing approaches for modeling naturalistic fMRI experiments. The first approach

is pre-convolving the stimuli using the canonical HRF (Canonical HRF), as is done in many naturalistic studies (Brennan et al., 2012; Willems et al., 2015; Henderson et al., 2015, 2016; Lopopolo et al., 2017, *inter alia*). This approach has advantages for parsimony, since it avoids the need to fit coefficients at multiple time offsets. But it assumes a fixed, universal hemodynamic response that may not accurately describe the response profile in a given brain region (Handwerker et al., 2004). The second approach is using piecewise linear interpolation to resample the predictors at timepoints that align with the fMRI scan times (Interpolated). This approach distorts the predictor series by treating it as a sequence of samples from an underlyingly continuous signal (see §2.2 for discussion). The third approach is using the fMRI scan times to define discrete temporal bins and then average the predictor values within each bin (Averaged). This approach has been used, for example, by Wehbe et al. (2020). The fourth approach follows Huth et al. (2016) in downsampling the predictor series to the temporal resolution of the fMRI signal by convolving it with a low-pass Lanczos filter with 3 lobes and a cutoff frequency of 0.25, the Nyquist frequency of the fMRI scanner (Lanczos).[16] This method essentially implements a "soft" variant of the averaging approach by taking a weighted sum of the stimuli in the neighborhood of an fMRI sample, weighted by a function (the Lanczos kernel) of the temporal distance between the stimulus and the sample. Mixed models with the same random effects structure as the CDR models described above were fitted using the `lme4` package. For the Canonical HRF baseline, the model contains a single fixed and by-fROI random slope per predictor, since the temporal modeling is implemented by the convolutional preprocess. For the Interpolated, Averaged, and Lanczos baselines, fourth-order FIR models are applied with fixed and by-fROI random slopes for the four TR's preceding an fMRI sample. These FIR kernels implement a discretized version of the HRF and estimate its (non-parametric) shape from data.

[16] Source code for this technique is available at `https://github.com/HuthLab/speechmodeltutorial`.

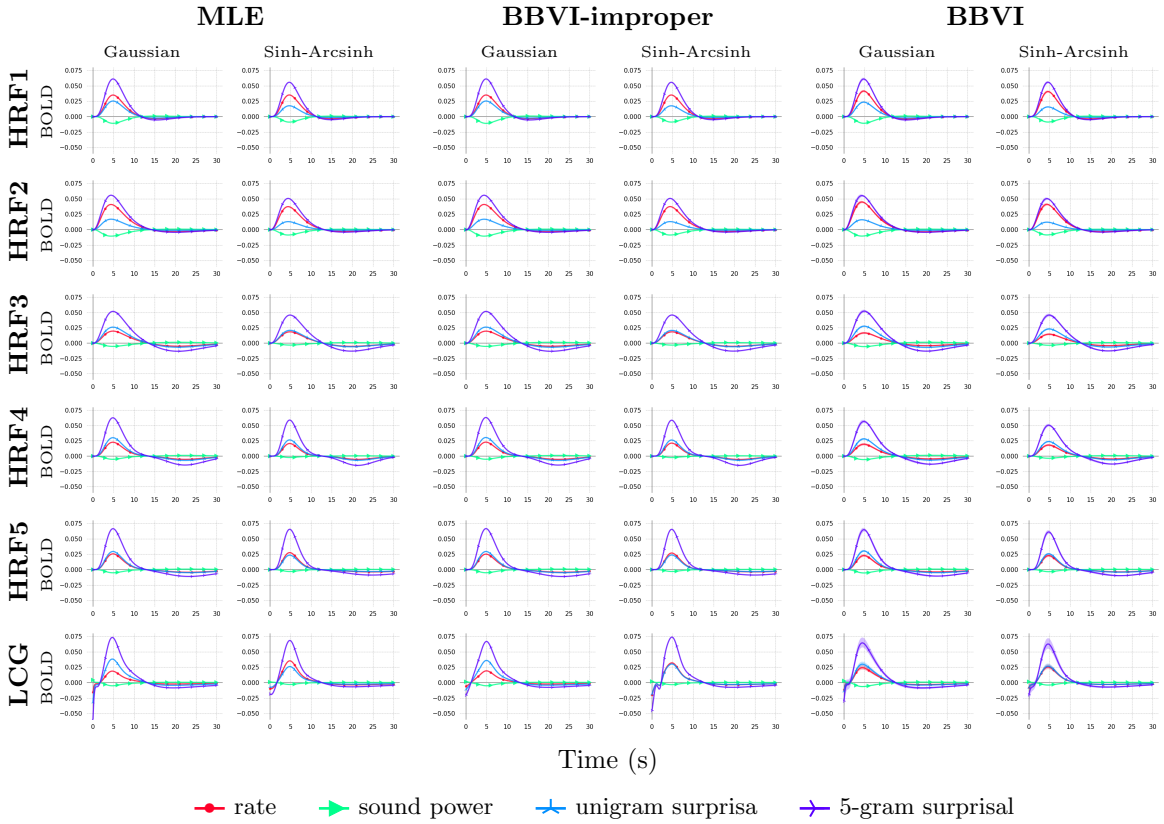


Figure 4.17: **Natural Stories fMRI:** HRF estimates using Gaussian vs. sinh-arcsinh error distributions and various hemodynamic response kernels.

4.3.3 Results and Discussion

Figure 4.17 shows plots of all CDR estimates on fMRI. As in synthetic and reading experiments, results are highly consistent across these dimensions. As in the reading data, sinh-arcsinh substantially improves goodness of fit of modeled error distribution (Figure 4.18).

Models estimate *5-gram surprisal* to have the largest influence on the language network's response, generally followed by *rate* and then by *unigram surprisal*. The *sound power* predictor tends to be assigned a small-magnitude negative response.

Figure 4.17 shows that all models find estimates that closely resemble the canonical HRF, and that these estimates do not change dramatically in the simplified reparameterizations of the HRF. At the same time, some models deviate from the canonical HRF in noteworthy

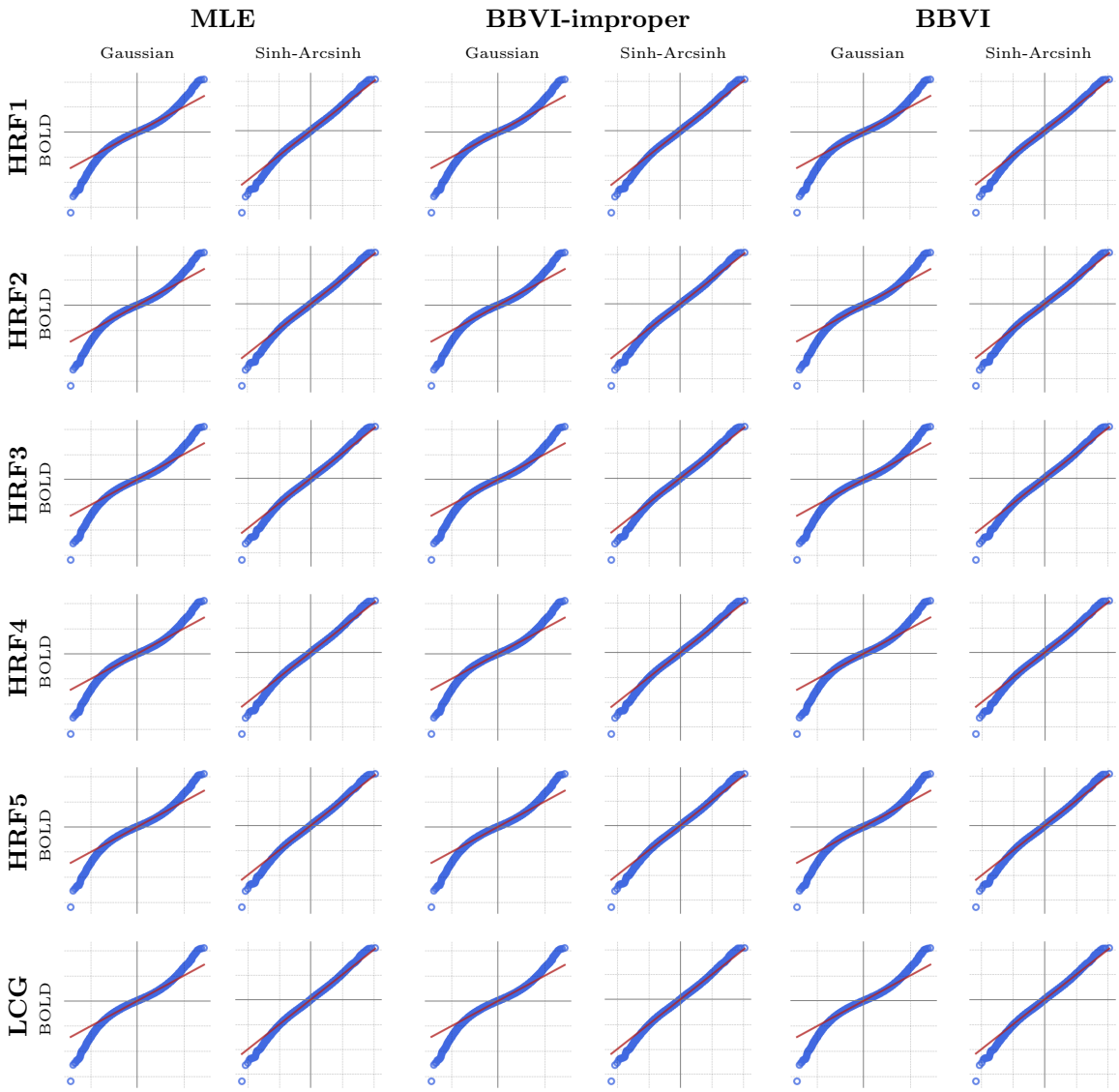


Figure 4.18: **Natural Stories fMRI:** Quantile-quantile plots of fitted (x-axis) to true (y-axis) error distributions on the training set, using Gaussian vs. sinh-arcsinh error distributions and various hemodynamic response kernels. The theoretical best-fit line is plotted in red.

	Natural Stories fMRI		
Model	Train	Expl	Test
Canonical HRF	11.3548†	*11.8263*†	*11.5661*†
Interpolated	11.4236†	11.9888†	11.6654†
Averaged	*11.3478*†	11.9280†	11.6090†
Lanczos	11.3536†	11.9059†	11.5871†
CDR-HRF1-MLE	11.3442	11.8729	—
CDR-HRF1-BBVI.imp	11.3442	11.8732	—
CDR-HRF1-BBVI	11.3469	11.8600	—
CDR-HRF2-MLE	11.3365	11.8551	—
CDR-HRF2-BBVI.imp	11.3365	11.8550	—
CDR-HRF2-BBVI	11.3386	11.8410	—
CDR-HRF3-MLE	11.2810	11.7131	—
CDR-HRF3-BBVI.imp	11.2809	11.7126	—
CDR-HRF3-BBVI	11.2840	11.7058	—
CDR-HRF4-MLE	11.2758	11.7033	—
CDR-HRF4-BBVI.imp	11.2757	11.7034	—
CDR-HRF4-BBVI	11.2808	11.7002	—
CDR-HRF5-MLE	11.2730	11.6956	—
CDR-HRF5-BBVI.imp	11.2730	11.6956	—
CDR-HRF5-BBVI	11.2774	11.6928	**11.5369**
CDR-LCG-MLE	***11.2585***	***11.6819***	—
CDR-LCG-BBVI.imp	11.2607	11.6861	—
CDR-LCG-BBVI	11.2762	11.7023	—

Table 4.8: **Natural Stories fMRI.** CDR vs. baselines, mean-squared error. CDR results shown using 1-, 2-, 3-, 4-, and 5-parameter double-gamma hemodynamic response kernels, along with non-parametric LCG response kernels, fitted using MLE, BBVI improper, and BBVI. Best-performing models within the sets of baseline and CDR models are shown in *italics*. Best-performing overall models are shown in **bold**. Daggers (†) indicate convergence failures.

	Permutation test	Binomial test	
Baseline	p	Success rate	p
Canonical HRF	**4.0e-4*** **	0.83	**3.5e-5*** **
Interpolated	**1.0e-4*** **	1.0	**1.5e-11*** **
Averaged	**1.0e-4*** **	1.0	**1.5e-11*** **
Lanczos	**1.0e-4*** **	1.0	**1.5e-11*** **

Table 4.9: **fMRI data model comparison**. Permutation tests of improvement from CDR-G-BBVI over baselines, along with binomial tests of the probability of CDR models improving over each baseline, aggregating over training and exploratory sets (those on which all CDR models are evaluated).

ways. First, more parameterized models (HRF3+, which allow tuning of the undershoot amplitude c) tend to find a larger-magnitude undershoot component (the negative dip at the tail of the response kernel) than that of the canonical HRF. The ability to tune c and thus find deeper undershoots also corresponds to a striking improvement in both training and generalization performance (see the drop in error from HRF2 to HRF3 shown in Table 4.8). Second, the LCG model finds an early negative response consistent with prior evidence of an "initial dip" in the HRF (Yacoub et al., 2001; Röther et al., 2002; Hu and Yacoub, 2012, *inter alia*). Such a dip is outside the solution space of the parametric kernels used here and could only be discovered by LCG.

These findings have several implications. First, they reassuringly show that CDR models discover patterns that resemble the canonical HRF and are thus consistent with decades of prior research on the hemodynamic response, even using kernels with highly unconstrained solution spaces (LCG). Second they bear on prior concerns that the hemodynamic response is neither strictly stationary (Logothetis, 2003) nor strictly additive (Friston et al., 1998b, 2000), possibly undermining the usefulness of fMRI measures from long-running naturalistic exposures where confounds from e.g. response saturation might be more pronounced (Lindquist et al., 2009). These results are reassuring for naturalistic fMRI modeling since they directly support the hypothesis that the double-gamma shape continues to characterize the HRF not only in short constructed sensory experiments where it is typically studied, but also in long-running naturalistic experiments using indicators of high-level cognitive processes like language comprehension. Third, the deeper undershoot components and better fits obtained using more heavily parameterized models suggest that the canonical HRF may underestimate the size of the undershoot in the functional language network during naturalistic sentence comprehension, although further experiments would be needed to bear this out more convincingly.

Table 4.8 compares the performance of CDR against that of the baselines described in §4.3.2. As shown, among the baseline models, pre-convolution with the canonical HRF

performs quite well in terms of both in-sample and out-of-sample error (achieving the best generalization performance of any baseline model), despite its reduced number of parameters and its inability to adapt the HRF to the data. However, CDR generally outperforms all baselines on all datasets. This indicates that CDR constitutes a substantial improvement over existing methods for modeling fMRI data in naturalistic experiments.

CDR-HRF5-BBVI is selected as the representative CDR model for the fMRI dataset because it is the most flexible and best-performing of the BBVI parametric models explored here. CDR-HRF5-BBVI generalization performance on the test set is compared to that of all baseline models.[17] As shown in Table 4.8, CDR-HRF5-BBVI outperforms all baselines on the test set, and as shown in the *Permutation test* column of Table 4.9, this performance improvement is significant. As in Experiment A, the combined training and development set results from all CDR models are compared against each baseline, using a binomial test of the success rate (rate at which any CDR model outperforms a baseline model). As shown in Table 4.9, the CDR success rate over each baseline is significantly greater than chance, indicating that CDR models generally achieve better in-sample and out-of-sample error than any of the baselines, even without careful tuning on an exploratory set.

4.4 Hypothesis Testing

This section compares null hypothesis significance testing (NHST) methods using CDR models on a familiar result from psycholinguistics: *surprisal* effects in human sentence processing, which have been argued to support the existence of a predictive coding component of the language comprehension architecture that generates expectations about upcoming words (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013). In particular, this section concerns statistical tests of CDR estimates of *5-gram surprisal* against the null hypothesis of no effect. One possible approach is a credible intervals test, checking

[17] The choice of BBVI for this analysis is motivated in §3.7.4.

whether Monte-Carlo-estimated 95% credible intervals of the effect size (§3.2) in CDR for *5-gram surprisal* include zero, which implements NHST at a 0.05 level of significance. This test rejects the null hypothesis for all comparisons, as shown in Table 4.10, where the upper and lower bounds on the 95% credible interval for the effect estimate g' are always positive. However, as argued in §3.6, such a test is anticonservative in a CDR setting because of non-convexity, since the credible interval estimates only consider the local neighborhood of the mode to which the model has converged. Furthermore, such single-model tests are viewed with increasing skepticism in psycholinguistics because they are influenced by multicollinearity, leading many researchers to favor ablative model comparisons against a baseline in which the fixed effect of interest is removed (Frank and Bod, 2011). Finally, in-sample tests such as a credible intervals test or a likelihood ratio test evaluate on the training data and are therefore unable to directly diagnose overfitting. This limitation can be addressed by the use of non-parametric out-of-sample tests, such as the paired permutation test.

For the purposes of this hypothesis testing demonstration, three testing paradigms (discussed in §3.6) are applied against the null hypothesis of no effect for *5-gram surprisal* in each of the datasets explored above:

1. **Direct PT:** Ablative held-out paired permutation test (PT) of CDR models with and without a fixed effect for *5-gram surprisal*.[18]

2. **2-step LRT:** Ablative likelihood ratio test (LRT) of LME models with and without a fixed effect for 5-gram surprisal, with models fitted to the training set using predictors convolved using the full CDR model.

3. **2-step PT:** Ablative held-out paired permutation test of LME models with and without a fixed effect for 5-gram surprisal, with models fitted to the training set using

[18]Tests are based on out-of-sample likelihood rather than mean-squared error to enable consistent application to models with asymmetric error distributions, since the latter do not optimize mean squared error.

predictors convolved using the full CDR model. Tests are based on out-of-sample mean squared error.

The key difference between the direct and 2-step approaches is that the 2-step tests use LME to estimate globally optimal intercepts and linear coefficients on the convolved data. The key difference between the LRT and PT approaches to 2-step testing is that PT is a non-parametric evaluation on out-of-sample data, while LRT is a parametric evaluation on in-sample data under asymptotic guarantees about the distribution of the likelihood ratio statistic (Wilks, 1938). In order to avoid evaluating multiple CDR models on the test set, tests use the same BBVI-estimated CDR models that were selected in the previously reported baseline comparisons (BBVI inference, *shifted gamma* kernels for reading data and *HRF5* kernels for fMRI data). LME models in 2-step LRT tests are fitted to data convolved using the full CDR model as a preprocess (see §3.6 for details). To facilitate convergence, the LME structure is simplified by using uncorrelated random intercepts and slopes (Bates et al., 2015). Since these analyses are primarily for demonstration purposes, they seek uniformity within testing procedures across datasets, and no further steps are taken to address LME convergence problems in 2-step tests, although in practice it is recommended to simplify LME models until convergence is obtained before using them in scientific tests (Barr et al., 2013). Out-of-sample permutation tests use likelihood difference (direct PT) or mean squared error difference (2-step PT) as the test statistic. The exploratory and test sets in each corpus are joined to create the PT evaluation set.

Dataset	Response	Error	g' Mean	g' 2.5%	g' 97.5%	p-value Direct PT	p-value 2-Step LRT	p-value 2-Step PT
Natural Stories	ms	Gaussian	0.729	0.706	0.753	1.0	**2.2e-16*****†	1.0†
Natural Stories	ms	sinh-arcsinh	2.47	2.42	2.53	1.0	**2.2e-16*****†	0.44†
+Natural Stories	log-ms	Gaussian	0.011	0.011	0.011	**3.0e-4*****	**2.2e-16*****	1.0
Natural Stories	log-ms	sinh-arcsinh	0.009	0.009	0.009	**1/0e-4*****	**2.2e-16*****	1.0
Dundee (FP)	ms	Gaussian	3.52	3.32	3.71	**1.0e-4*****	**2.6e-14*****†	0.074†
Dundee (FP)	ms	sinh-arcsinh	1.75	1.69	1.80	**1.0e-4*****	**3.1e-8*****†	1.0†
Dundee (FP)	log-ms	Gaussian	0.012	0.011	0.013	**2.0e-4*****	**2.6e-14*****	**0.016***
Dundee (FP)	log-ms	sinh-arcsinh	0.011	0.010	0.011	**1.0e-4*****	**2.9e-12*****	0.50
Dundee (GP)	ms	Gaussian	5.75	5.42	6.05	**1.0e-4*****	**4.0e-13*****†	0.99†
Dundee (GP)	ms	sinh-arcsinh	1.93	1.88	1.98	**1.0e-4*****	**8.7e-8*****†	0.39†
+Dundee (GP)	log-ms	Gaussian	0.018	0.017	0.019	**2.0e-4*****	**6.9e-15*****	0.52
Dundee (GP)	log-ms	sinh-arcsinh	0.014	0.013	0.015	**1.0e-4*****	**9.7e-11*****	0.58
Dundee (SP)	ms	Gaussian	3.88	3.78	3.98	**1.0e-4*****	**2.2e-16**†	0.76†
Dundee (SP)	ms	sinh-arcsinh	0.703	0.592	0.816	**1.0e-4*****	1.0†	0.42 †
Dundee (SP)	log-ms	Gaussian	0.009	0.008	0.009	**1.0e-4*****	**1.7e-13*****	0.38
Dundee (SP)	log-ms	sinh-arcsinh	0.006	0.006	0.007	**1.0e-4*****	**9.0e-12*****	0.41
fMRI	BOLD	Gaussian	0.180	0.175	0.184	**1.0e-4*****	**1.0e-9*****	**1.0e-4*****
fMRI	BOLD	sinh-arcsinh	0.140	0.137	0.144	**1.0e-4*****	**2.0e-8*****	**5.0e-4*****

Table 4.10: Ablative testing results against null hypothesis of no effect for *5-gram surprisal*, using three different CDR-appropriate testing procedures. Mean and 95% credible intervals for the CDR effect estimate g' for *5-gram surprisal* are shown in the g' columns. Rejections of the null are shown in **bold**. Daggers (†) indicate convergence failures in one or both models. Plusses (+) indicate conceptual reproductions of tests in Shain (2019).

Results are shown in Table 4.10. The p-values of 1.0 observed in some cells indicate that the ablated model outperformed the full model on the evaluation set. Despite the use of simpler LME models with uncorrelated random intercepts and slopes, convergence failures affect the 2-step results for the linear response (ms) models of Natural Stories and Dundee.

The 2-step LRT test is the least conservative, rejecting the null (and supporting the existence of surprisal effects) in all models. This is unsurprising both because the likelihood ratio test is maximally powerful (Neyman and Pearson, 1933) and because it is in-sample and therefore unable to directly account for external validity, unlike PT, which is based on generalization quality. Generalization-based tests like PT are arguably more likely to favor replicable findings than in-sample tests like LRT, since an effect that is significant by LRT on the training data but does not generalize to a different sample is of limited scientific interest. The direct PT test rejects the null for all models except Natural Stories (ms), where ablated models outperform the full model. Further exploration of this exception is left to future research. Nonetheless, direct PT results overall support the existence of surprisal effects on all three kinds of experimental measures considered here. The 2-step PT test appears to be the most conservative of the testing procedures evaluated here, only rejecting the null for two out of five comparisons. These results suggest that LME models fitted to CDR-convolved data generalize less well than the underlying CDR model itself. In light of this finding and the evidence from §3.6 that CDR-estimated coefficients are near globally optimal, the added complexity of the 2-step approach may be of limited value.

In sum, using multiple testing procedures, CDR models generally reveal evidence for surprisal effects across all datasets considered here. Although all three procedures described here are appropriate for testing scientific hypotheses, the direct test may be a reasonable default because of its simplicity and in light of the combined evidence that (1) CDR-estimated coefficients are near-optimal (§3.6), (2) LME models in 2-step tests are prone to convergence problems, and (3) LME estimates in 2-step PT tend to generalize less well than CDR estimates.

4.5 General Recommendations

The foregoing results suggest certain empirically-motivated best practices for future CDR analyses of psycholinguistic time series. These best practices are used as defaults in the CDR implementation proposed here, although they can easily be overridden on a model-by-model basis as motivated by the experimental design.

Inference type: BBVI. Of the three inference types examined here (MLE, BBVI-improper, and BBVI), results show that BBVI tends to converge more quickly (§3.7.4). As shown in §4.1, they also tend to yield more conservative estimates of uncertainty. For these reasons, I suggest BBVI inference as a general default. MLE and BBVI-improper inference modes are still useful for sanity checking and sometimes obtain better error.

Kernel type: parametric. Results show a strong tendency for low-dimensional parametric response kernels to perform at least as well as high-dimensional LCG kernels in terms of synthetic IRF recovery and generalization error. Furthermore, depending on compute architecture, LCG kernels can be many times slower per iteration than parametric ones, in part because they contain many more parameters (Table 4.1). At the same time, over-constrained kernels can lead to high model bias (see e.g. *exponential* kernels fitted to *shifted gamma* responses in Simulation D). For this reason, I recommend the use of parametric kernels for research purposes whenever possible (i.e. whenever domain knowledge suggests a parametric kernel that covers the space of plausible solutions), although sanity checking the results against LCG estimates can be a useful step for datasets rich enough to support discovery of LCG models.

Convergence criterion: time-loss correlation. I have proposed and applied a CDR convergence criterion based on statistical tests for non-decreasing loss. All parametric CDR models in this study met this criterion within a reasonable number of training iterations, suggesting that it is robust and scale-independent, as argued in §3.7.2.[19] The use of a

[19] While the criterion is robust to the scale of the loss, it is sensitive to low-level training parameters like the learning rate, optimizer, and batch size. For example, the 500-iteration window used for convergence

reliable automatic stopping criterion reduces the number of experimenter degrees of freedom by eliminating the need for researchers to decide when model training has completed.

diagnosis in these experiments may lead to unnecessarily long training times at smaller batch sizes or learning rates. Users who manipulate these optimization settings should also revisit the convergence parameters in order to ensure that they are still appropriate.

Part III

Studying Human Language Processing with CDR

Chapter 5

Word Frequency and Predictability in Reading

This chapter applies CDR to answer the question: are there distinct effects of a word's frequency versus predictability in human sentence comprehension? Recent evidence implicates prediction as a major organizing principle in cognition (Bubic et al., 2010; Singer et al., 2018; Keller and Mrsic-Flogel, 2018), and psycholinguists have long studied the role of prediction in human sentence processing and its relation to other comprehension mechanisms (Marslen-Wilson, 1975; Kutas and Hillyard, 1984; MacDonald et al., 1994; Tanenhaus et al., 1995; Hale, 2001; Norris, 2006; Levy, 2008; Frank and Bod, 2011). Some prominent theories of word recognition claim that ease of lexical access is modulated by the strength of a word's representation in memory, independently of contextual factors that guide prediction (Seidenberg and McClelland, 1989; Coltheart et al., 2001; Harm and Seidenberg, 2004). Other theories hold that apparent effects of frequency are underlyingly effects of predictability (Norris, 2006; Levy, 2008; Rasmussen and Schuler, 2018).

A number of studies using constructed stimuli that factorially manipulate word frequency and predictability have found separable additive effects of each, suggesting distinct influences on lexical processing (see Staub, 2015 for a review). This study examines the generalizability of these findings to typical sentence comprehension by searching for separable effects of frequency and n-gram predictability using CDR models fitted to three large naturalistic reading corpora: Natural Stories (Futrell et al., 2020), Dundee (Kennedy et al., 2003), and UCL (Frank et al., 2013). While results show evidence of both frequency and predictability effects in isolation, they show no effect of frequency over predictability and thus do not support the existence of separable effects. They are instead consistent with

either (1) an account of apparent frequency effects as epiphenomena of predictive processing (Norris, 2006; Levy, 2008) or (2) a more circumscribed role for frequency effects in naturalistic reading than constructed experiments suggest.

5.1 Background and Related Work

5.1.1 Frequency and Predictability in Human Sentence Processing

It has long been recognized that low-frequency words are harder to process (Inhoff and Rayner, 1986). For example, in a neutral context, the more frequent *bottle* should on average be processed more quickly than the less frequent *kettle*:

I have a **bottle**. (5.1)

I have a **kettle**. (5.2)

However, context can dramatically alter these patterns by changing words' predictability (Ehrlich and Rayner, 1981):

the pot calling the **bottle** black (5.3)

the pot calling the **kettle** black (5.4)

Some models of word recognition (Seidenberg and McClelland, 1989; Coltheart et al., 2001; Harm and Seidenberg, 2004) posit a context-independent lexical retrieval mechanism, distinct from any mechanisms for predictive coding, with processing cost proportional to the strength of a word's representation in memory (a function of lexical frequency). Such a view predicts separable effects of frequency and predictability in human language comprehension. Other models (Hale, 2001; Norris, 2006; Levy, 2008; Rasmussen and Schuler, 2018) posit no such context-independent retrieval mechanism, and instead propose a unified comprehension

mechanism that incrementally reallocates resources between possible interpretations of the unfolding sentence, with processing cost proportional to the amount of information (resource reallocation) contributed by each new word. Such a view predicts no separable effects of frequency and predictability because lexical frequencies are subsumed into the incremental probability model.

Consistently with the first hypothesis, previous studies have shown separable additive effects of frequency and predictability by factorially manipulating corpus frequency and cloze predictability (Rayner et al., 2004a; Ashby et al., 2005; Gollan et al., 2011; Staub and Benatar, 2013, see Staub, 2015 for a review). However, cloze estimates poorly distinguish degrees of low contextual probability (Smith and Levy, 2013), and constructed stimuli, while affording direct control over linguistic variables, may fail to reflect the typical distributional characteristics of the language, lack context, and/or inadvertently trigger suspension of the usual processes of pragmatic inference due to the absence of an overarching discourse (Demberg and Keller, 2008; Hasson and Honey, 2012; Shain et al., 2018). It is therefore not yet clear whether frequency and predictability effects can be separated in a more realistic setting.

5.1.2 The Naturalistic Experimental Paradigm

As discussed in §2.1.3, concerns about the ecological validity of constructed stimuli can be addressed by the use of naturalistic stimuli (e.g. stories, newspaper articles, persuasive pieces, etc.). Naturalistic experiments are therefore an important complement to constructed experiments in the study of cognitive processes (Hasson and Honey, 2012).

However, naturalistic experiments introduce their own challenges. Without the ability to factorially manipulate frequency and predictability, naturalistic studies must confront the natural collinearity between these two variables in ordinary language (Demberg and Keller, 2008). Furthermore, because naturalistic stimuli do not define a critical region of the stimulus, responses are generally modeled word-by-word (Demberg and Keller, 2008;

	Effect estimate (log-ms)							
Corpus	SentPos	Trial	Rate	WordLen	SaccLen	PrevFix	Unigram	5-gram
Natural Stories	0.0098	-0.0216	-0.3069	—	—	0.0158	-0.0018	0.0174
Dundee	-0.0085	-0.0052	-0.0277	0.0068	-0.0021	-0.0178	-0.0067	0.0117
UCL	0.0524		-0.1330	0.0023	0.0221	0.0778	0.0005	0.0184

Table 5.1: Effect estimates in log-ms by corpus, computed as the IRF integral over the longest time offset seen in training. Following psycholinguistic convention, unigrams and 5-grams have opposite sign (log prob vs. surprisal). In UCL, *sentence position* and *trial* are identical (sentences were shuffled).

Frank and Bod, 2011; Smith and Levy, 2013; van Schijndel and Schuler, 2015). It is standard psycholinguistic practice to do so through ablative likelihood ratio testing (LRT) of linear mixed effects regression (LMER) models (Bates et al., 2015) fitted to the dependent variable of interest (e.g. fixation duration) (Demberg and Keller, 2008; Frank and Bod, 2011; van Schijndel and Schuler, 2015; Shain et al., 2016). However, this approach has important disadvantages. First, delayed effects in human sentence processing can have a severe impact on results obtained using this approach (Chapter 2). Second, LRT implicitly evaluates on in-sample data, making it challenging to diagnose overfitting and to assess external validity (Vasishth et al., 2018). This can be addressed through out-of-sample non-parametric tests, such as those explored in §4.4.

5.2 Experimental Setup

This study seeks to complement constructed stimulus experiments by searching for separable effects of frequency and predictability during naturalistic reading, using methods designed to address the challenges of Section 5.1.2. The problem of temporal diffusion is addressed by using CDR models rather than LMER. The problem of external validity is addressed by using held-out paired permutation testing rather than LRT, thus basing the hypothesis test directly on generalization error. The possibility that cloze probabilities are poor estimates of predictability for low-frequency words is addressed by operationalizing predictability as 5-gram surprisal generated by a large-vocabulary statistical language model. The natural

Corpus	ρ
Natural Stories	-0.78
Dundee	-0.73
UCL	-0.74

Table 5.2: Pearson's correlation between *5-gram surprisal* and *unigram log probability* by corpus.

		Corpus		
Comparison	Pooled	Natural Stories	Dundee	UCL
5-gram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
Unigram only vs. baseline	**0.0001*** **	**0.0001*** **	**0.0001*** **	**0.0001*** **
5-gram + Unigram vs. Unigram-only	**0.0001*** **	**0.0001*** **	0.0626	**0.0006*** **
5-gram + Unigram vs. 5-gram-only	0.1515	0.1831	0.0105	0.1491

Table 5.3: Held-out paired permutation testing results, both pooled (left) and by corpus (right).

collinearity of frequency and predictability is addressed through the use of large-scale data that should permit subtle differentiation of collinear effects. Taken together, the corpora examined in this study contain over one million fixations generated by 243 human subjects. Although there is a large-magnitude correlation between *unigram log probability* (frequency) and *5-gram surprisal* (predictability) in these corpora, as shown in Table 5.2, synthetic experiments show that CDR can faithfully identify models from much smaller data than that used here, even when all predictors are correlated at the 0.75 level (Chapter 4). Given the size of the data, failure to distinguish effects of frequency and predictability would raise doubts about the existence of such a separation in naturalistic reading.

5.2.1 Statistical Procedure

CDR models are fitted separately to each of the Natural Stories (Futrell et al., 2018), Dundee (Kennedy et al., 2003), and UCL (Frank et al., 2013) corpora.[1]

[1] Natural Stories is a self-paced reading corpus containing 848,768 word fixations from 181 subjects reading narrative and informational texts. Dundee is an eye-tracking corpus containing 260,065 word fixations from 10 subjects reading newspaper editorials. UCL is an eye-tracking corpus containing 53,070 fixations from 42 subjects reading sentences taken from novels by amateur authors. Although the sentences in UCL were randomized and presented in isolation — and therefore subject to some of the concerns about constructed stimuli raised in Section 5.1 — they are included here because the stimuli are naturally occurring

CDR Models used variational inference to fit the means and variances of independent normal posterior distributions over all model parameters assuming an improper uniform prior. Convolved predictors used the three-parameter ShiftedGamma impulse response function (IRF) kernel:

$$f(x;\alpha,\beta,\delta) = \frac{\beta^{\alpha}(x-\delta)^{\alpha-1}e^{-\beta(x-\delta)}}{\Gamma(\alpha)} \tag{5.5}$$

Posterior means for the IRF parameters were initialized at $\alpha = 2$, $\beta = 5$, and $\delta = -0.2$, which defines a decreasing IRF with peak centered at $t = 0$ that decays to near-zero within about 1s. Models were fitted using the Adam optimizer (Kingma and Ba, 2014) with Nesterov momentum (Nesterov, 1983; Dozat, 2016), a constant learning rate of 0.01, and minibatches of size 1024. For computational efficiency, histories were truncated at 128 timesteps. Prediction from the network used an exponential moving average of parameter iterates with a decay rate of 0.999, and models were evaluated using *maximum a posteriori* estimates obtained by setting all parameters to their posterior means.[2] Convergence was visually diagnosed.

Following previous investigations of this question (Rayner et al., 2004a; Ashby et al., 2005; Gollan et al., 2011, *inter alia*), frequency is estimated from corpus statistics — in this case, KenLM (Heafield et al., 2013) unigram models trained on the Gigaword 3 corpus (Graff and Cieri, 2003). Unlike previous studies using close estimates of predictability (Rayner et al., 2004a; Ashby et al., 2005; Gollan et al., 2011, *inter alia*), predictability is statistically estimated, again using KenLM models (5-gram) trained on Gigaword 3. This is both because (1) cloze norming all words contained in thousands of naturalistic sentences is prohibitive and (2) statistical language models trained on large data can more reliably differentiate low probability continuations (Smith and Levy, 2013). Following recent work on

rather than constructed for a particular experimental purpose. The UCL results replicate the overall pattern of significance (Table 5.3), and excluding them has no impact on the overall results.

[2]Since all parameters have independent normal distributions in the variational posterior, the law of large numbers guarantees that samples from the posterior converge in probability to the posterior mean.

prediction effects in naturalistic sentence comprehension (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013), predictability estimates are encoded as *surprisal* by negating the 5-gram log probabilities.

ShiftedGamma impulse response functions are assumed for each of these variables, as well as for the nuisance variables *word length*, *saccade length* and an indicator variable for whether the previous word was fixated.[3] To capture trends in the response at different timescales, the models also include linear effects for the word's index in the sentence (*sentence position*) and document (*trial*). In addition to the intercept, the models contain a convolved intercept (*rate*) designed to capture effects of stimulus timing. The response used in all corpora is log fixation duration (go-past for eye-tracking).[4] Outlier filtering is performed in each corpus following the procedures described in Chapter 4.

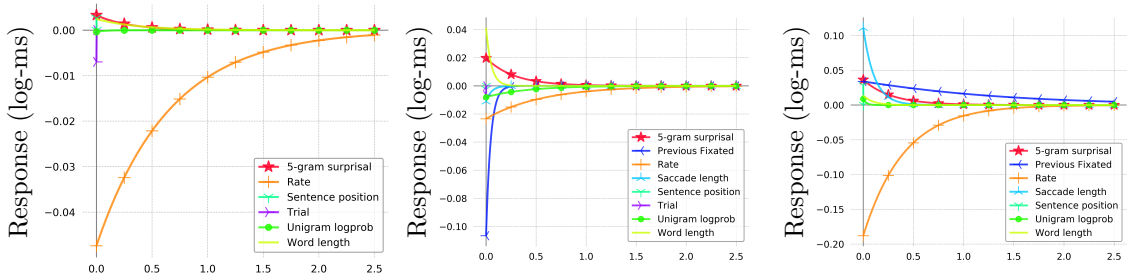
Approximately half the data in each corpus is used for training, with the remaining half reserved for held-out evaluation. Models include by-subject random intercepts as well as by-subject random slopes and impulse response parameters for each predictor.[5] Held-out hypothesis testing uses a "diamond" ablative structure first ablating fixed effects for *5-gram surprisal* and *unigram log probability* individually and then ablating both. All random effects are retained in all models. Comparisons use paired permutation tests of the by-item losses on the evaluation set, pooling across all corpora.[6] Note that the non-parametric permutation test permits this pooling procedure to unify the models from all three corpora into a single test, since (unlike LRT) permutation testing supports out-of-sample comparison. Data processing was performed using the ModelBlocks toolchain (van Schijndel and Schuler, 2013), available at `https://github.com/modelblocks/modelblocks-release`. Model fitting was performed using the CDR software library (`https://github.com/coryshain/`

[3]The variables *saccade length* and *previous was fixated* are only used for eye-tracking since they are not relevant to self-paced reading.

[4]The overall pattern of significance does not change when first-pass durations are used.

[5]By-word random intercepts are not included because of their potential to subsume frequency effects.

[6]To correct for different error variances, errors are rescaled by the joint standard deviation of the errors from the full and ablated models by corpus.



(a) Natural Stories (b) Dundee (c) UCL

Figure 5.1: IRF estimates by corpus

`cdr`). See the citations above for data access instructions.

5.3 Results

Estimated impulse response shapes by corpus are plotted in Figures 5.1a–5.1c. Plotted curves describe the estimated change in the response t seconds after having observed a unit impulse of each predictor. For example, in the Dundee estimates, observing a word with one standard deviation of *5-gram surprisal* (red curve) is expected to increase reading time by about 0.04 log-ms instantaneously, and by about 0.01 log-ms at a subsequent word observed 0.5s later. Positive IRFs (curves above 0) mean that predictors are estimated to increase reading time (and, by assumption, comprehension difficulty), and negative IRFs (curves below 0) mean that predictors are estimated to decrease reading time. For more detailed psycholinguistic interpretation of IRF estimates like these, see Chapter 4.

Overall effect estimates from the full models are presented in Table 5.1 and pooled statistical comparisons are presented in the *Pooled* column of Table 5.3. If predictability and frequency effects are additive, all four comparisons in Table 5.3 should be significant. As shown, this is not the case. There is evidence that both frequency (*unigram log probability*) and predictability (*5-gram surprisal*) in isolation reliably index processing difficulty, as shown by the significance of both effects over the baseline. However, when the effects are

compared to each other, predictability explains significantly more variance than frequency but not vice versa.

This general pattern of results further obtains for each corpus individually, as shown by the *Corpus* column breakdown in Table 5.3. One minor exception is that neither predictability nor frequency improves significantly over the other in Dundee.[7] The Dundee results are nevertheless consistent with an interpretation in which frequency and predictability do not index distinct processing phenomena and inconsistent with an interpretation in which they do. These results thus provide no evidence of separable frequency and predictability effects, whether the corpora are considered together or individually.

5.4 Discussion

As described in Section 5.3, results show no evidence of separable effects of frequency and predictability in naturalistic reading. One possible explanation for this outcome is that 5-gram surprisal tracks human prediction effort better than cloze probabilities, in part because cloze probabilities are less reliable for infrequent words. Although countervailing evidence exists in the literature (e.g. Smith and Levy, 2011 found effects of cloze but not n-gram probabilities in human reading times), in general this evidence is based on weak statistical competitors to cloze (e.g. Smith and Levy, 2011 used tri-grams). By contrast, recent trends in cognitive modeling point toward a correlation between the linguistic and psycholinguistic performance of language models, such that more powerful models with lower perplexity also tend to correlate more strongly with measures of cognitive effort (Goodkind and Bicknell, 2018; van Schijndel and Linzen, 2018). This suggests that apparent frequency effects may arise in part from poor estimates of predictability. Note that by using 5-gram surprisal rather than more powerful neural language models (Jozefowicz et al., 2016), the analysis described in this study is conservative in its attribution of variance to predictability. The

[7] The p-value of 0.0105 observed for frequency over predictability does not achieve significance at the 0.05 level under 6-way Bonferroni correction (2 variables $\times$ 3 corpora).

failure of frequency is thus all the more compelling, since replacing 5-gram surprisal with surprisals obtained from more powerful language models would be unlikely to increase the explanatory power of frequency.

Another potential explanation for the lack of separable effects of frequency and predictability is the use of naturalistic rather than constructed stimuli. Neuroscientific evidence shows that domain-general executive control regions activate during the processing of some artificially constructed language stimuli (Kaan and Swaab, 2002; Kuperberg et al., 2003; Novick et al., 2005; January et al., 2009) but fail to activate during the processing of naturalistic stimuli (Blank and Fedorenko, 2017). Such results have led some to argue that artificially constructed experimental stimuli may increase general cognitive load by coercing comprehension into problem solving, thereby engaging mechanisms that play little role in everyday sentence processing (Campbell and Tyler, 2018, Wehbe et al., in prep; Diashek et al., in prep). It is possible that the language comprehension mechanisms that implement linguistic prediction (Chapter 6) are relatively less engaged while domain general executive control mechanisms are relatively more engaged during the processing of constructed stimuli presented without context, perhaps suppressing the influence of preceding words on participants' reading behavior. Further investigation is needed in order to explore this hypothesis.

In any case, it is a statistical truism that negative results do not motivate acceptance of the null hypothesis. Thus, it is possible that frequency effects exist in naturalistic reading but are too small to be detected here. Nevertheless, the failure to find frequency effects in large naturalistic data indicates that any such effects are greatly attenuated in the processing of naturalistic texts in comparison to the processing of constructed stimuli, which circumscribes the importance that any such effects might have in driving comprehension effort during typical reading.

5.5 Conclusion

This chapter explored whether effects of word frequency and predictability are distinguishable in naturalistic sentence processing. Despite the size of the combined dataset, results showed no evidence of separable effects in naturalistic reading, contrary to previous findings of separable effects in studies using constructed stimuli. This investigation thus shows no evidence of a distinct, context-independent lexical retrieval mechanism modulated by strength of memory representation (Seidenberg and McClelland, 1989; Coltheart et al., 2001; Harm and Seidenberg, 2004), and instead favors a view in which sentence processing effort is driven by a mechanism that incrementally reallocates resources between competing interpretations, subsuming any effects of raw lexical frequency (Norris, 2006; Levy, 2008; Rasmussen and Schuler, 2018). The discrepancy between constructed and naturalistic experimental settings presents a puzzle for our understanding of the mental processes that underlie human language comprehension, and is perhaps linked to recent evidence that artificially constructed linguistic stimuli can spuriously engage non-linguistic executive mechanisms by increasing general cognitive load as compared to naturalistic settings (Blank and Fedorenko, 2017; Campbell and Tyler, 2018). Further investigation into the precise sources of the discrepancy may shed new light on the interplay between prediction and memory in human sentence processing.

Chapter 6

fMRI Evidence of Domain-Specific, Structure-Sensitive Prediction During Naturalistic Language Processing

This study uses CDR to probe the role of prediction in brain responses to naturalistic language stimuli. In the domain of language comprehension, various results show that listeners and readers actively predict upcoming words and structures (e.g. Kutas and Hillyard, 1984; MacDonald et al., 1994; Tanenhaus et al., 1995; Rayner et al., 2004a; Frank and Bod, 2011; Smith and Levy, 2011, 2013; Staub and Benatar, 2013; Frank et al., 2015; Kuperberg and Jaeger, 2016). However, the cognitive and neural mechanisms that support predictive language processing are not well understood. Under one widely held view, predictive language processing is implemented by domain-general executive (inhibitory control and working memory) resources. This perspective receives support from numerous studies showing that prediction effects during language comprehension are absent or less pronounced for populations with reduced executive resources, such as children, older individuals, and non-native speakers (e.g. Federmeier et al., 2002; Federmeier and Kutas, 2005; Dagerman et al., 2006; Federmeier et al., 2010; Mani and Huettig, 2012; Wlotko and Federmeier, 2012; Martin et al., 2013; Kaan, 2014; Mitsugi and MacWhinney, 2016; Gambi et al., 2018; Payne and Federmeier, 2018, cf. Dave et al., 2018; Havron et al., 2019). Furthermore, several neuroimaging studies have reported sensitivity to linguistic manipulations in what appear to be cortical regions thought to support domain-general executive function (e.g. Kaan and Swaab, 2002; Kuperberg et al., 2003; Novick et al., 2005; Rodd et al., 2005; Novais-Santos et al., 2007; January et al., 2009; Peelle et al., 2009; Rogalsky and Hickok, 2011; Nieuwland et al., 2012; Wild et al., 2012; McMillan et al., 2012, 2013), suggesting that such regions

may also be implicated in language processing, including perhaps prediction. These results have led some to conclude that predictive coding for language is implemented by domain-general executive control resources (Linck et al., 2014; Huettig and Mani, 2016; Pickering and Gambi, 2018; Strijkers et al., 2019).

However, this interpretation is subject to several objections. First, most prior work on linguistic prediction has relied on behavioral and electrophysiological measures which are well suited for identifying global response patterns but cannot spatially localize the source of these effects in the brain to a certain functional region or network (Mather et al., 2013, e.g.). Second, the (alleged) between-population differences in prediction noted above are consistent with accounts that do not directly invoke executive resources, including (1) possible qualitative differences between populations in the kind of information that is being predicted and the consequent need for population-specific norms to detect prediction effects, or (2) differences in how often predictions are correct, which may modulate the likelihood of engaging in predictive behavior (Ryskin et al., 2020). And third, past studies that did employ neuroimaging tools with high spatial resolution and consequently reported linguistic prediction responses — typically neural response increases for violations of linguistic structure — localized to executive control regions (e.g. Newman et al., 2001; Kuperberg et al., 2003; Nieuwland et al., 2012; Schuster et al., 2016) may have been influenced by task artifacts; indeed, some have argued that artificially constructed laboratory stimuli and tasks increase general cognitive load in comparison to naturalistic language comprehension (e.g. Blanco-Elorrieta and Pylkkänen, 2017; Blank and Fedorenko, 2017; Campbell and Tyler, 2018; Wehbe et al., 2020; Diachek et al., 2020). To ensure that findings from the laboratory paradigms truly reflect the cognitive phenomenon of interest, it is important to validate them in more naturalistic experimental settings that better approximate the typical conditions of human sentence comprehension (Hasson and Honey, 2012; Hasson et al., 2018).

Despite the growing number of fMRI studies of naturalistic language comprehension

(e.g. Speer et al., 2007; Yarkoni et al., 2008; Speer et al., 2009; Whitney et al., 2009; Wehbe et al., 2014; Hale et al., 2015; Henderson et al., 2015; Huth et al., 2016; Sood and Sereno, 2016; Brennan, 2016; Desai et al., 2016; de Heer et al., 2017; Dehghani et al., 2017; Bhattasali et al., 2018), only a handful have directly investigated effects of word predictability (Willems et al., 2015; Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017), a well-established predictor of behavioral measures in naturalistic language comprehension (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013; van Schijndel and Schuler, 2015). These previous naturalistic studies of linguistic prediction effects in the brain using estimates of prediction effort such as surprisal (Hale, 2001; Levy, 2008), the negative log probability of a word given its context, or entropy (Hale, 2006), an information-theoretic measure of the degree of constraint placed by the context on upcoming words have yielded mixed results on the existence, type, and functional location of such effects. For example, of the lexicalized and unlexicalized (part-of-speech) bigram and trigram models of word surprisal explored in Brennan et al. (2016), only part-of-speech bigrams positively modulated neural responses in most regions of the functionally localized language network. Lexicalized bi- and trigrams and part-of-speech trigrams yielded generally null or negative results (16 out of 18 comparisons). By contrast, Willems et al. (2015) found lexicalized trigram effects in regions typically associated with language processing (e.g., anterior and posterior temporal lobe). In addition, Willems et al. (2015) and Lopopolo et al. (2017) found prediction effects in regions that are unlikely to be specialized for language processing, including (aggregating across both studies) the brain stem, amygdala, putamen, and hippocampus, as well as in superior frontal areas more typically associated with domain-general executive functions like self-awareness and coordination of the sensory system (Goldberg et al., 2006). It is therefore not yet clear whether predictive coding for language relies on domain-general mechanisms in addition to, or instead of, language-specific ones, especially in naturalistic contexts.

In addition to questions about the functional localization of linguistic prediction, substantial prior work has also investigated the structure of the predictive model, seeking to shed light on the nature of linguistic representations in the mind. If effects from theoretical constructs like hierarchical natural language syntax can be detected in online processing measures, this would constitute evidence that such constructs are present in human mental representations and used to comprehend language. This position is widely supported by behavioral and electrophysiological experiments using constructed stimuli (see Lewis and Phillips, 2015, for review) and by some behavioral (Roark et al., 2009; Fossum and Levy, 2012; van Schijndel and Schuler, 2015; Shain et al., 2016), electrophysiological (Brennan and Hale, 2019) and neuroimaging (Brennan et al., 2016) experiments using naturalistic stimuli. However, other naturalistic studies reported null or negative syntactic effects (Frank and Bod, 2011; van Schijndel and Schuler, 2013, Chapter 2 contra Shain et al., 2016), or mixed syntactic results within the same set of experiments (Demberg and Keller, 2008; Henderson et al., 2016), leading some to argue that the representations used for language comprehension (in the absence of task artifacts from constructed stimuli) contain little hierarchical structure (Frank and Bod, 2011; Frank et al., 2015; Frank and Christiansen, 2018). Furthermore, the few naturalistic fMRI studies that have explored structural prediction effects have yielded inconsistent localization of these effects. For example, Brennan et al. (2016) found context-free grammar surprisal effects throughout the functional language network except in inferior frontal gyrus, whereas inferior frontal gyrus is the only region in which Henderson et al. (2016) found such effects.

The current study used fMRI to determine whether a signature of predictive coding during language comprehension increased response to less predictable words, i.e. surprisal (e.g. Smith and Levy, 2013) is primarily evident during naturalistic sentence processing in (1) the domain-specific, fronto-temporal language (LANG) network (Fedorenko et al., 2011), or (2) the domain-general, fronto-parietal multiple demand (MD) network (Duncan, 2010). The MD network supports executive functions (e.g., inhibitory control, attentional selection,

conflict resolution, maintenance and manipulation of task sets) across both linguistic and non-linguistic tasks (e.g. Duncan and Owen, 2000; Fedorenko et al., 2013; Hugdahl et al., 2015, for discussion, see: Fedorenko, 2014) and has been shown to be sensitive to surprising events (Corbetta and Shulman, 2002).

On the one hand, given that the language network plausibly stores linguistic knowledge, including the statistics of language input, it might directly carry out predictive processing. Such a result would align with a growing body of cognitive neuroscience research supporting prediction as a "canonical computation" (Keller and Mrsic-Flogel, 2018) locally implemented in domain-specific circuits (Montague et al., 1996; Rao and Ballard, 1999; Alink et al., 2010; Bubic et al., 2010; Bastos et al., 2012; Wacongne et al., 2011, 2012; Singer et al., 2018). This hypothesis is also supported by prior findings of linguistic prediction effects in portions of the language network (Bonhage et al., 2015; Willems et al., 2015; Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017; Matchin et al., 2018). On the other hand, given that the MD network has been argued to encode predictive signals across domains and relay them as feedback to other regions (Strange et al., 2005; Cristescu et al., 2006; Egner et al., 2008; Wacongne et al., 2011; Chao et al., 2018), it might be recruited to predict upcoming words and structures in language.

Prior fMRI studies using hand-constructed sentences to probe effects of linguistic expectation have not yielded a clear answer as to the mechanisms – language-specific vs. domain-general – that support linguistic prediction. Numerous such studies have observed responses in areas of the language network to manipulations of word predictability (Kuperberg et al., 2000; Baumgaertner et al., 2002; Kiehl et al., 2002; Friederici et al., 2003; Gold et al., 2006; Obleser et al., 2007; Dien et al., 2008; Obleser and Kotz, 2009; Bonhage et al., 2015; Schuster et al., 2016; Hartwigsen et al., 2017; Matchin et al., 2018; Schuster et al., 2019). However, many studies have also reported linguistic prediction effects in frontal, parietal, and cingulate cortical regions typically associated with the MD network (Kuperberg et al., 2000; Baumgaertner et al., 2002; Gold et al., 2006; Bonhage et al., 2015;

Hartwigsen et al., 2017), as well as in other parts of the brain like the fusiform gyrus (Kuperberg et al., 2000; Gold et al., 2006) and the cerebellum (Lesage et al., 2017). Although it is certainly possible that predictive coding for language is carried out by both the LANG and the MD networks, with additional contributions from other brain areas, it is important to ensure that the foregoing results are not due to task artifacts induced by the use of artificially constructed stimuli (see §6.3), through validation of these findings in more naturalistic comprehension conditions (Hasson et al., 2018).

To distinguish the hypotheses above in a naturalistic comprehension paradigm, this study searched for neural responses in LANG vs. MD regions to the contextual predictability of words as estimated by two model implementations of surprisal: a surface-level 5-gram model and a hierarchical probabilistic context-free grammar (PCFG) model. N-gram surprisal estimates are sensitive to word co-occurrence patterns but are limited in their ability to model hierarchical natural language syntax, since they contain no explicit representation of grammatical categories or syntactic composition and have limited memory for preceding words in the sentence (in the present study, up to four preceding words). PCFG surprisal estimates, by contrast, are based on structured syntactic representations of the unfolding sentence but do not directly encode surface-level word co-occurrence patterns. Correlations between each of these measures and human neural responses would shed light on the relative importance assigned to these two information sources (word co-occurrences and syntactic structures) in computing predictions about upcoming words. Although surprisal is not the only extant measure of linguistic prediction (others include PCFG entropy, Roark et al., 2009; entropy reduction, Hale, 2006; and successor surprisal, Kliegl et al., 2006), surprisal has received extensive consideration in the experimental literature (e.g. Demberg and Keller, 2008; Frank and Bod, 2011; Fossum and Levy, 2012; Frank et al., 2015; van Schijndel and Schuler, 2015; Brennan et al., 2016; Henderson et al., 2016; Brennan and Hale, 2019). Related measures were not considered in order to avoid excessive statistical comparisons.

Note that the use of surprisal to estimate prediction effects implicitly assumes a notion

of linguistic prediction as a distributed pre-activation process, following e.g. Kuperberg and Jaeger (2016), rather than as an all-or-nothing commitment to a specific upcoming word. Thus, this study investigated the degree to which the statistics of the local lexical (n-gram) and structural (PCFG) linguistic context modulate the sentence processing response, and where in the brain this modulation occurs, leaving aside questions about the underlying mechanisms by which these effects arise: e.g. the extent to which they are active or passive, or the extent to which integrative structure-building operations (e.g. composing words into syntactic constituents, constructing dependencies, etc.) underlie the observed facilitation effects (Altmann, 1998; Hale, 2014). See §6.3 for elaboration on this point.

To avoid the problem of reverse inference from anatomy to function (Poldrack, 2006, 2011), the LANG and MD networks were functionally identified in each individual participant using an independent localizer task (Saxe et al., 2006; Fedorenko et al., 2010), and then the response of those functional regions to each estimate of surprisal was examined. Results show significant independent effects of 5-gram and PCFG surprisal in LANG, but no such effects in MD, as well as significant differences in surprisal effect sizes between the two networks. This finding supports the hypothesis that predictive coding for language is primarily carried out by language-specialized rather than domain-general cortical circuits and exploits both surface-level and structural cues.

6.1 Materials and Methods

6.1.1 General Approach

Several features set the current study apart from prior cognitive neuroscience investigations of linguistic prediction in naturalistic stimuli. First, this study used naturalistic language stimuli rather than controlled stimuli constructed for a particular experimental goal. Naturalistic stimuli improve ecological validity compared to isolated constructed stimuli, which may introduce task artifacts that do not generalize to everyday cognition (Demberg and

Keller, 2008; Hasson and Honey, 2012; Richlan et al., 2013; Schuster et al., 2016; Campbell and Tyler, 2018), and prior work indicates that naturalistic stimuli yield more reliable BOLD signals than artificial tasks (Hasson et al., 2010). Minimizing such artifacts is crucial in studies of the MD network, which is highly sensitive to task variables (Miller and Cohen, 2001; Sreenivasan et al., 2014; D'Esposito and Postle, 2015; Diachek et al., 2020).

Second, this study used participant-specific functional localization to identify regions of interest constituting the LANG and MD networks (Fedorenko et al., 2010). This approach is crucial because many functional regions do not exhibit a consistent mapping onto macroanatomical landmarks (Frost and Goebel, 2012), especially in the frontal (Amunts et al., 1999; Tomaiuolo et al., 1999), temporal (Jones and Powell, 1970; Gloor, 1997; Wise et al., 2001) and parietal (Caspers et al., 2006, 2008; Scheperjans et al., 2008) lobes, which house the language and MD networks. Due to this inconsistent functional-to-anatomical mapping, a given stereotactic coordinate might belong to the language network in some participants but to the MD network in others (Fedorenko et al., 2012; Blank et al., 2017; Fedorenko and Blank, 2020). Such inter-individual variability severely compromises the validity of both anatomical localization (Juch et al., 2005; Poldrack, 2006; Fischl et al., 2007; Frost and Goebel, 2012; Tahmasebi et al., 2012) and group-based functional localization (Saxe et al., 2006; Fedorenko and Kanwisher, 2009). In contrast, participant-specific functional localization allows this study to pool data from a given functional region across participants even in the absence of perfect anatomical alignment and is therefore better suited for the kind of questions studied here (Nieto-Castañón and Fedorenko, 2012). Both networks probed here have been extensively functionally characterized in prior work, so responses to linguistic surprisal therein can be taken to index the engagement of linguistic processing mechanisms vs. domain-general executive mechanisms (e.g. Mather et al., 2013).

Third, this study uses CDR to overcome the problems in hemodynamic response modeling that are presented by naturalistic experiments. As discussed in Chapter 4, the variable spacing of words in naturalistic language prevents direct application of discrete-time,

data-driven techniques for hemodynamic response function (HRF) discovery, such as finite impulse response modeling (FIR) or vector autoregression. Because CDR is a parametric continuous-time deconvolutional method, it can infer the hemodynamic response directly from naturalistic time series, without distortionary preprocessing steps such as predictor interpolation (cf. e.g. Huth et al., 2016). Thus, unlike prior naturalistic fMRI studies of prediction effects in language processing, the shape of the HRF is not assumed.

Fourth, unlike related studies, hypotheses were evaluated using non-parametric statistical tests of model fit to held-out (out-of-sample) data, an approach which builds external validity directly into the statistical test and should thereby improve replicability (e.g. Demšar, 2006).

Finally, to my knowledge, this is the largest fMRI investigation to date (78 subjects) of prediction effects in naturalistic language comprehension.

Seventy-eight native English speakers (30 males), aged 18-60 (M±SD = 25.8±9, Med-SIQR = 23±3), from MIT and the surrounding Boston community participated for payment. Each participant completed a story comprehension task (the critical experiment) and a functional localizer task designed to identify the language and MD networks.

6.1.2 Experimental Design

Data collection was carried out by my collaborators at the Massachusetts Institute of Technology: Evelina Fedorenko, Idan Blank, and their associates.

Participants

Of the 78 participants, 8 were left-handed per self-report or Edinburgh handedness inventory (Oldfield, 1971), and of these, only one showed a right-lateralized language network (the language networks of all other participants were left-lateralized). Responses from the remaining participant were retained in analyses for completeness and generalizability to the larger population (see Willems et al., 2014, for discussion). All participants gave informed

consent in accordance with the requirements of MITs Committee on the Use of Humans as Experimental Subjects (COUHES).

Stimuli and Tasks

Each participant completed one or more localizer tasks as well as a the critical listening task.

The sentences > non-words (S>N) localizer of Fedorenko et al. (2010) was used to identify the regions of the functional language network. This language localizer has been extensively validated by prior work across tasks (passive comprehension vs. memory probe), presentation modality (visual vs. auditory), participants, and scanning sessions within participants (Fedorenko et al., 2010; Braze et al., 2011; Vagharchakian et al., 2012; Fedorenko and Thompson-Schill, 2014; Mahowald and Fedorenko, 2016; Blank and Fedorenko, 2017, *inter alia*).

The converse of this localizer (N>S) was used to identify the regions of the MD network. This localizer is effective for MD because processing lists of non-words requires more effort than processing valid sentences, and it picks out highly similar regions to those identified by a wide range of effort-related contrasts, including non-linguistic contrasts targeting working memory and inhibitory control (Fedorenko et al., 2013; Mineroff et al., 2018).

In the main story comprehension task, participants heard between one and eight auditorily presented stories from the Natural Stories corpus (Futrell et al., 2020). Stories were recorded by two native English speakers (one male, one female) at a 44.1 kHz sampling rate, ranged in length from 4m46s to 6m29s (983-1099 words), and were played over scanner-safe headphones (Sensimetrics, Malden, MA).

Thirty participants completed the main task as a passive listening task, while the remainder answered either six or twelve comprehension questions following each story. Prior evidence indicates that the presence or absence of comprehension questions does not measurably affect the responses in the language and MD systems (Blank and Fedorenko, 2017).

Data Acquisition, Preprocessing, and Functional Localization

Data acquisition, preprocessing, and functional localization used the same equipment and procedures described in Diachek et al. (2020). Following established fMRI preprocessing steps over the spatial and temporal dimensions (Diachek et al., 2020), activity in the fROIs of interest in each participant were extracted by (1) applying broad fROI-specific binary masks computed from a large independent sample (`https://evlab.mit.edu/funcloc/download-parcels`), then (2) averaging activity in the top 10% of voxels within each mask in each participant with the largest localizer contrast (S > N for LANG or N > S for MD) according to the beta estimates of voxel-wise generalized linear models. The language network contains six fROIs: inferior frontal gyrus (IFG) and its orbital part (IFGorb), middle frontal gyrus (MFG), anterior temporal cortex (AntTemp), posterior temporal cortex (PostTemp), and angular gyrus (AngG). The MD network contains ten fROIs: posterior (PostPar), middle (MidPar), and anterior (AntPar) parietal cortex, precentral gyrus (PrecG), superior frontal gyrus (SFG), middle frontal gyrus (MFG) and its orbital part (MFGorb), opercular part of the inferior frontal gyrus (IFGop), the anterior cingulate cortex and pre-supplementary motor cortex (ACC/pSMA), and the insula (Insula).

6.1.3 Statistical Analysis

Predictor Definitions

Word predictability was estimated using an information-theoretic measure known as *surprisal* (Shannon, 1948; Hale, 2001): the negative log probability of a word given its context. Surprisal can be computed in many ways, depending on the choice of probability model. Three previous naturalistic fMRI studies (Willems et al., 2015; Brennan et al., 2016; Lopopolo et al., 2017) searched for surface-level n-gram surprisal effects, using words and/or parts of speech as the token-level representation. In addition, two previous naturalistic fMRI studies (Brennan et al., 2016; Henderson et al., 2016) probed structure-sensitive

PCFG surprisal measures (Hale, 2001; Roark et al., 2009). As discussed in §6, results from these studies failed to converge on a clear answer as to the nature and functional location of surprisal effects. In this study, the following surprisal estimates were used:

- **5-gram surprisal:** 5-gram surprisal for each word in the stimulus set from a KenLM (Heafield et al., 2013) language model with default smoothing parameters trained on the Gigaword 3 corpus (Graff et al., 2007). 5-gram surprisal quantifies the predictability of words as the negative log probability of a word given the four words preceding it in context.

- **PCFG surprisal:** Lexicalized probabilistic context-free grammar surprisal computed using the incremental left-corner parser of (van Schijndel et al., 2013) trained on a generalized categorial grammar (Nguyen et al., 2012) reannotation of Wall Street Journal sections 2 through 21 of the Penn Treebank (Marcus et al., 1993).

Models also included the control variables Sound Power, Repetition Time (TR) Number, Rate, Frequency, and Network, which were operationalized as follows:

- **Sound power:** Frame-by-frame root mean squared energy (RMSE) of the audio stimuli computed using the Librosa software library (McFee et al., 2015).

- **TR Number:** Integer index of the current fMRI sample within the current scan.

- **Rate:** Deconvolutional intercept. A vector of ones time-aligned with the word onsets of the audio stimuli. Rate captures influences of stimulus timing independently of stimulus properties (see e.g. Brennan et al., 2016).

- **Frequency:** Corpus frequency computed using a KenLM unigram model trained on Gigaword 3. For ease of comparison to surprisal, frequency is represented here on a surprisal scale (negative log probability), such that larger values index less frequent words (and thus greater expected processing cost).

- **Network:** Numeric predictor for network ID, 0 for MD and 1 for LANG.

Models additionally included the mixed-effects random grouping factors Participant and fROI. Prior to regression, all predictors were rescaled by their standard deviations in the training set except *rate* (which has no variance) and *network* (which is an indicator variable). Reported effect sizes are therefore in standard units.

Continuous-time deconvolutional regression

Naturalistic fMRI studies of language processing are an important use case for CDR, since the variable duration of spoken words prevents direct temporal alignment between word times and scanner acquisition times in order to deconvolve the HRF (Boynton et al., 1996), which is known to vary between brain regions (Handwerker et al., 2004). For this reason, this study employed CDR models with the following two-parameter HRF kernel based on the widely-used double-gamma canonical HRF (Lindquist et al., 2009):

$$h(x;\alpha,\beta) \stackrel{\text{def}}{=} \frac{\beta^{\alpha}x^{\alpha-1}e^{\frac{-1}{\beta}}}{\Gamma(\alpha)} - \frac{1}{6}\frac{\beta^{\alpha+10}x^{\alpha+9}e^{\frac{-1}{\beta}}}{\Gamma(\alpha+10)} \tag{6.1}$$

where α and β are initialized to the SPM defaults of 6 and 1, respectively. More complex kernels (e.g., that fit the amplitude of the second term, rather than fixing it at 1/6) were avoided because of their potential to overfit. Predictors in these models were given their own coefficients (which rescale h above), but the parameters α and β of h were tied across predictors within each region of interest, modeling the assumption of a fixed-shape blood oxygenation response to neural activity in a given cortical region.

The CDR models applied in this study assumed improper uniform priors over all parameters in the variational posterior and were optimized using a learning rate of 0.001 and stochastic minibatches of size 1024. Following standard practice from linear mixed-effects regression (Bates et al., 2015), random effects were L2-regularized toward zero at a rate of 1.0. Convergence was declared when the loss was uncorrelated with training time by t-test

at the 0.5 level for at least 250 of the past 500 training epochs. For computational efficiency, predictor histories were truncated at 256 timesteps (words), which yields a maximum temporal coverage in these data of 48.34s (substantially longer than the effective influence of the canonical HRF). Prediction from the network used an exponential moving average of parameter iterates (Polyak and Juditsky, 1992) with a decay rate of 0.999, and models were evaluated using maximum a posteriori estimates obtained by setting all parameters in the variational posterior to their means. This approach is valid because all parameters are independent Gaussian in the CDR variational posterior.

Model Specification

The following CDR model specification was fitted to responses from each of the LANG and MD fROIs, where italics indicate predictors convolved using the fitted HRF and bold indicates predictors that were ablated for hypothesis tests:

$$\begin{aligned}&\text{BOLD} \sim \text{TRNumber} + \textit{soundPower} + \textit{Rate} + \textit{Frequency} + \boldsymbol{\textit{5gram}} + \boldsymbol{\textit{PCFG}} \\ &+ (\text{TRNumber} + \textit{soundPower} + \textit{Rate} + \textit{Frequency} + \textit{5gram} + \textit{PCFG} \mid \text{fROI}) \\ &+ (1 \mid \text{Participant})\end{aligned}$$

The random effect by fROI indicates that the model included zero-centered by-fROI random variation in response amplitude and HRF parameters for each functional region of interest. As shown, the model also included a random intercept by participant.[1] The above model can test whether the surprisal variables help predict neural activation in a given cortical region. However, it cannot be used to compare the magnitudes of response to surprisal across networks (Nieuwenhuis et al., 2011). Therefore, differences in effects between networks were tested by fitting the combined responses from both LANG and MD using the following model specification with the indicator variable *network*:

[1]The data do not appear to support richer by-participant random effects, e.g. including random slopes and HRF shapes, since such models explained no held-out variance in early analyses, indicating overfitting.

BOLD ~ TRNumber + *soundPower* + *Rate* + *Frequency* + *5gram* + *PCFG* + Network + TRNumber:Network + *soundPower*:Network + *Rate*:Network + *Frequency*:Network + ***5gram*:Network** + ***PCFG*:Network** + (1 | fROI) + (1 | Participant)

The random effects by fROI were simplified in comparison to that of the single-network models because the *network* variable exactly partitions the fROIs. Thus ablated models can fully capture network differences as long as they have by-fROI random effects for surprisal. Indeed, initial tests showed virtually no difference in held-out likelihood between full and ablated combined models when those models included full by-fROI random effects despite large-magnitude estimates for the interactions with Network in the full model. Furthermore, the fitted parameters suggested that the by-fROI term was being appropriated in ablated models to capture between-network differences. In the full model, the *5-gram surprisal* estimates for 50% of LANG fROI and 45% of MD fROI were positive, while in the model with *5gram:network* ablated, 100% of LANG fROI and only 20% of MD fROI were positive, indicating that differences in response to *5-gram surprisal* had been pushed into the by-fROI random term. For this reason, this study used simpler models for the combined test, despite their insensitivity to by-fROI variation in HRF shape or response amplitude.

In interactions between *network* and convolved predictors, the interaction was computed following convolution but prior to rescaling with that predictors coefficient. Thus, the interaction term represents the offset in the estimated coefficient from the MD network to the LANG network, as is the case for binary interaction terms in linear regression models.

Finally, as discussed in §3.1, exact deconvolution from continuous predictors like *sound power* is not possible, since such predictors do not have an analytical form that can be integrated. Instead, *sound power* was sampled at fixed intervals (100ms), in which case the event-based CDR procedure reduces to a Riemann sum approximation of the continuous convolution integral. Note that the word-aligned predictors (e.g. *5-gram Surprisal*) therefore have different timestamps than *Sound Power*, and as a result the history window spans

different regions of time (up to 256 words into the past for the word-aligned predictors and up to 100ms $\times$ 256 = 25.6s of previous *Sound Power* samples).

Ablative Statistical Testing

In order to avoid confounds from (1) collinearity in the predictors and/or (2) overfitting to the training data, this study followed a standard testing protocol from machine learning of evaluating differences in prediction performance on out-of-sample data using ablative non-parametric paired permutation tests for significance (Demšar, 2006). This approach can be used to assess the presence of an effect by comparing the prediction performance of a model that contains the effect against that of an ablated model that does not contain it. Specifically, given two pre-trained nested models, out-of-sample by-item likelihoods are computed from each model over the evaluation set and used to construct an empirical p value for the likelihood difference test statistic by randomly swapping by-item likelihoods n times (where n=10,000) and computing the proportion of obtained likelihood differences whose magnitude exceeded that observed between the two models. To ensure a single degree of freedom for each comparison, only fixed effects were ablated, with all random effects retained in all models.

The data partition was created by cycling TR numbers e into different bins of the partition with a different phase for each subject u:

$$\text{partition}(e, u) \stackrel{\text{def}}{=} \left\lfloor \frac{e+u}{30} \right\rfloor \mod 2 \tag{6.2}$$

assigning output 0 to the training set and 1 to the evaluation set. Since TR duration is 2s, this procedure splits the BOLD times series into 60 second chunks, alternating assignment of chunks into training and evaluation sets with a different phase for each participant. Partitioning in this way allowed these analyses to (1) obtain a model of each participant, (2) cover the entire time series, and (3) sub-sample different parts of the time series for

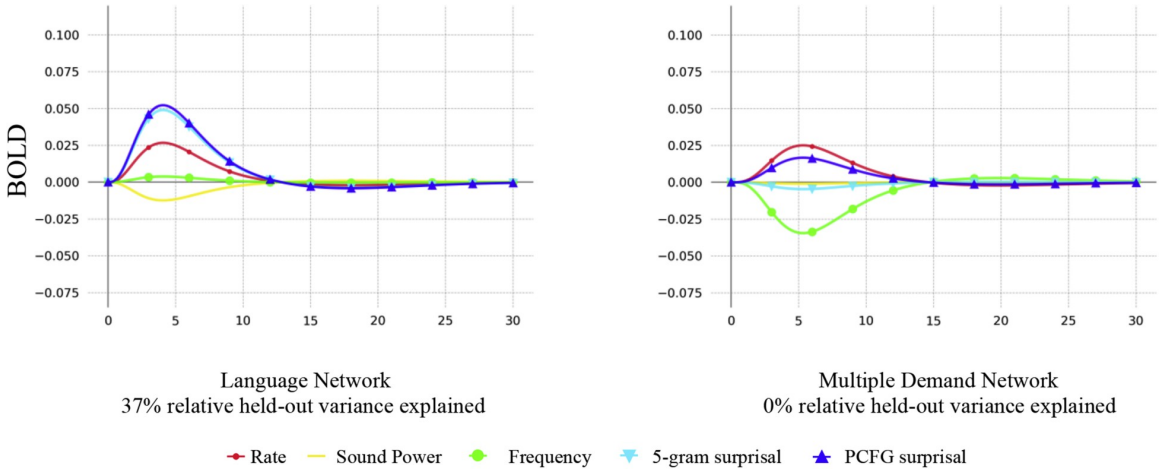


Figure 6.1: Estimated overall double-gamma hemodynamic response functions (HRFs) by network

each participant during training, while at the same time suppressing correlation between the training and evaluation responses by using a relatively long period of alternation (30 TRs or 60s).

6.1.4 Accessibility

Access instructions for software and supplementary data needed to replicate these experiments (e.g. librosa, CDR, KenLM, Gigaword 3, etc.) are given in the publications cited above. Post-processed fMRI timeseries are publicly available at the following URL: `https://osf.io/eyp8q/`. These experiments were not pre-registered.

6.2 Results

The CDR-estimated mean double-gamma hemodynamic response functions (HRFs) for the LANG and MD networks are given in Figure 6.1, the estimated HRFs by fROI in LANG regions are shown in Figure 6.2, surprisal estimates and percent variance explained by region are given in Tables 6.1 and 6.2, and population-level effect estimates (i.e. areas under the estimated HRFs) are reported in Table 6.3. MD estimates by fROI are of little relevance

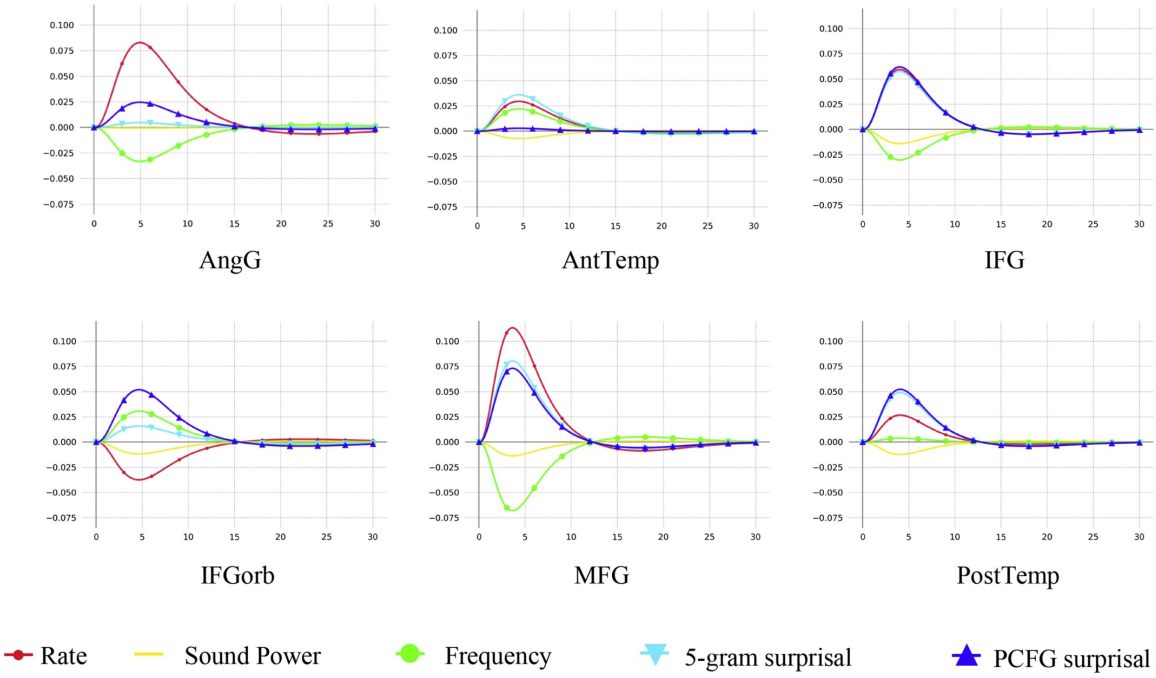


Figure 6.2: Estimated language-network HRFs by fROI

fROI	Hemisphere	5-gram estimate	PCFG estimate	% held out variance explained
AngG	L	0.030	0.156	0.0%
AntTemp	L	0.215	0.017	5.1%
IFG	L	0.287	0.309	2.2%
IFGorb	L	0.010	0.318	1.3%
MFG	L	0.382	0.346	2.3%
PostTemp	L	0.242	0.258	6.1%

Table 6.1: **LANG surprisal estimates by fROI.** Estimates given are the area under the fitted HRF. Models explain held-out variance in all regions but AngG.

fROI	Hemisphere	5-gram estimate	PCFG estimate	% held out variance explained
AntPar	L	0.102	-0.523	0.0%
IFGop	L	0.009	0.141	0.0%
Insula	L	-0.200	0.284	0.0%
MFG	L	0.074	-0.026	0.0%
MFGorb	L	-0.215	0.252	0.5%
MidPar	L	0.116	-0.051	0.0%
mPFC	L	-0.125	0.257	0.0%
PostPar	L	0.083	-0.006	0.0%
PrecG	L	0.078	0.048	0.0%
SFG	L	0.180	0.025	0.0%
AntPar	R	0.016	-0.077	0.0%
IFGop	R	-0.011	0.075	0.0%
Insula	R	-0.185	0.227	0.0%
MFG	R	0.058	-0.006	0.0%
MFGorb	R	-0.004	0.019	0.0%
MidPar	R	0.040	-0.110	0.0%
mPFC	R	-0.321	0.440	0.0%
PostPar	R	-0.312	0.434	0.0%
PrecG	R	0.034	0.118	0.0%
SFG	R	0.066	-0.034	0.0%

Table 6.2: **MD surprisal estimates by fROI.** Estimates given are the area under the fitted HRF. Models explain no held-out variance in any region except left MFGorb.

	Coefficient		
Predictor	LANG	MD	Combined
Sound Power	-0.055	-0.006	-0.003
TR Number	-0.148	0.048	-0.005
Rate	0.242	0.146	0.048
Frequency	-0.060	-0.199	-0.134
5-gram Surprisal	0.209	-0.025	0.003
PCFG Surprisal	0.235	0.097	0.038
Network	–	–	-1.32
Sound Power by Network	–	–	-0.050
TR Number by Network	–	–	-0.008
Rate by Network	–	–	0.269
Frequency by Network	–	–	0.040
5-gram Surprisal by Network	–	–	0.212
PCFG Surprisal by Network	–	–	0.193

Table 6.3: Model effect estimates.

	LANG		MD		Combined	
	%Total	%Relative	%Total	%Relative	%Total	%Relative
Ceiling 6.18%	100%	1.34%	100%	2.63%	100%	
Model (train)	3.68%	59.5%	0.75%	56.0%	1.18%	44.9%
Model (evaluation)	2.30%	37.2%	0.00%	0.00%	0.71%	27.0%

Table 6.4: Model percent variance explained compared to a "ceiling" linear model regressing against the mean response of all other participants for a particular story/fROI. "% Total" columns show absolute percent variance explained, while "% Relative" columns show percent variance explained relative to the ceiling.

Comparison	p	LL Improvement	Effect Estimate
5-gram over neither	**0.0001*** **	182	0.307
PCFG over neither	**0.0001*** **	183	0.352
5-gram over PCFG	**0.0001*** **	61	0.209
PCFG over 5-gram	**0.0001*** **	61	0.235

Table 6.5: **LANG result.** Significance in LANG by paired permutation test of log-likelihood improvement on the evaluation set from including a fixed effect for each of 5-gram Surprisal and PCFG Surprisal, over (1) a baseline with neither fixed effect and (2) baselines containing the other fixed effect only. The Effect Estimate column shows the estimated effect size from the model containing the fixed effect (i.e. the area under the estimated HRF).

Comparison	p	LL Improvement	Effect Estimate
5-gram over neither	0.137	3	0.019
PCFG over neither	1.0	-29	0.081
5-gram over PCFG	1.0	-8	-0.025
PCFG over 5-gram	1.0	-40	0.097

Table 6.6: **MD result.** Significance in MD by paired permutation test of log-likelihood improvement on the evaluation set from including a fixed effect for each of 5-gram Surprisal and PCFG Surprisal, over (1) a baseline with neither fixed effect and (2) baselines containing the other fixed effect only. A p-value of 1.0 is assigned by default to comparisons in which held-out likelihood improved under ablation. The Effect Estimate column shows the estimated effect size from the model containing the fixed effect (i.e. the area under the estimated HRF).

Comparison	p	LL Improvement	Effect Estimate
5-gram:Network over neither	**0.0001*** **	144	0.212
PCFG:Network over neither	**0.0001*** **	144	0.193
5-gram:Network over PCFG:Network	**0.0001*** **	53	0.301
PCFG:Network over 5-gram:Network	**0.0001*** **	53	0.317

Table 6.7: **Combined result**. Significance in the combined data by paired permutation test of log-likelihood improvement on the evaluation set from including a fixed interaction for each of 5-gram Surprisal and PCFG Surprisal with Network, over (1) a baseline with neither fixed interaction and (2) baselines containing the other fixed interaction only. The Effect Estimate column shows the estimated interaction size from the model containing the fixed interaction (i.e. the difference in effect estimate between LANG and MD).

Comparison	Median LL Improvement by Participant	% Participants Improved	Num Removable Participants
5-gram over neither	1.236	71.8%	19
PCFG over neither	0.732	64.1%	14
5-gram over PCFG	0.335	61.5%	7
PCFG over 5-gram	0.498	60.3%	5

Table 6.8: **Generality of LANG surprisal effects across participants**. Median likelihood improvement in LANG on the evaluation set by participant, percent of participants whose held-out predictions improved due to surprisal effects, and the number of participants with the largest held-out improvement whose data can be removed without changing the significance of the effect at a 0.05 level. Held-out likelihood improves for most participants in every comparison, and at least 5 of the most responsive participants can be removed in each comparison without changing the significance of the effect.

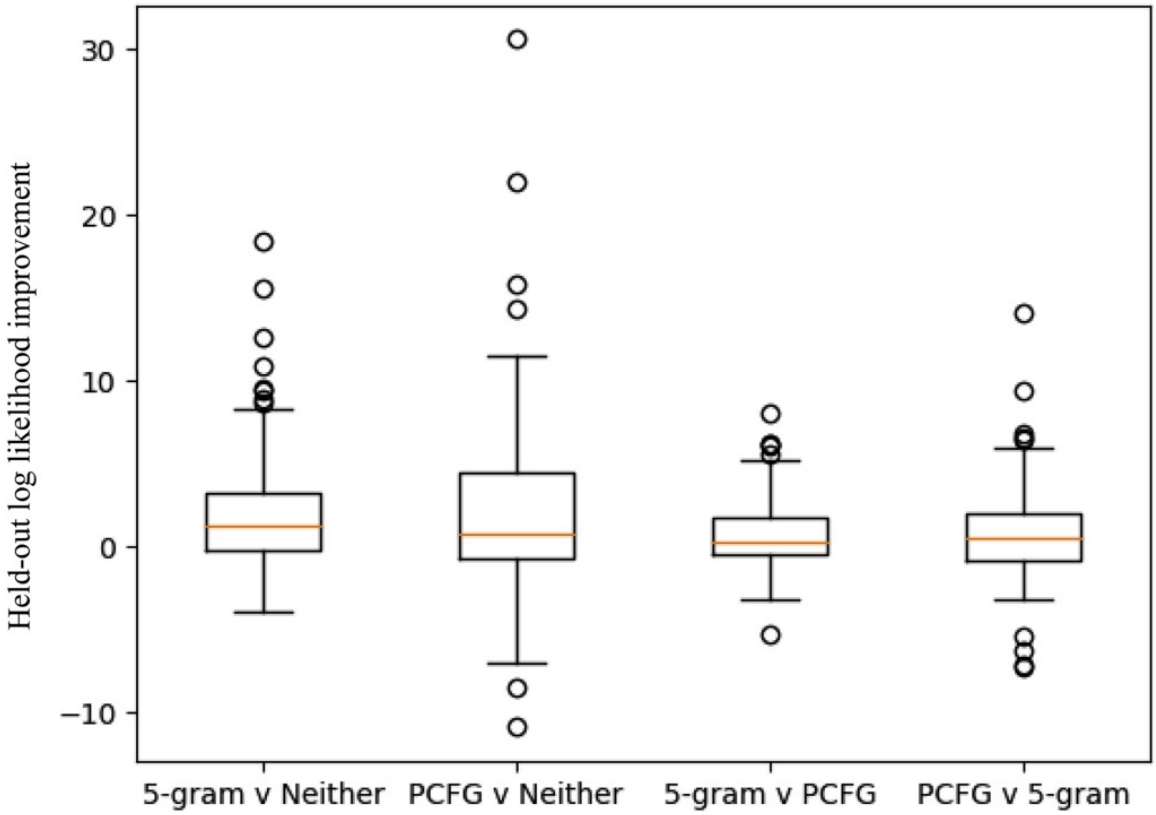


Figure 6.3: **LANG likelihood improvement by participant.** Spread of by-participant likelihood improvements in each comparison. Most improvements are positive, and effects are not driven by large positive outliers (see Table 6.8).

because they do not generalize (Tables 6.2 and 6.6). As shown, HRF shapes resemble but deviate slightly from the canonical HRF (Boynton et al., 1996) to varying degrees in each region, highlighting both consistency with HRF estimates established by prior research as well as the potential of CDR to discover subtle differences in HRF shape between cortical regions (Handwerker et al., 2004) in naturalistic data. The models find positive effects of similar strength for both *5-gram surprisal* and *PCFG surprisal* in LANG, and smaller effects of surprisal (even negative in the case of *5-gram surprisal*) in MD.

At the level of individual regions, the models explained held-out variance in all but one of the language fROIs (the exception was the AngG fROI). In contrast, the models explained no held-out variance in any but one MD fROI (the left MFGorb fROI). These two exceptions are left to future research, but overall, the results demonstrate that surprisal effects are generally present throughout the language network and generally absent throughout the MD network. The differences between the individual-network models are largely replicated in the Combined model (Table 5), where main effects represent the estimated mean response in MD while interactions with Network represent the estimated difference in mean response between LANG and MD. As shown, Combined model estimates of both *5-gram:network* and *PCFG:network* are positive and large-magnitude, indicating that the model estimates these variables to yield greater increases in neural activity in LANG over MD.

Table 6.4 reports model percent variance explained compared to a theoretical ceiling computed by regressing responses against responses from the same brain region in all other participants exposed to that stimulus. This ceiling is designed to quantify the variance that can be explained based on the stimuli alone, independently of inter-participant variation. As shown, models explain a substantial amount of the available variance in LANG. MD models explain no variance on the evaluation set, suggesting that the MD model did not learn generalizable patterns.

Because fROIs were modeled as random effects in these analyses, pairwise statistical testing of between-region differences in effect amplitude is not straightforward, and systematic

investigation of regions/subnetworks within each broader functional network is left to future work. However, a qualitative examination of the by-region estimates suggests potentially interesting functional differences within the language network (Table 6.1). In particular, the IFG, MFG, and PostTemp fROIs all responded roughly equally to both measures of surprisal. The IFGorb fROI responded more to *PCFG* than *5-gram surprisal* (an unexpected finding given that this is not the language region that is traditionally most strongly associated with syntactic processing; e.g., Friederici, 2011; Blank et al., 2016). The AngG fROI showed a similar pattern, but the models did not explain held-out variance for this fROI. And the AntTemp fROI responded more to *5-gram* than *PCFG surprisal* (a finding which bears on debates about the functional role of this brain region in language processing, see §6.3). Although the differences in effect sizes between the two surprisals are significant in each of IFGorb, AngG, and AntTemp by Monte Carlo estimated credible intervals tests, such tests are anticonservative in CDR (see §3.6). Nonetheless, they suggest that different regions of the language network might be differentially sensitive to surface-level vs. structural properties of language. The internal architecture of the language network has been long debated, and a number of proposals have been put forward (e.g. Friederici, 2011, 2012; Baggio and Hagoort, 2011; Tyler et al., 2011; Duffau et al., 2014; Ullman, 2016). However, no consensus has yet been reached about whether different regions support different aspects of language processing, and, if so, which regions support which linguistic computations (see e.g. Fedorenko et al., 2018, for discussion). Perhaps neural investigations of naturalistic language comprehension, combined with the power of the novel CDR approach and stringent statistical evaluation, can help inform this ongoing debate.

Tables 6.5–6.7 show the main finding of this study: fixed effects for *5-gram surprisal* and *PCFG surprisal* significantly improve held-out likelihood in the LANG network over a model containing neither, as well as over one another. The difference in effect size between the LANG and MD networks is statistically significant, as shown by the significant likelihood improvements yielded by interactions of the surprisal variables with *network*.

As shown in Figure 6.1, the effects signs for *frequency* in both networks are negative. The lack of a positive effect of *frequency* is not what would be expected if word frequency modulated neural activity (Staub, 2015), but it is consistent with recent naturalistic behavioral evidence against distinct effects of frequency and predictability (Chapter 5), as well as with previous theoretical claims that apparent frequency effects are underlyingly effects of predictability (Levy, 2008). Negative effects like these indicate suppression of the BOLD response and pose a challenge for interpretation (Harel et al., 2002). Prior work has suggested that such negative effects can arise from increased processing load elsewhere in the brain through hemodynamic factors ("vascular steal" Lee et al., 1995; Saad et al., 2001; Harel et al., 2002; Kannurpatti and Biswal, 2004) and/or neuronal ones such as inhibition by an attention mechanism (Smith et al., 2000; Shmuel et al., 2002, 2006). The means by which such mechanisms might give rise to negative frequency effects in these experiments are not currently clear. Since frequency effects are not central to the present research question, targeted investigation of their existence and direction is left to future research.

Figure 6.3 and Table 6.8 assess the generalizability of surprisal effects across participants. Figure 6.3 shows most by-participant improvements clustered around a positive median, without strong visual indication of large-magnitude positive outliers that might exclusively drive the effect. This intuition is quantified in Table 6.8. As shown, held-out likelihood improves for most participants in all comparisons. Furthermore, at least 5 of the most responsive participants in each comparison can be removed without changing the significance of the effect. Participant removal is a stringent criterion not only because it excludes the most responsive participants from consideration but also because it reduces the power of the permutation test by shrinking the evaluation set. These participant-level analyses demonstrate that surprisal effects in LANG are not merely driven by e.g. one or two outlier participants.

6.3 Discussion

The current study examined signatures of predictive processing during naturalistic story comprehension in two functionally distinct cortical networks: the domain-specific language (LANG) network, and the domain-general multiple demand (MD) network. Specifically, this study tested which of these networks increased their responses with lower word predictability, operationalized using both 5-gram and probabilistic context-free grammar (PCFG) surprisal. The main results, yielded by continuous-time deconvolutional regression (CDR) analysis of surprisal effects in the two networks, are shown in Tables 6.5–6.7: in LANG, both *5-gram surprisal* and *PCFG surprisal* have positive effects that yield statistically significant improvements to held-out likelihood, both over a baseline containing neither fixed effect as well as over one another. By contrast, in MD, neither surprisal effect is significant in any comparison. A direct test for a difference in surprisal effects across the two networks (Table 6.7) shows that the interactions of both surprisals with *network* are positive and statistically significant, indicating that the BOLD response to both surface-level (5-gram) and structural (PCFG) word predictability is larger in LANG than MD. These results are over a baseline that includes an effect for lexical frequency (log unigram probability), which is notable given the strong natural correlation between surprisal and frequency, both generally (Demberg and Keller, 2008) and in the current experimental materials ($r = 0.78$ overall). This finding suggests that the surprisal effects reported here are indeed driven by predictive coding and not merely by the cost of retrieving infrequent words. Together, these results demonstrate that predictive coding for upcoming words is primarily a canonical computation carried out by domain-specific cortical circuits, rather than by feedback from higher, domain-general executive control circuits, and that these predictions depend on both surface-level and structural information sources. The finding of a generalized effect of *PCFG surprisal* throughout the language network aligns with prior findings of evidence for linguistic prediction (e.g. Kuperberg et al., 2000; Baumgaertner et al., 2002; Friederici

et al., 2003; Obleser et al., 2007) and syntactic processing (e.g. Blank et al., 2016, see Zaccarella et al., 2017 for review) in these regions, but suggests that prior evidence of linguistic prediction effects in MD (e.g. Kuperberg et al., 2000; Baumgaertner et al., 2002; Gold et al., 2006; Bonhage et al., 2015; Hartwigsen et al., 2017) may have been influenced by the use of artificially constructed linguistic stimuli and/or task artifacts.

This finding bears on an ongoing discussion in cognitive neuroscience about the compartmentalization of language processing. Early investigations of the functional organization of the brain argued for the existence of neuroanatomical modules dedicated to specific linguistic functions, from lower-level perceptual and motor components of language to higher-level ones like phonological, lexical, and combinatorial syntactic and semantic processing (Broca, 1861; Dax, 1863; Wernicke, 1874; Fodor, 1983; Petersen et al., 1988; Levelt, 1989; Pinker, 1994). This position has been called into question by subsequent work stressing the distributed nature of cognition (e.g. Mesulam, 1998; Thompson-Schill et al., 2005; Blumstein and Amso, 2013), based on evidence both (1) that brain regions conventionally believed to be language-specific are also recruited for non-linguistic tasks (e.g. Dehaene et al., 1999; Stanescu-Cosson et al., 2000; Maess et al., 2001; Kaan and Swaab, 2002; Koelsch et al., 2002; Koechlin and Jubault, 2006; Hein and Knight, 2008; Blumstein, 2009; January et al., 2009), and (2) that brain regions conventionally believed to support domain-general cognitive control are also recruited for language processing, especially under difficult comprehension conditions (e.g. Kaan and Swaab, 2002; Kuperberg et al., 2003; Novick et al., 2005; Rodd et al., 2005; Novais-Santos et al., 2007; January et al., 2009; Peelle et al., 2009; Rogalsky and Hickok, 2011; Nieuwland et al., 2012; Wild et al., 2012; McMillan et al., 2012, 2013; Hsu and Novick, 2016). Although such results might raise doubts about the necessity and sufficiency of the putative language network for language processing, they are counterbalanced by rigorous non-replications of (1) the engagement of language regions in arithmetic, working memory, or cognitive control tasks (Fedorenko et al., 2011; Monti et al., 2012; Amalric and Dehaene, 2018), and (2) the engagement of cognitive control (MD) regions in language

processing (Blank and Fedorenko, 2017; Wehbe et al., 2020). Based on this evidence, some have concluded that there does indeed exist a functionally specific cortical language network (Fedorenko, 2014; Fedorenko and Thompson-Schill, 2014, see also Hagoort, 2005; Friederici et al., 2011; Matchin et al., 2014; Rogalsky et al., 2015; Matchin et al., 2017, for proposals that are compatible with the idea that at least some of the language-responsive areas are specific to language) and that MD engagement in many previous studies of language processing was induced by experimental task artifacts (Campbell and Tyler, 2018; Diachek et al., 2020; Wehbe et al., 2020).

The aforementioned debate about the compartmentalization of language processing has largely focused on controlled experimental paradigms, which are prone to induce task artifacts that confound functional differentiation of neural structures. By showing strong prediction-based functional differentiation between the LANG and MD networks during naturalistic language comprehension, the present study provides evidence that predictive coding for language is primarily carried out by language-specific rather than domain-general mechanisms.

This finding also contributes to the growing literature on predictive coding in the mammalian brain, which has recently produced evidence that neurons are tuned to predict upcoming inputs but has also primarily focused on low-level perceptual processing (Rao and Ballard, 1999; Alink et al., 2010; Bubic et al., 2010; Keller and Mrsic-Flogel, 2018; Singer et al., 2018). The present study suggests that prediction extends to high-level cognitive functions like language comprehension and is similarly implemented as a domain-specific canonical computation in regions that plausibly store linguistic knowledge (e.g. Hagoort, 2005; Fedorenko, 2014).

The finding that surprisal computed by marginalizing over syntactic structures (PCFG Surprisal) modulates the LANG response independently of surface-level n-gram surprisal is evidence that participants are indeed computing such structures during incremental sentence processing (Hale, 2001; Levy, 2008; Fossum and Levy, 2012; Rasmussen and Schuler, 2018)

and is inconsistent with previous arguments that the human sentence processing response is largely insensitive to such structures (Frank and Bod, 2011; Frank et al., 2012; Frank and Christiansen, 2018). At the same time, the finding that *5-gram surprisal* modulates the LANG response independently of *PCFG surprisal* is evidence that the human sentence processing mechanism is sensitive to word co-occurrence patterns in ways that are not well captured by a strictly context-free parser. This suggests either (1) that the human parser is not strictly context-free (see e.g., tree-adjoining grammars, Joshi, 1985; combinatory categorial grammars, Steedman, 2000; and other context-sensitive grammar formalisms for natural language), or (2) that participants track both hierarchical structure and word co-occurrence patterns separately and simultaneously when generating predictions, and that these two kinds of processes take place in overlapping brain areas. In addition, (2) is compatible with either distinct predictive mechanisms that track word-level and syntactic features separately, or with a unified mechanism that leverages both kids of cues. Evaluating these hypotheses is left to future work. However, note that the two variants of (2) can be distinguished by the functional form of effects. Separate mechanisms should make roughly additive contributions to prediction error (see e.g. Fedorenko et al., 2006), whereas a unified mechanism should primarily experience prediction difficulty when both string-level and syntactic cues are poor. Using a deep neural generalization of CDR that can estimate non-linear interactions, Chapter 8 presents some preliminary evidence in favor of the latter (unified mechanism) view. The lack of structured prediction effects in MD is of interest given prior proposals that ground structural effects in constraints on working memory (Abney and Johnson, 1991; Resnik, 1992; Rasmussen and Schuler, 2018). These theories view the processing of hierarchical language structures as a special case of a domain general capacity for hierarchic sequential prediction (Botvinick, 2007), which is at least consistent with the hypothesis that the resources recruited for prediction are also domain general (see e.g. Smith and Levy, 2013). However, to the extent that the memory resources used for prediction are also expected to activate in response to prediction error (e.g., by undergoing model

revision, Chao et al., 2018), the failure to find such a signal in MD suggests that these memory resources may also be specific to the functional language network, rather than domain general (e.g. Caplan and Waters, 1999; Matchin et al., 2017).

Estimates at the fROI level shed light on results from prior naturalistic fMRI experiments (Willems et al., 2015; Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017). The present results show strong effects of both surface-level and structural estimates of word predictability in roughly the union of left-hemisphere language regions for which such effects have been reported in prior work (e.g., temporal and inferior frontal regions). At the same time, this study did not find clear evidence of predictive coding in regions linked with the multiple demand network, like superior frontal gyrus (cf. Lopopolo et al., 2017), in part because the use of held-out significance tests helped avoid reporting MD surprisal effects that fail to generalize (e.g., left-hemisphere SFG, Table 6.2). The lack of held-out testing in earlier studies may therefore have contributed to prior findings of surprisal effects in MD regions. Finally, this study obtained significant positive effects for surprisal implementations in language regions that have previously been reported null or negative (e.g., lexicalized trigrams in IFG and posterior temporal cortex or PCFG surprisal in IFG, per Brennan et al., 2016; PCFG surprisal in the temporal lobe, per Henderson et al., 2016). It is possible that the size of the present study increased sensitivity to these effects, since studies using less data are more likely to yield sign and magnitude errors (Gelman and Carlin, 2014). The picture that emerges more clearly from the current results than from those of prior studies is of a predictive coding mechanism that is specific to the functional language network, generalized throughout it, and sensitive to both surface-level word co-occurrence patterns and hierarchical structure.

Although these analyses focused on prediction effects, language comprehension involves a good deal more than simply minimizing surprise — meanings conveyed by partially-complete words and syntactic structures are rapidly and incrementally recognized, stored, and integrated into existing knowledge representations as the discourse unfolds (Tanenhaus

et al., 1995; Altmann and Kamide, 1999). Numerous studies have probed the computations involved in storage, retrieval, and integration during human sentence comprehension (MacDonald et al., 1992; Kluender and Kutas, 1993; Gibson and Ko, 1998; Felser et al., 2003; Hsiao and Gibson, 2003; Aoshima et al., 2004; Grodner and Gibson, 2005; Lewis and Vasishth, 2005; Fiebach et al., 2005; Fedorenko et al., 2006, 2007; Rasmussen and Schuler, 2018), and several memory-based estimators of structural processing have been investigated across behavioral and cognitive neuroscience investigations, including embedding difference (Wu et al., 2010), the number of open nodes based on a particular parsing strategy (top-down, bottom-up, or left-corner; Nelson et al., 2017; Brennan and Pylkkänen, 2017), dependency locality costs (storage or integration cost from maintaining and retrieving syntactic dependencies; Gibson, 2000), and encoding or retrieval interference (i.e. processing costs in the ACT-R framework; Lewis and Vasishth, 2005). Effort due to memory storage and retrieval is plausibly distinct from effort due to reallocating resources between competing structural interpretations of the unfolding sentence (a standard interpretation of surprisal effects, e.g. Hale, 2001; Levy, 2008), and a complete account of human language processing will likely involve both prediction-based and integration-based computations (Levy et al., 2013; Levy and Gibson, 2013). Although these kinds of integration effects are outside the scope of the present study of predictive coding (though see Chapter 7), some have argued that prediction may subserve memory retrieval and therefore interact with integrative processing (Altmann, 1998). Therefore, the prediction effects reported here may, to some extent, be amenable to interpretation as effects of integration. That is, researchers who view "prediction" as a conscious lexically specific activity may view these results as evidence of conceptual pre-activation or preparedness that eases integration once a word is observed (see Ferreira and Chantavarin, 2018, for an overview of this distinction). Fuller investigation of this distinction is left to future work. For the purposes of the present study, simply note that these results indicate that any such pre-activation processes appear to be restricted to the LANG network, rather than invoking the MD network, and are strongly

correlated with probabilistic measures of word predictability.

The present emphasis on structural influences on *prediction*, rather than sensitivity to syntactic structure more generally, is a possible explanation for one apparent discrepancy between these results and those of some previous studies. In particular, results do not show evidence of *PCFG surprisal* effects in the AntTemp language fROI (the *PCFG surprisal* estimate in AntTemp is virtually 0, Figure 6.2), whereas numerous previous studies have argued for syntactic effects in left anterior temporal cortex, both using hand-constructed stimuli (Mazoyer et al., 1993; Stowe et al., 1998; Friederici et al., 2003; Vandenberghe et al., 2002; Dronkers et al., 2004; Humphries et al., 2006; Rogalsky and Hickok, 2009; Pallier et al., 2011; Brennan and Pylkkänen, 2012; Nelson et al., 2017) and naturalistic stimuli (Brennan et al., 2016; Brennan and Pylkkänen, 2017; Bhattasali et al., 2018, 2019). The role of left anterior temporal cortex in syntactic processing has been called into question by an absence of syntactic deficits in patients with anterior temporal damage (e.g. Wilson et al., 2012), and some have argued that parts of the anterior temporal lobe primarily carry out lexical and semantic processing, including perhaps semantic composition (e.g. Bemis and Pylkkänen, 2011), rather than syntactic structure building (Visser et al., 2010; Wilson et al., 2014; Lambon Ralph et al., 2017, see also Matchin et al., 2018). However, even granting that left anterior temporal cortex is implicated in syntactic processing, prior studies by and large have focused on structural measures that are arguably integrative in nature (syntactic node count, number of parser operations, etc.) or have used manipulations that are too broad to target prediction vs. integration (sentences vs. list of words or "Jabberwocky" sentences). Indeed, claims about syntactic processing in left anterior temporal cortex tend to focus on composition rather than on structured prediction. The present results thus do not preclude a role for left anterior temporal cortex in structure-building broadly construed; they simply fail to show strong evidence in this brain area of effects of structural context on word predictability. Prior studies of structured prediction effects in left anterior temporal cortex have yielded mixed results; although Brennan et al. (2016) found evidence of part-of-

speech n-gram and PCFG surprisal in anterior temporal cortex over bi- and tri-gram effects, Lopopolo et al. (2017) did not find a response to part-of-speech n-gram surprisal, and the response to syntactic PCFG surprisal in Henderson et al. (2016) was too weak to achieve significance. Prediction effects based on lexical context in left anterior temporal cortex (i.e. lexical n-grams) are better attested (Willems et al., 2015; Lopopolo et al., 2017), and some have explicitly argued that left anterior temporal cortex plays a central role in lexical-semantic prediction (Lau et al., 2016). The present findings in the AntTemp fROI (esp. large effects of *5-gram surprisal*) contribute to this debate, suggesting that lexical prediction does occur in left anterior temporal cortex (among other regions) while syntactic prediction likely occurs elsewhere. Left anterior temporal cortex may therefore be an important object of study in teasing apart predictive vs. integrative processing during language comprehension, and further investigation is warranted.

In summary, the present findings based on a large-scale naturalistic fMRI experiment support a view of linguistic prediction as implemented by domain-specific cortical circuits, sensitive to both surface-level and syntactic information sources, and generalized across the functional language network.

Chapter 7

fMRI Evidence of Domain-Specific Working Memory Retrieval During Naturalistic Language Processing

Chapter 6 reported fMRI evidence that predictive processes that underlie human language comprehension (Hale, 2001; Kuperberg and Jaeger, 2016; Levy, 2008). However, many theories of human language comprehension also posit integrative processes that retrieve and update representations in working memory (Gibson, 2000; Hawkins, 1994; Lewis and Vasishth, 2005). This chapter revisits the same fMRI dataset used in Chapter 6 in order to evaluate relationships between working memory and brain activity during naturalistic language processing.

Prior behavioral and neuroimaging experiments have supported effects of integration (Fiebach et al., 2001; Grodner and Gibson, 2005) and prediction (Bonhage et al., 2015; Frisson et al., 2005) in isolation using carefully constructed stimuli. Empirically distinguishing these two hypotheses is challenging (Levy, 2008), and many previously reported effects of syntactic processing (e.g. syntactic anomalies Osterhout and Holcomb, 1992) are consistent with both theories. Further complicating matters are poorly-understood influences of standard experimental designs in studies of human language comprehension; although studies overwhelmingly investigate syntax using carefully selected or hand-crafted sentences presented in isolation, the normal conditions of language use are richly varied, contextualized, and meaningful, and growing neuroscientific evidence indicates that artificial stimuli and tasks may engage cognitive mechanisms that are not central to language processing (Campbell and Tyler, 2018; Diachek et al., 2020; Hasson et al., 2018; Hasson and Honey, 2012). Importantly, recent naturalistic behavioral studies have generally not shown

strong evidence of memory retrieval costs (Demberg and Keller, 2008; van Schijndel and Schuler, 2013), casting doubt on theories that hypothesize syntax-driven memory operations as central to comprehension (Gibson, 2000; Hawkins, 1994; Lewis and Vasishth, 2005; McElree et al., 2003). In addition, terminological and conceptual overlap between prediction and integration presents a challenge when attempting to dissociate them (Ferreira and Chantavarin, 2018; Kuperberg and Jaeger, 2016; Levy et al., 2013).

This study pursues two central questions about the role of working memory (WM) in language comprehension. **Question 1 (Q1):** Do the mechanisms that update representations in WM operate (at least partially) independently from any mechanisms that predict upcoming words? **Question 2 (Q2):** Are the WM resources used for language comprehension shared with other domains of cognition? The first question follows from ongoing debate about memory and expectation in human sentence processing (Gibson, 2000; Levy, 2008; Lewis and Vasishth, 2005). The second follows from ongoing debate about the degree to which WM resources used in language comprehension are domain-specific (Caplan and Waters, 1999; Fiebach et al., 2001) vs. domain-general (Amici et al., 2007; Fedorenko et al., 2006, 2007; Stowe et al., 1998).

This study investigates these questions using data from a large-scale naturalistic fMRI study (Chapter 6), estimating memory demands via integration cost as hypothesized by the Dependency Locality Theory (DLT Gibson, 2000) under rigorous controls for word predictability (van Schijndel and Linzen, 2018). To probe domain specificity, influences of integration cost are examined in participant-specific functionally-localized cortical regions of the language-selective network (Fedorenko et al., 2010, LANG) on the one hand and of the domain-general multiple-demand (MD) executive control network (Duncan, 2010, 2013) on the other. Results show a strong effect of integration cost in LANG only, supporting the use of language-specialized working memory resources for incremental language comprehension. Based on these results, this chapter argues (a) that the human language processor uses syntactic cues to compose representations in working memory; (b) that this kind of

composition is active in everyday naturalistic language processing; and (c) that the working memory resources used to compose these representations are specialized for the language domain.

7.1 Materials and Methods

The materials, methods, data acquisition, preprocessing, and control predictors (*Sound Power*, *TR Number*, *Rate*, *Frequency*, and *Network*) are identical to those described in Chapter 6, with one difference: any TRs that follow the final word of each story were removed, in order to reduce the influence of response decay following the end of the stimulus. The pattern of significance reported in Chapter 6 holds under these stricter exclusion criteria.

7.1.1 Control Predictors

Because points of predicted retrieval cost may partially overlap with prosodic breaks between clauses, models additionally include two prosodic controls:

- **End of Sentence:** Indicator for whether a word terminates a sentence.
- **Pause Duration:** Length (in ms) of pause following a word, as indicated by hand-corrected word alignments over the auditory stimuli. Words that are not followed by a pause take the value 0ms.

The pattern of significance reported in Chapter 6 holds in the presence of these additional controls.

In addition, inspired by evidence that word predictability strongly influences blood oxygen level dependent (BOLD) responses in the language network, models additionally include the critical *5-gram Surprisal* and *PCFG Surprisal* predictors from Chapter 6. *PCFG* and *5-gram Surprisal* were investigated Chapter 6 because their interpretable structure permits

testing of hypotheses of interest in that study. However, their strength as language models has now been outstripped by less interpretable but better performing incremental language models based on deep neural networks (Gulordava et al., 2018; Jozefowicz et al., 2016; Radford et al., 2019). In the present investigation, predictability effects are a control rather than an object of study and are therefore not bound by the same interpretability considerations. To strengthen the case for independence of retrieval processes from prediction processes, models additionally include the following predictability control:

- **Adaptive Surprisal:** Word surprisal as computed by the adaptive recurrent neural network (RNN) of van Schijndel and Linzen (2018). This network is equipped with a cognitively-inspired mechanism that allows it to adjust its expectations to the local discourse context at inference time, rather than relying strictly on knowledge acquired during the training phase. Compared to strong baselines, results show both improved model perplexity and improved fit between model-generated surprisal estimates and measures of human reading times. Because the RNN can in principle learn both (1) the local word co-occurrence patterns exploited by 5-gram models and (2) the structural features exploited by PCFG models, it competes for variance in our regression models with the other surprisal predictors, whose effects are consequently attenuated relative to Chapter 6.

As in Chapter 6, models additionally include the mixed-effects random grouping factors *Participant* and *fROI*. The ability to treat *fROI* as a coherent unit of analysis derives from the functional localization strategy, which aligns regions functionally rather than stereotactically across participants. Prior to regression, all predictors are rescaled by their standard deviations in the training set except *Rate* (which has no variance) and the indicators *End of Sentence* and *Network*. Reported effect sizes are therefore in standard units.

7.1.2 Measuring Working Memory Involvement

There are a number of existing theories concerning the kind and relative difficulty of working memory operations involved in human sentence processing, but most focus on costs related to incremental storage and/or retrieval of intermediate representations as required by the syntax of the sentence being processed. For example, early accounts of processing effects like nesting complexity posited *storage* costs for simultaneously maintaining multiple incomplete elements (e.g. relative clauses) in memory (Miller and Chomsky, 1963). Relatedly, Hawkins (1994) posits *locality*-based processing costs proportional to the number of words required in order to recognize a syntactic constituent. An influential theory unifying these notions of storage and locality is the Dependency Locality Theory (DLT) of Gibson (2000). According to the DLT, storage cost arises from maintenance in memory of syntactic dependencies to expected future syntactic heads, and occurs in proportion to the number of such dependencies. Integration cost, by contrast, arises from retrieval from working memory of previously mentioned discourse referents in order to construct new dependencies to them, as required by the syntax of the unfolding sentence. Integration cost is proportional to the number of discourse referents that intervene in a dependency and could compete as retrieval targets, giving rise to the notion of locality: longer dependencies with more intervening discourse referents incur a larger integration cost. The DLT is a computational level (Marr, 1982) theory that does not explicitly commit to a parsing algorithm or memory structure.

Subsequent algorithmic-level theories have hypothesized related principles of working memory use in sentence comprehension. Lewis and Vasishth (2005) propose a parser grounded in the *adaptive control of thought-rational* (ACT-R) framework (Anderson, 2004). ACT-R composes representations in memory through a content-addressable retrieval operation that is subject to similarity-based interference (McElree et al., 2003; Van Dyke and Lewis, 2003), with memory representations that decay with time unless reactivated through retrieval. The decay function enforces a locality-like notion (retrievals triggered by long

dependencies will on average cue targets that have decayed more), but this effect can be attenuated by intermediate retrievals of the target. Unlike the DLT, ACT-R has no notion of active maintenance in memory (items are simply retrieved as needed) and therefore does not predict a storage cost.

Another line of research (Johnson-Laird, 1983; Resnik, 1992; van Schijndel et al., 2013; Rasmussen and Schuler, 2018) frames incremental sentence comprehension as left-corner parsing (Aho and Ullman, 1972) under a stack-based implementation of working memory. Under this view, incomplete derivation fragments representing the hypothesized structure of the sentence are assembled word by word, with working memory required to (1) push new derivation fragments to the stack, (2) retrieve and compose derivation fragments from the stack, and (3) maintain incomplete derivation fragments on the stack. For a detailed presentation of a recent instantiation of this framework, see Rasmussen and Schuler (2018). In principle, costs could be associated with any of the parse operations computed by left-corner models, as well as (1) with DLT-like notions of storage (maintenance of multiple derivation fragments on the stack) and (2) with ACT-R-like notions of retrieval and reactivation, since items in memory (corresponding to specific derivation fragments) are incrementally retrieved and updated. Unlike ACT-R, left-corner frameworks do not necessarily enforce activation decay over time, and they do not inherently specify expected processing costs.

Prior work has investigated the empirical predictions of some of these theories using computer simulations (Lewis and Vasishth, 2005; Rasmussen and Schuler, 2018) and human responses to constructed stimuli (Grodner and Gibson, 2005) and reported robust WM effects. Related work has also shown effects of dependency length manipulations in measures of comprehension and online processing difficulty (Gibson et al., 1996; McElree et al., 2003; Van Dyke and Lewis, 2003; Makuuchi et al., 2009; Meyer et al., 2013). In light of these findings, evidence from more naturalistic human sentence processing settings for working memory effects of any kind is surprisingly weak. Demberg and Keller (2008) report DLT integration cost effects in the Dundee eye-tracking corpus (Kennedy et al., 2003), but only

when the domain of analysis is restricted — overall DLT effects are actually negative (longer dependencies yield shorter reading times, also known as an *anti-locality* effect, Konieczny, 2000). van Schijndel and Schuler (2013) also report anti-locality effects in Dundee, even under rigorous controls for word predictability phenomena that have been invoked to explain anti-locality effects in other experiments (Konieczny, 2000; Vasishth and Lewis, 2006). It is therefore not yet settled how central syntactically-related working memory involvement is to human sentence processing in general, rather than perhaps being driven by the stimuli and tasks commonly used in experiments designed to test these effects (Hasson and Honey, 2012; Hasson et al., 2018; Campbell and Tyler, 2018; Diachek et al., 2020). Few prior fMRI studies of naturalistic sentence processing have investigated working memory (although some of the syntactic predictors in Brennan et al., 2016, especially syntactic node count, may be amenable to a memory-based interpretation).

In light of this, the present study includes both an exploratory and a confirmatory component. The exploratory component seeks to determine whether the signatures of working memory involvement as predicted by *any* of the above theories register during naturalistic language processing in either of the brain networks examined here (LANG and MD). The confirmatory component then evaluates a single representative WM measure on an unseen dataset, in order to assess the degree to which the effect is generalizable in LANG and/or MD.

Because these measures are being deployed on naturalistic texts, this study is constrained to consider only those WM measures that are amenable to broad-coverage implementation. Unfortunately, no broad-coverage implementation of the Lewis and Vasishth (2005) ACT-R model exists, and it cannot be evaluated here. This study therefore focuses on measures from the DLT and from left-corner theories, both of which can be deterministically inferred from a hand-corrected (Shain et al., 2018) generalized categorial grammar (Nguyen et al., 2012) reannotation of the stimulus sentences originally annotated using the Penn Treebank standard (Marcus et al., 1993).

For the DLT predictors, following prior work (Demberg and Keller, 2008), this study focuses on measures of integration (retrieval) cost. Integration cost is computed as the number of discourse referents that intervene in a backward-looking syntactic dependency, where "discourse referent" is operationalized for simplicity as any noun or finite verb. In addition, all three implementation variants of integration cost proposed by Shain et al. (2016) are considered:

- ±**V**: Verbs are more expensive. Non-finite verbs receive a cost of 1 (instead of 0) and finite verbs receive a cost of 2 (instead of 1).
- ±**C**: Coordination is less expensive. Dependencies out of coordinate structures skip preceding conjuncts in the calculation of distance, and dependencies with intervening coordinate structures assign that structure a weight equal to that of its heaviest conjunct.
- ±**M**: Exclude modifier dependencies. Dependencies to preceding modifiers are ignored.

These variants are motivated by the following considerations. First, the reweighting in +V is motivated by the possibility (1) that finite verbs may require more information-rich representations than nouns, especially tense and aspect (Binnick, 1991), and (2) that non-finite verbs may still contribute eventualities to the discourse context, albeit with underspecified tense (Lowe, 2019). As in Gibson (2000), the precise weights are unknown, and the weights used here are simply heuristic approximations that instantiate a hypothetical overall pattern: non-finite verbs contribute to retrieval cost, and finite verbs contribute more strongly than other classes.

Second, the discounting of coordinate structures under +C is motivated by the possibility that conjuncts are incrementally integrated into a single representation of the overall coordinated phrase, and thus that their constituent nouns and verbs no longer compete

as possible retrieval targets. Anecdotally, this possibility is illustrated by the following sentence:

> Today I bought a cake, streamers, balloons, party hats, candy, and several gifts **for** my niece's birthday.

In this example, the dependency from *for* to its modificand *bought* does not intuitively seem to induce a large processing cost, yet it spans 6 coordinated nouns, yielding an integration cost of 6, which is similar in magnitude to that of some of the most difficult dependencies explored in Grodner and Gibson (2005). The +C variant treats the entire coordinated direct object as one discourse referent, yielding an integration cost of 1.

Third, the discounting of preceding modifiers in +M is motivated by the possibility that modifier semantics may be integrated early, alleviating the need to retrieve the modifier once the head word is encountered. Anecdotally, this possibility is illustrated by the following sentence:

> (Yesterday,) my coworker, whose cousin drives a taxi in Chicago, **sent** me a list of all the best restaurants to try during my upcoming trip.

The dependency between the verb *sent* and the subject *coworker* spans a finite verb and 3 nouns, yielding an integration cost of 4 (plus a cost of 1 for the discourse referent introduced by *sent*). If the sentence includes the pre-sentential modifier *Yesterday*, which under the syntactic annotation used in this study is also involved in a dependency with the main verb *sent*, then the DLT predicts that it should double the structural integration cost at *sent* because the same set of discourse referents intervenes in two dependencies rather than one. Intuitively, this does not seem to be the case, possibly because the temporal information contributed by *Yesterday* may already be integrated with the incremental semantic representation of the sentence before *sent* is encountered, eliminating the need for an additional retrieval operation at that point. The +M modification instantiates this possibility.

A superficial consequence of the +C and +M variants is that they tend to attenuate large integration costs. Thus, if they improve fit to human measures, it may simply be the case that the DLT in its original formulation overestimates the costs of long dependencies. To account for this possibility, this study additionally considers a log-transformed variant of (otherwise unmodified) DLT integration cost.

Full description of left-corner parsing models of sentence comprehension is beyond the scope of this presentation (see e.g. Rasmussen and Schuler, 2018), which is restricted to the minimum details needed to define the predictors covered here. At a high level, phrasal structure derives from a sequence of fork (±F) and join (±J) decisions made at each word. In terms of memory structures, the fork decision depends on whether a new element (representing the current word and its hypothesized part of speech) matches current expectations about the upcoming syntactic category; if so, it is composed with the derivation at the front of the memory store (−F), and if not, it is pushed to the store as a new derivation fragment (+F). Following the fork decision, the join decision depends on whether the two items at the front of the store can be composed (+J) or not (−J). In terms of phrasal structures, fork decisions index the ends of multiword constituents (−F at the end of a multiword constituent, +F otherwise), and join decisions index the ends of left-child (center-embedded) constituents (+J at the end of a left child, −J otherwise). These composition operations (−F and +J) instantiate the notion of syntactic "integration" as envisioned by e.g. the DLT, since structures are retrieved from memory and updated by these operations. They each may thus plausibly contribute a memory cost (Shain et al., 2016), leading to the following left-corner predictors:

- **End of Constituent (−F).** Indicator for whether a word terminates a multiword constituent (i.e. generates a "no fork" decision by the parser).
- **End of Center Embedding (+J).** Indicator for whether a word terminates a center embedding (left child) of one or more words (i.e. generates a "yes join" decision by

the parser).

- **End of Multiword Center Embedding (−F, +J).** Indicator for whether a word terminates a multiword center embedding (i.e. generates both "no fork" and "yes join" decisions).

In addition, the difficulty of retrieval operations could in principle be modulated by locality, possibly due to activation decay and/or interference as argued by Lewis and Vasishth (2005). To account for this possibility, this study also explores distance-based left-corner predictors:

- **Length of Constituent (−F).** When a word terminates a multiword constituent, distance from most recent retrieval (including creation) of the derivation fragment at the front of the store (otherwise, 0).

- **Length of Multiword Center Embedding (−F, +J).** When a word terminates a multiword center embedding, distance from most recent retrieval (including creation) of the derivation fragment at the front of the store (otherwise, 0).

The notion of distance must be defined, and three definitions are explored here. One simply counts the number of words (WD). However, this complicates comparison to the DLT, which then differs not only in its conception of memory usage (constructing dependencies vs. retrieving/updating derivations in a stack), but also in its notion of locality (the DLT defines locality in terms of nouns and finite verbs, rather than words). To enable direct comparison, DLT-like distance metrics are also used in the above left-corner locality-based predictors, in particular, both using the original DLT definition of discourse referents (DR), as well as the modified variant +V that reweights finite and non-finite verbs (DRV). All three distance variants are explored for both distance-based left-corner predictors.

Note that these left-corner distance metrics more closely approximate ACT-R retrieval cost than DLT integration cost, since, as stressed by Lewis and Vasishth (2005), decay in ACT-R is determined by the recency with which an item in memory was previously

activated, rather than overall dependency length. Left-corner predictors can therefore be used to test one of the motivating insights of the ACT-R framework: the influence of reactivation on retrieval difficulty.

Note also that because the parser incrementally constructs expected dependencies between as-yet incomplete syntactic representations, at most two retrievals are cued per word (up to one for each of the fork and join decisions), no matter how many dependencies the word participates in. This property makes left-corner parsing a highly efficient form of incremental processing, a feature that has been argued to support its psychological plausibility (Johnson-Laird, 1983; van Schijndel et al., 2013; Rasmussen and Schuler, 2018).

The aforementioned left-corner predictors are roughly analogous to DLT integration cost (retrieval). This study additionally considers the number of incomplete derivation fragments that must be held in memory (similar to the number of incomplete dependencies in DLT storage cost), can be read off the stack depth of the parser state:

- **Embedding depth.** The number of incomplete derivation fragments left on the stack once a word has been processed.

This study additionally considers the possibility that pushing a new fragment to the stack may incur a cost:

- **Start of embedding.** Indicator for whether *Embedding Depth* increased from one word to the next.

i As with retrieval-based predictors, the primary difference between left-corner *embedding depth* and DLT *storage cost* is the efficiency with which the memory store is used by the parser. Because expected dependencies between incomplete syntactic derivations are constructed as soon as possible, a word can contribute at most one additional item to be maintained in memory (*vis-a-vis* DLT storage cost, which can in principle increase arbitrarily at words which introduce multiple incomplete dependencies). As mentioned above,

ACT-R does not posit storage costs at all, and thus investigation of such costs potentially stands to empirically differentiate ACT-R from DLT/left-corner accounts.

As an initial step, this study examined effect estimates and goodness of fit to the training set (see §7.1.3 for model design and §7.1.4 for ablative testing procedures) for many variants of the aforementioned hypotheses computed from a hand-corrected deep syntactic annotation of the experimental materials (Shain et al., 2018). In general, in LANG, exploratory results showed strong training-set effects of DLT variants and weak or null training-set effects of left-corner predictors, while in MD, results showed little reliable effect of either family of predictors. The strongest-performing DLT variant in the training set in both networks was the one that uses all three manipulations (+V, +C, +M) mentioned above, and it was therefore selected for our critical analysis (henceforth "DLT"). Critical analyses additionally include a control variable for the strongest left-corner predictor across networks (*end of center embedding*, i.e. +J, as above), an indicator marking words that terminate center embedding regions (single- or multi-word left children of right children in a phrase structure tree), where items in working memory are hypothesized to be composed during incremental parsing (Rasmussen and Schuler, 2018).

7.1.3 Model Design

Model design and HRF parametric family follow Chapter 6, except that improper uniform priors are replaced with proper priors as described in Chapter 3, motivated by evidence (Chapter 4) that models with proper priors converge more quickly and estimate uncertainty more conservatively.

The CDR model is as follows, where *italics* indicate convolved predictors and **bold** indicates the ablated predictor (the fixed effect for DLT integration cost):

$$\text{BOLD} \sim \text{TRNumber} + \textit{Rate} + \textit{SoundPower} + \textit{EndOfSentence} + \textit{PauseDuration} + \textit{Frequency} + \textit{5gramSurp} + \textit{PCFGSurp} + \textit{AdaptiveSurp} + \textit{yesJ} + \boldsymbol{\textit{DLT}} + (\text{TRNumber} + \textit{Rate} + \textit{SoundPower} + \textit{EndOfSentence} + \textit{PauseDuration} +$$

Frequency + *5gramSurp* + *PCFGSurp* + *AdaptiveSurp* + *yesJ* + *DLT* | fROI)
+ (1 | Participant)

that is, a linear coefficient for the index of the TR in the experiment, convolutions of the remaining predictors with the fitted HRF, by-fROI random variation in effect size and shape, and by-participant random variation in base response level. This model is used to test for significant effects of the DLT in each of LANG and MD. To test for significant differences between LANG and MD in DLT effect size, we additionally fitted the following model to the combined responses from both LANG and MD:

BOLD ∼ TRNumber + *Rate* + *SoundPower* + *EndOfSentence* + *PauseDuration* + *Frequency* + *5gramSurp* + *PCFGSurp* + *AdaptiveSurp* + *yesJ* + *DLT* + TRNumber:Network + *Rate*:Network + *SoundPower*:Network + *EndOfSentence*:Network + *PauseDuration*:Network + Frequency:Network + *5gramSurp*:Network + *PCFGSurp*:Network + *AdaptiveSurp*:Network + *yesJ*:Network + ***DLT***:Network + (1 — fROI) + (1 — Participant)

By-fROI random effects are simplified from the individual network models in order to support model identifiability (see Chapter 6).

7.1.4 Ablative Statistical Testing

Data partitioning follows that described in Chapter 6. Model quality is quantified as the Pearson sample correlation, henceforth ρ, between model predictions on the evaluation set and the true response. Fixed effects are tested by paired permutation test (Demšar, 2006) of the difference in correlation $\rho_{\text{diff}} = \rho_{\text{full}}\rho_{\text{ablated}}$, where ρ_{full} is the ρ of a model containing the fixed effect of interest while ρ_{ablated} is the ρ of a model lacking it. Paired permutation testing requires an elementwise performance metric that can be permuted between the two models, whereas Pearson correlation is a global metric that applies to the entire prediction-response matrix. To address this, tests exploited the fact that the sample correlation can be

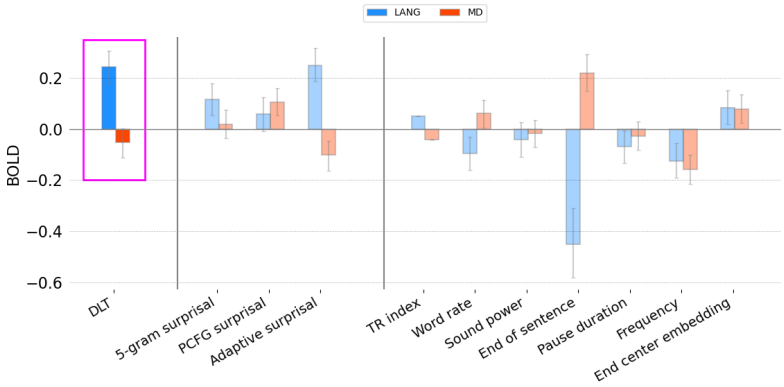


Figure 7.1: Effect sizes (HRF integrals) by network, with 95% Monte Carlo estimated credible intervals. DLT integration cost is associated with a strong increase in LANG activation, but is not associated with MD activation.

converted to an elementwise performance statistic as long as both variables are standardized (i.e. have sample mean 0 and sample standard deviation 1):

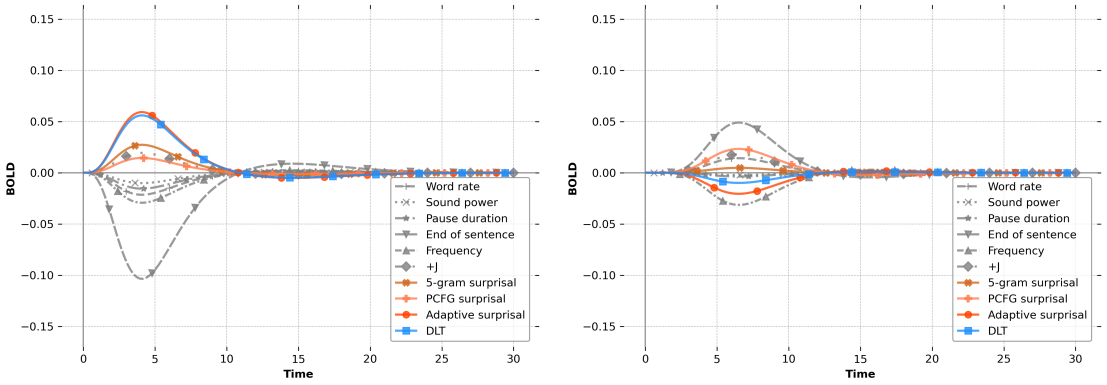
$$\rho\left(X,Y\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i \tag{7.1}$$

As a result, an elementwise performance metric can be derived as the elements of a Hadamard product between independently standardized prediction and response vectors. These products are then permuted in the usual way, using 10,000 resampling iterations. Each test involves a single ablated fixed effect, retaining all random effects in all models.

7.2 Results

7.2.1 Principal Analysis

Effect sizes (HRF integrals) by predictor are plotted in Figure 7.1, with the underlying CDR-estimated HRF shapes by network plotted in Figure 7.2. As shown, the DLT effect is strongly positive in LANG (among the largest effects in the model), but null in MD. Table 7.1 shows correlations between model predictions and true responses by network in both

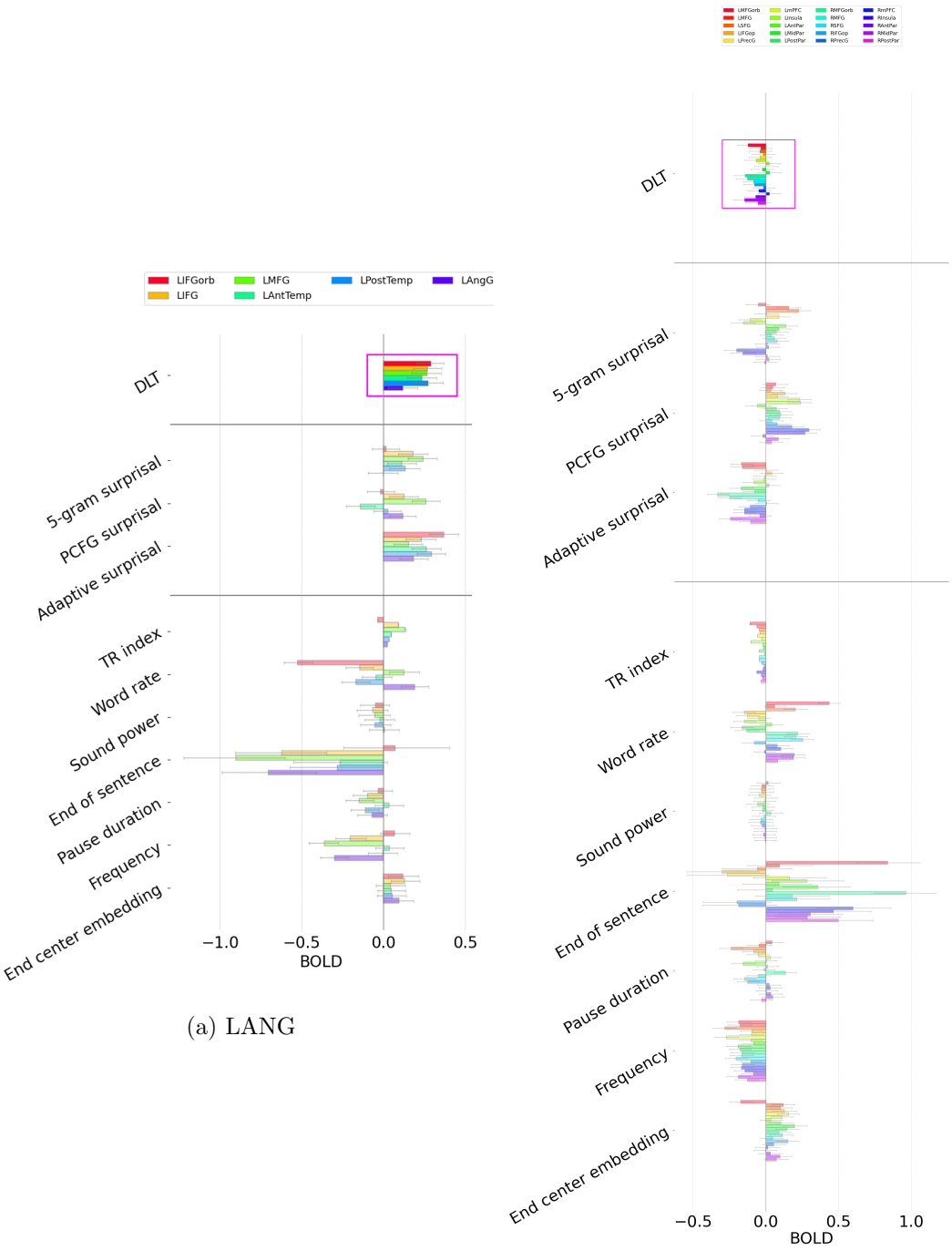


(a) LANG (b) MD

Figure 7.2: Overall CDR-estimated HRF shapes in LANG (left) and MD (right).

	LANG		MD		Combined	
	ρ-absolute	ρ-relative	ρ-absolute	ρ-relative	ρ-absolute	ρ-relative
Ceiling	0.221	1.0	0.116	1.0	0.152	1.0
Model (train)	0.142	0.643	0.072	0.621	0.084	0.553
Model (evaluation)	0.098	0.443	0.019	0.164	0.057	0.375

Table 7.1: Correlation ρ of model predictions with the true response under document-based repartitioning, compared to a ceiling measure correlating the true response with the mean response of all other participants for a particular story/fROI. "ρ-absolute" columns show absolute percent variance explained, while "ρ-relative" columns show the ratio of ρ-absolute to the ceiling.



(a) LANG

(b) MD

Figure 7.3: Effect sizes by fROI in LANG (left) vs. MD (right), with 95% Monte Carlo estimated credible intervals. The DLT effect is consistently positive across LANG fROIs, indicating a spatially distributed effect, whereas the DLT effect is not statistically greater than zero in any MD region.

fROI	Network	ρ_{diff}
AngG	LANG	0.004
AntTemp	LANG	0.018
IFG	LANG	0.006
IFGorb	LANG	0.009
MFG	LANG	0.005
PostTemp	LANG	0.008
LAntPar	MD	0.000
LIFGop	MD	-0.003
LInsula	MD	-0.004
LMFG	MD	-0.002
LMFGorb	MD	0.001
LMidPar	MD	-0.002
LPostPar	MD	-0.003
LPrecG	MD	0.000
LSFG	MD	0.001
LmPFC	MD	-0.002
RAntPar	MD	0.002
RIFGop	MD	0.002
RInsula	MD	-0.001
RMFG	MD	0.001
RMFGorb	MD	0.002
RMidPar	MD	0.001
RPostPar	MD	-0.001
RPrecG	MD	0.002
RSFG	MD	0.004
RmPFC	MD	-0.003

Table 7.2: By fROI out of sample correlation improvement from adding a fixed DLT effect. DLT models are systematically better correlated with the true response in LANG (6/6 regions) but not in MD (11/20 regions).

halves of the partition (training and evaluation), relative to a "ceiling" estimate of stimulus-driven correlation, computed as the correlation between (a) the responses in each region of each participant at each story exposure and (b) the average response of all other participants in that region, for that story. Consistent with prior studies (Blank and Fedorenko, 2017), LANG exhibits stronger language-driven synchronization across participants than MD. Our models also explain a greater share of that correlation in LANG vs. MD, especially on the out-of-sample evaluation set (40% relative correlation for LANG vs. 5% relative anticorrelation in MD). This finding is borne out by permutation tests on the out-of-sample evaluation set: the DLT significantly improves ρ_{diff} in LANG ($p < 0.0001^{***}$) but not MD ($p = 0.651$). The DLT effect is statistically larger in LANG than MD, since the increased DLT response in LANG (the positive interaction DLT:Network, Table 3) significantly improves ρ_{diff} in a combined model of responses from both networks ($p < 0.0001^{***}$).

Estimates also show a spatially distributed positive effect of DLT integration cost across the regions of LANG (Figure 7.3a) that systematically improves generalization quality (Table 4), indicating that all LANG fROIs are implicated to some extent in the processing costs associated with the DLT. This pattern does not hold in MD, where no region shows a statistically positive DLT response (Figure 7.3b), and the sign of ρ_{diff} across regions is roughly at chance (Table 7.2). These results converge to support strong, spatially distributed sensitivity in the language-selective network to an estimate of WM retrieval difficulty, but no clear evidence of such sensitivity anywhere within the domain-general multiple-demand network.

In addition to the critical analyses above, follow-up analyses were conducted to address possible alternative interpretations of the data pattern. First, because the data partition of Chapter 6 distributes materials across the training and evaluation sets, it is possible that item-level confounds may have affected results in ways that generalize to the test set. To address this possibility, the data were repartitioned so that the training and test sets contain non-overlapping materials and the critical analyses above were re-run (Reanalysis 1). The critical result is unchanged: estimated credible intervals for DLT effects overlap

between the two splits for both networks, with a significant positive DLT effect in LANG ($p < 0.0001^{***}$) but not MD ($p = 0.154$) and a significantly larger DLT effect in LANG than MD in a combined model of both networks responses ($p < 0.0001^{***}$). Evidence thus supports the existence of memory costs that generalize to new materials.

A second reanalysis addresses the possibility that the use of the flipped language localizer (Non-words > Sentences) to define MD may have ruled out linguistic effects in MD by construction. Note first that this is not entailed: regions that are not selective for sentences may still be implicated in language-related WM operations. Such a pattern in fact follows from the hypothesis targeted in this study that domain-general WM resources are recruited for sentence processing. In addition, the MD voxels selected by the Nonwords > Sentences contrast overlap heavily with voxels localized by a non-linguistic spatial WM task (and other executive tasks) and respond robustly to non-linguistic cognitive difficulty (Fedorenko et al., 2013). Nonetheless, this study addresses the concern directly by relocalizing the MD regions using a non-linguistic spatial WM contrast (Hard > Easy) and re-running the critical MD analyses (Reanalysis 2). In other words, rather than defining MD regions as those voxels with the greatest increase in activation for non-word lists over sentences, MD regions are redefined as those voxels with the greatest increase in activation for the hard condition over the easy condition in a (non-linguistic) spatial working memory task. Under the new split on materials, the DLT does not yield a held-out performance improvement (*de facto* $p = 1.0$). Under the original split from Chapter 6, the DLT contributes a significant performance improvement ($p = 0.002^{**}$), but its effect is negative (i.e. it is associated with a decrease in BOLD signal), which is not indicative of a retrieval cost.

The ensemble of follow-up analyses thus support the central claim: WM retrieval difficulty registers in the language network, with little effect in the multiple-demand network.

7.3 Discussion

This study investigated two key questions about the role of working memory (WM) retrieval during naturalistic sentence comprehension: does hypothesized retrieval difficulty register in brain activity (Q1), and, if so, are domain-general WM resources involved (Q2)? To do so, this study analyzed a large publicly available dataset of fMRI responses to naturalistic stories (Chapter 6) with respect to theory-driven estimates of syntactically modulated WM retrieval difficulty (Gibson, 2000) under rigorous controls for word predictability (van Schijndel and Linzen, 2018)), functionally localizing a domain-specific, language-selective network (LANG; Fedorenko et al., 2010) and a domain-general, multiple-demand network (MD; Duncan, 2010), implicated in executive functions, in each participant. Results show a strong, spatially distributed, positive association between LANG activation and syntactic dependency length during naturalistic language comprehension (the DLT; Gibson, 2000), but no such association in MD, the most plausible location for domain-general WM resources (Assem et al., 2020; Camilleri et al., 2018; Cole and Schneider, 2007; Duncan, 2010; Duncan and Owen, 2000; Gläscher et al., 2010; Goldman-Rakic, 1988; Kimberg and Farah, 1993; Owen et al., 1990; Prabhakaran et al., 2000; Rottschy et al., 2012). The DLT effect in LANG emerges on top of multiple strong controls for word frequency and predictability from context, including surprisal estimates from a recent cognitively-inspired recurrent neural network language model (van Schijndel and Linzen, 2018). This result withstands multiple reanalyses that (a) control for item-level confounds by resplitting the data document-wise (Reanalysis 1) and/or (b) use an alternative way to localize the MD network, with a non-linguistic WM task (Reanalysis 2). Results thus respectively support the answers "yes" and "no" to Q1 and Q2: WM retrieval difficulty registers strongly in brain activity during naturalistic sentence comprehension, but such retrieval is implemented in language-specialized cortical circuits, with little recruitment of domain-general WM resources housed within the MD network.

DLT estimates of retrieval difficulty are critically linked to the syntactic structure of sentences: retrieval operations are hypothesized to occur at words which terminate syntactic dependencies, with difficulty proportional to the number of intervening discourse referents that might compete strongly with the retrieval target. The present evidence that such difficulty measures robustly describe the brain response to naturalistic stories provides compelling support for syntactic processing as a core subroutine of typical language comprehension. This finding challenges to some extent prior arguments that representations computed during typical language comprehension are largely approximate and shallow (Frank et al., 2015; Frank and Bod, 2011; Frank and Christiansen, 2018; Swets et al., 2008), although further research is needed to determine more precisely the level of syntactic detail present in human mental representations (Brennan et al., 2016; Brennan and Hale, 2019; Lopopolo et al., 2020). Our results also suggest a spatially distributed burden of syntactic processing throughout the regions of the language network (see also Chapter 6) and are not consistent with prior arguments for the existence of one or two dedicated syntactic processing centers (Bemis and Pylkkänen, 2011; Friederici et al., 2006; Hagoort, 2005; Matchin et al., 2017; Matchin and Hickok, 2020; Pallier et al., 2011; Vandenberghe et al., 2002).

The DLT effects shown here furthermore fail to be accounted for by multiple strong measures of word surprisal, which has repeatedly been shown in prior work to describe naturalistic human sentence processing responses across modalities, including behavioral (Aurnhammer and Frank, 2019; Demberg and Keller, 2008; Fossum and Levy, 2012; Frank and Bod, 2011; Smith and Levy, 2013; van Schijndel and Schuler, 2015), electrophysiological (Armeni et al., 2019; Frank et al., 2015), and fMRI (Brennan et al., 2016; Henderson et al., 2016; Lopopolo et al., 2017; Willems et al., 2015, also, Chapter 6). In its strong form, surprisal theory (Levy, 2008) equates sentence comprehension with allocating activation between (potentially infinite) possible interpretations of the unfolding sentence, in proportion to their probability given the currently observed string. Under such a view, structured representations are assumed to be available, and the primary work of comprehension is

(probabilistically) selecting among them. However, according to integration-based theories (Gibson, 2000; Lewis and Vasishth, 2005), incremental effort is required to compute the available interpretations in the first place (i.e. by storing, retrieving, and updating representations in memory). By showing integration costs that are not well explained by word predictability, the current study joins recent arguments in favor of complementary roles played by integration and prediction in language comprehension (Ferreira and Chantavarin, 2018; Levy et al., 2013). Strong word predictability controls are of course a perpetually moving target: the present methods cannot rule out the possibility that some other current or future statistical language model might explain apparent WM effects. However, such an objection effectively renders surprisal theory unfalsifiable. This study has attempted to address such concerns by deploying a surprisal control (adaptive surprisal) that, at the time of writing, is recent, high performance, and cognitively motivated.

This study bears on the role of working memory in functionally identified cortical networks, but does not rule out a domain-general role for sub-cortical structures in WM during language processing, especially hippocampus (Leszczynski, 2011; Olson et al., 2006; Olton et al., 1979; Yonelinas, 2013; Yoon et al., 2008). While the contribution of MD regions to WM is well established, the role of hippocampus in WM (as opposed to long-term episodic memory) is less so, given several prior studies that have not found an association between WM and hippocampus (Baddeley et al., 2010; Baddeley and Warrington, 1970; Jeneson et al., 2010; Nadel and MacDonald, 1980; Shrager et al., 2008). Investigation of possible subcortical WM contributions to naturalistic language processing is left to future research.

Taken together with prior work, these results also bear on the computational nature of the memory systems that support language comprehension. The DLT itself is agnostic toward the design of the retrieval mechanism by which intervening discourse referents increase retrieval cost. One possible mechanism that is consistent with behavioral evidence of DLT-driven reading slowdowns is recency-based serial search, whereby the processor iterates backward over possible dependency targets by recency of mention until a good match is

encountered. This interpretation is inconsistent with findings from speed-accuracy trade-off experiments, which do not show improved retrieval accuracy at longer latencies and thus instead support a direct, content-addressable retrieval mechanism whereby similar competitors degrade the retrieved representation but do not increase retrieval time (McElree et al., 2003). A plausible hypothesis based on these speed-accuracy trade-off results is that all retrieval operations require the same amount of computation and differ only in the fidelity of the retrieved representation. The present findings from fMRI disconfirm this hypothesis: computing dependencies with more intervening competitors requires measurably more neuronal activity (hence, more computation). If this is not the result of serial search (McElree et al., 2003), a plausible alternative hypothesis is that computationally-intensive clean-up mechanisms exist to identify and repair low-quality retrieval outputs (Van Dyke and Lewis, 2003).

In conclusion, results of this study (1) support the hypothesis that incremental retrieval and integration of syntactic structures in memory is a core component of human sentence comprehension and (2) indicate that the memory resources responsible for these computations reside primarily in language-specialized cortical circuits, rather than in domain-general executive control areas.

Part IV

CDRNN: A Deep Neural Extension of CDR

Chapter 8

CDRNN Motivation, Definition, and Evaluation

The CDR approach proposed in this thesis has been shown to address the problem of temporal diffusion in studies of human language processing (Chapters 1–4) and to be useful for investigating the existence and timecourse of processing effects (Chapters 5–7). However, CDR retains a number of simplifying assumptions (e.g. that the IRF is fixed over time) that may not hold of the human language processing system (linearity, additivity, homosketasticity, stationarity, and context-independence, see Section 8.1).

Deep neural networks (DNNs), widely used in natural language processing (NLP), can relax these strict assumptions. Indeed, psycholinguistic regression analyses and NLP systems share a common structure: both fit a function from word features to some quantity of interest. However, psycholinguistic regression models face an additional constraint: they must be interpretable enough to allow researchers to study relationships between variables in the model. This requirement may be one reason why black box DNNs are not generally used to analyze psycholinguistic data, despite the tremendous gains DNNs have enabled in natural language tasks (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020, *inter alia*), in part by better approximating the complex dynamics of human cognition as encoded in natural language (Linzen et al., 2016; Gulordava et al., 2018; Tenney et al., 2019; Hewitt and Manning, 2019; Wilcox et al., 2019; Schrimpf et al., 2020).

This chapter proposes an attempt to leverage the flexibility of DNNs for psycholinguistic data analysis. The continuous-time deconvolutional regressive neural network (CDRNN) is an extension of CDR that reimplements the impulse response function as a DNN describing the expected influence of preceding events (e.g. words) on future responses (e.g. reading

times) as a function of their properties and timing. CDRNN retains the deconvolutional design of CDR while relaxing many of its simplifying assumptions, resulting in a highly flexible model. Nevertheless, CDRNN can also shed light on the underlying data generating process. Results on reading and fMRI measures show substantial generalization improvements from CDRNN over baselines, along with detailed insights about the underlying dynamics that cannot easily be obtained from existing methods.

8.1 Background

As argued in Chapter 2, psycholinguists long been aware for decades that processing effects may lag behind the words that trigger them (Morton, 1964; Bouma and De Voogd, 1974; Rayner, 1977; Erlich and Rayner, 1983; Mitchell, 1984; Rayner, 1998; Vasishth and Lewis, 2006; Smith and Levy, 2013), possibly because cognitive "buffers" may exist to allow higher-level information processing to catch up with the input (Bouma and De Voogd, 1974; Baddeley et al., 1975; Just and Carpenter, 1980; Ehrlich and Rayner, 1981; Mollica and Piantadosi, 2017). They have also recognized the potential for non-linear, interactive, and/or time-varying relationships between word features and language processing (Smith and Levy, 2013; Baayen et al., 2017, 2018). No prior regression method can jointly address these concerns in non-uniform time series (e.g. words with variable duration) like naturalistic psycholinguistic experiments. Discrete-time methods (e.g. lagged/spillover regression, Sims, 1971; Erlich and Rayner, 1983; Mitchell, 1984) ignore potentially meaningful variation in event duration, even if some (e.g. generalized additive models, or GAMs, Hastie and Tibshirani, 1986; Wood, 2006) permit non-linear and non-stationary (time-varying) feature interactions (Baayen et al., 2017). CDR addresses this limitation by fitting continuous-time IRFs, but it assumes that the IRF is stationary (time invariant), that features scale linearly and combine additively, and that the response variance is constant (homoskedastic). By implementing the IRF as a time-varying neural network, CDRNN relaxes all of these

assumptions, incorporating the featural flexibility of GAMs while retaining the temporal flexibility of CDR.

Previous studies have investigated latency and non-linearity in human sentence processing. For example, Smith and Levy (2013) attach theoretical significance to the functional form of the relationship between word surprisal and processing cost, using GAMs to show that this relationship is linear and arguing on this basis that language processing is highly incremental. This claim is under active debate (Brothers and Kuperberg, 2021), underlining the importance of methods that can investigate questions of functional form. Smith and Levy (2013) also investigate the timecourse of surprisal effects using spillover and find a more delayed surprisal response in self-paced reading (SPR) than in eye-tracking. Chapter 4 supports the latter finding using CDR, and in addition show evidence of strong inertia effects in SPR, such that participants who have been reading quickly in the recent past also read more quickly now. However, this outcome may be an artifact of the stationarity assumption: CDR may be exploiting its estimates of rate effects in order to capture broad non-linear negative trends (e.g. task adaptation, Prasad and Linzen, 2019) in a stationary model. Similarly, the generally null word frequency estimates reported in Chapter 4 may be due in part to the assumption of additive effects: word frequency and surprisal are related, and they may coordinate interactively to determine processing costs (Norris, 2006). Thus, in general, prior findings on the timecourse and functional form of effects in human sentence processing may be influenced by methodological limitations: the GAM models of Smith and Levy (2013) ignore variable event duration, the CDR models of Chapter 4 ignore non-linearity, and both approaches assume stationarity, context-independence, constant variance, and additive effects. By jointly relaxing these potentially problematic assumptions, CDRNN stands to support more reliable conclusions about human language comprehension, while also possibly enabling new insights into cognitive dynamics.

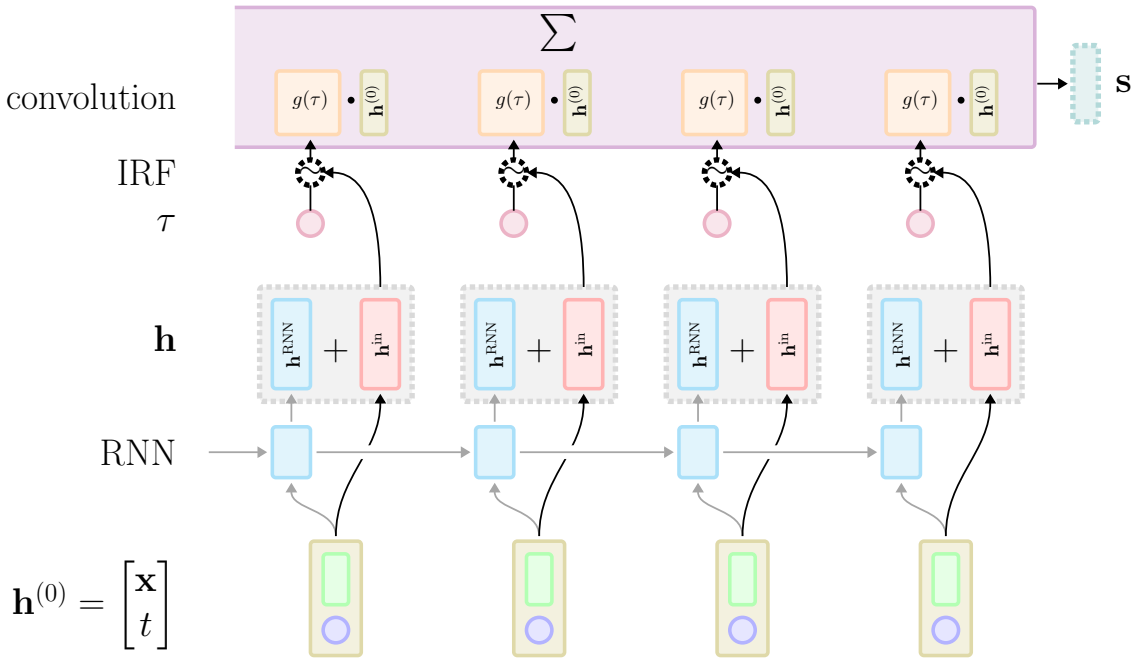


Figure 8.1: **CDRNN model.** Subscripts omitted to reduce clutter. The IRF $g(\tau)$ at an event computes the expected contribution of each feature of the event vector $\mathbf{h}^{(0)}$ to each element of the parameter vector $\mathbf{e}$ of the predictive distribution for a particular response value. The first layer of the IRF depends non-linearly on the properties of the event via $\mathbf{h}^{\text{in}}$ and (optionally) on context via $\mathbf{h}^{\text{RNN}}$, which requires the recurrent connections in gray. Elements with random effects have dotted outlines. For variable definitions, see §8.2.2.

8.2 Model

8.2.1 Architecture

This section presents a high-level description of the model design (formal definition follows). The CDRNN architecture is represented schematically in Figure 8.1. The primary goal of estimation is to identify the deep neural IRF (top) that computes the influence of a preceding event on a subsequent response as a function of their distance in time τ. The IRF depends on the properties of preceding events via the deep neural projections in the lower part of the figure. First, the predictors $\mathbf{x}$ are concatenated with their timestamps t and submitted to the model as input. Inputs are cast to a hidden state for each preceding event as the

Dataset	CDR	CDRNN-FF	CDRNN-RNN
Synth	662	8,162	17,890
NatStor (SPR)	21,845	13,261	31,213
Dundee	2,080	8,012	15,980
NatStor (fMRI)	331	10,352	22,582

Table 8.1: Number of trainable parameters by model and dataset.

sum of three quantities: a feedforward projection $\mathbf{h}^{\text{in}}$ of each input, a forward-directional RNN projection $\mathbf{h}^{\text{RNN}}$ of the events up to and including each input, and random effects $\mathbf{h}^{\text{Z}}$ containing offsets for the relevant random effects level(s) (e.g. for each participant in an experiment). In this study, the recurrent component is treated as optional (gray arrows). Without the RNN, the model is non-stationary (via input t) but cannot capture context-dependence.

Each input $h^{(0)}$ is then premultiplied its corresponding IRF $g(\tau)$ to produce the expected contribution of each element of the input vector to each element of the parameter vector $\mathbf{e}$ of the predictive distribution for a particular response (e.g. mean and variance of a univariate normal response). The summation over the time dimension at the top of the figure ensures that the model is deconvolutional: each preceding input contributes to the response in some proportion, with that proportion determined by the features, context, and relative timing of that input. The dependence of all elements of $\mathbf{e}$ on the sequence of timestamped predictors permits a time-varying (non-constant) response distribution: based on the sequence of preceding events, the model determines the structure of its own uncertainty about the response. Because the IRF depends on a deep neural projection of the current stimulus as well as (optionally) the entire sequence of preceding stimuli, it constitutes a manifold over the predictor dimensions and time, implicitly estimating all interactions between these variables in governing the response. Predictors may thus coordinate in a non-linear, non-additive, and time-varying manner.

Despite their flexibility (above) and task performance (Section 8.4), CDRNN models used in this study have few parameters (Table 8.1) by current deep learning standards (in

one case, even fewer parameters than CDR) because they are relatively shallow and small (§8.2.6).

8.2.2 Mathematical Definition

This section formally defines the CDRNN model.[1] CDRNN assumes the following quantities as input:[2]

- $X \in \mathbb{N}$: Number of predictor observations (e.g. word exposures)
- $Y \in \mathbb{N}$: Number of response observations (e.g. fMRI scans)
- $Z \in \mathbb{N}$: Number of random grouping factor levels (e.g. distinct participants)
- $K \in \mathbb{N}$: Number of predictors
- $\mathbf{X} \in \mathbb{R}^{X \times K}$: Design matrix of X predictor observations of K dimensions each.
- $\mathbf{y} \in \mathbb{R}^{Y}$: Vector of Y response observations
- $\mathbf{Z} \in \{0, 1\}^{Y \times Z}$: Boolean matrix indicating random grouping factor levels associated with each response observation
- $\mathbf{t} \in \mathbb{R}^{X}$: Vector of timestamps associated with each observation in $\mathbf{X}$
- $\mathbf{t}' \in \mathbb{R}^{Y}$: Vectors of timestamps associated with each observation in $\mathbf{y}$
- $S \in \mathbb{N}$: Number of parameters in predictive distribution (e.g. 2 for a normal distribution: mean and variance)

For simplicity of exposition, $\mathbf{X}$ and $\mathbf{y}$ are assumed to contain data from a single time series (e.g. a single participant performing a single experiment). The definition below can be

[1] These equations represent the most recent state of the system. However, results reported in this chapter were obtained using a slightly different definition. Reproduction using the new definition was not able to complete in time for inclusion in this thesis, but early outcomes suggest that they yield similar estimates.

[2] Throughout these definitions, vectors and matrices are notated in **bold** lowercase and uppercase, respectively (e.g. $\mathbf{u}$, $\mathbf{U}$). Objects with indexed names are designated using subscripts (e.g. $\mathbf{v}_r$). Vector and matrix indexing operations are notated using subscript square brackets, and slice operations are notated using $*$ (e.g. $\mathbf{X}_{[*,k]}$ denotes the k^{th} column of matrix $\mathbf{X}$). Hadamard (pointwise) products are notated using $\odot$. The notations $\mathbf{0}$ and $\mathbf{1}$ designate conformable column vectors of 0's and 1's, respectively. Superscripts are used for indexation and do not denote exponentiation.

applied without loss of generality to data containing multiple time series by concatenating the output of the model as applied to multiple $\mathbf{X}$, $\mathbf{y}$ pairs. $\mathbf{X}, \mathbf{y}$ and their associated satellite data $\mathbf{Z}, \mathbf{t}, \mathbf{t}'$ must be temporally sorted.

Given these inputs, CDRNN estimates a latent impulse response function that relates timestamped predictors to all parameters of the assumed predictive distribution. For example, assuming a univariate normally distributed response, CDRNN learns an IRF with two output dimensions, one for the predictive mean, and one for the predictive variance. Such a design (i.e. modeling dependencies on the predictors of all parameters of the predictive distribution) has previously been termed *distributional regression* (Bürkner, 2018).

CDRNN contains a recurrent neural network (RNN), neural projections that map inputs and RNN states to a hidden state for each preceding event, and neural projections that map the hidden states to predictions about (1) the influence of each event on the response (IRF) and (2) the parameter(s) of the error distribution (e.g. the variance of a Gaussian error). The definition assumes the following quantities:

- L_{in}, L_{RNN}, L_{IRF}, $L_{\epsilon} \in \mathbb{N}$: Number of layers in the input projection, RNN, IRF, and error parameter function, respectively
- $D_{\text{in}(\ell)}$, $D_{\text{RNN}(\ell)}$, D_{h}, $D_{\text{IRF}(\ell)}$, $D_{\text{error}(\ell)} \in \mathbb{N}$: Number of output dimensions in the ℓ^{th} layer of the input projection, RNN, hidden state, IRF, and error parameter function, respectively

The following values are deterministically assigned:

- $D_{\text{IRF}(L_{\text{IRF}})} = S(K+1)$ (the IRF generates a convolution weight for every predictor dimension, plus the timestamp, for each parameter of the predictive distribution)
- $D_{\text{in}(0)} = K + 1$ (input is predictors + time)
- $D_{\text{in}(L_{\text{in}})} = D_{\text{h}}$

In these definitions, integers x, y respectively refer to row indices of $\mathbf{X}$, $\mathbf{y}$. Let $\mathbf{z}_y$ be the vector $\left(\mathbf{Z}_{[y,*]}\right)^{\top}$ of random effects associated with the response at y. Let $\mathbf{W}^{\text{h,Z}} \in \mathbb{R}^{D_{\text{h}} \times Z}$,

$\mathbf{W}^{\text{IRF}(1),\text{Z}} \in \mathbb{R}^{2D_{\text{IRF}(1)} \times Z}$, and $\mathbf{W}^{\text{s,Z}} \in \mathbb{R}^{S \times Z}$ be an embedding matrix for $\mathbf{z}_y$. Random effects offsets at response step y for the hidden state ($\mathbf{h}_y^{\text{Z}}$), the weights and biases of the first layer of the IRF ($\mathbf{w}_y^{\text{IRF}(1),\text{Z}}$, $\mathbf{b}_y^{\text{IRF}(1),\text{Z}}$), and the parameters of the predictive distribution ($\mathbf{e}_y^{\text{Z}}$, i.e. random intercepts and variance parameters) are generated as follows:

$$\mathbf{h}_y^{\text{Z}} \overset{\text{def}}{=} \mathbf{W}^{\text{h,Z}}\mathbf{z}_y \tag{8.1}$$

$$\begin{bmatrix} \mathbf{w}_y^{\text{IRF}(1),\text{Z}} \\ \mathbf{b}_y^{\text{IRF}(1),\text{Z}} \end{bmatrix} \overset{\text{def}}{=} \mathbf{W}^{\text{IRF}(1),\text{Z}}\mathbf{z}_y \tag{8.2}$$

$$\mathbf{s}_y^{\text{Z}} \overset{\text{def}}{=} \mathbf{W}^{\text{s,Z}}\mathbf{z}_y \tag{8.3}$$

Following prior work in mixed effects models (Bates et al., 2015), to ensure that population-level estimates reliably encode central tendency, each output dimension of $\mathbf{W}^{\text{h,Z}}$, $\mathbf{W}^{\text{IRF}(1),\text{Z}}$, and $\mathbf{W}^{\text{s,Z}}$ is constrained to have mean 0 across the levels of each random grouping factor (e.g. across participants in the study).

The neural IRF is applied to a temporal offset τ representing the delay at which to query the response to an input (e.g. $\tau = 1$ queries the response to an input 1s after the input occurred). The output of the neural IRF $g_{x,y}^{\ell}(\tau) \in \mathbb{R}^{D_{\text{IRF}(\ell)}}$ applied to τ at layer ℓ is defined as:

$$g_{x,y}^{(1)}(\tau) \overset{\text{def}}{=} s_{\text{IRF}(1)}\left(\mathbf{w}_{x,y}^{\text{IRF}(1)}\tau + \mathbf{b}_{x,y}^{\text{IRF}(1)}\right) \tag{8.4}$$

$$g_{x,y}^{(\ell)}(\tau) \overset{\text{def}}{=} s_{\text{IRF}(\ell)}\left(\mathbf{W}^{\text{IRF}(\ell)}g_{x,y}^{(\ell-1)}(\tau) + \mathbf{b}^{\text{IRF}(\ell)}\right), \quad \ell > 1 \tag{8.5}$$

$$\mathbf{w}_{x,y}^{\text{IRF}(1)} \overset{\text{def}}{=} \mathbf{w}^{\text{IRF}(1)} + \mathbf{w}_y^{\text{IRF}(1),\text{Z}} + \mathbf{W}_{\Delta}^{\text{IRF}(1)}\mathbf{h}_{x,y} \tag{8.6}$$

$$\mathbf{b}_{x,y}^{\text{IRF}(1)} \overset{\text{def}}{=} \mathbf{b}^{\text{IRF}(1)} + \mathbf{b}_y^{\text{IRF}(1),\text{Z}} + \mathbf{B}_{\Delta}^{\text{IRF}(1)}\mathbf{h}_{x,y} \tag{8.7}$$

$\mathbf{W}_{x,y}^{\text{IRF}(\ell)}$, $\mathbf{b}_{x,y}^{\text{IRF}(\ell)}$, and $s_{\text{IRF}(\ell)}$ are respectively the ℓ^{th} IRF layer's weight matrix at predictor timestep x and response timestep y, bias vector at time x, y, and squashing function, and $g_{x,y}^{(0)}(\tau) = \tau$. $\mathbf{w}^{\text{IRF}(1)}$, $\mathbf{b}^{\text{IRF}(1)}$ are respectively globally applied initial weight and bias vectors

for the first layer of the IRF, which transforms scalar τ, each of which is shifted by its corresponding random effects. $\mathbf{W}_{\Delta}^{\text{IRF}(1)}$, $\mathbf{B}_{\Delta}^{\text{IRF}(1)}$ are respectively weight matrices used to compute additive modifications to $\mathbf{W}^{\text{IRF}(1)}$ from CDRNN hidden state $\mathbf{h}_{x,y}$, similar in spirit to a residual network (He et al., 2016). Non-initial IRF layers are treated as stationary (i.e. their parameters are independent of x, y). The final output of the IRF is given by:

$$g_{x,y}(\tau) \stackrel{\text{def}}{=} \text{reshape}\left(g_{x,y}^{(L_{\text{IRF}})}(\tau), (S, K+1)\right) \tag{8.8}$$

The hidden state $\mathbf{h}_{x,y}$ is computed as the squashed sum of several quantities: a global bias $\mathbf{h}^{\text{bias}}$, random effects $\mathbf{h}^{\text{Z}}$, a neural projection $\mathbf{h}_{x,y}^{\text{in}}$ of the inputs at x, y, and a neural projection $\mathbf{h}_{x,y}^{\text{RNN}}$ of the hidden state of an RNN over the sequence of predictors up to and including timestep x:

$$\mathbf{h}_{x,y} \stackrel{\text{def}}{=} s_{\text{h}}\left(\mathbf{h}^{\text{bias}} + \mathbf{h}_{y}^{\text{Z}} + \mathbf{h}_{x,y}^{\text{in}} + \mathbf{h}_{x,y}^{\text{RNN}}\right) \tag{8.9}$$

The IRF $g_{x,y}$ is therefore feature-dependent via the neural projection $\mathbf{h}_{x,y}^{\text{in}}$ of the input at x, y and context-dependent via the neural projection $\mathbf{h}_{x,y}^{\text{RNN}}$ of an RNN over the input up to x for the response at y. This design relaxes stationarity assumptions while also sharing structure across timepoints. The definitions of $\mathbf{h}_{x,y}^{\text{in}}$ and $\mathbf{h}_{x,y}^{\text{RNN}}$ are given below.

Let t_x be the element $\mathbf{t}_{[x]}$ and $\mathbf{x}_x$ be the x^{th} predictor vector $\left(\mathbf{X}_{[x,*]}\right)^{\top}$. The inputs $\mathbf{h}_{x,y}^{(0)}$ to the CDRNN model are defined as the vertical concatenation of the predictors $\mathbf{x}_x$ and the event timestamp t_x:

$$\mathbf{h}_{x,y}^{(0)} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{x}_x \\ t_x \end{bmatrix} \tag{8.10}$$

The output of the input projection at layer l and time x, y is defined as:

$$\mathbf{h}_{x,y}^{\text{in}(\ell)} \stackrel{\text{def}}{=} s_{\text{in}(\ell)}\left(\mathbf{W}^{\text{in}(\ell)}\mathbf{h}_{x,y}^{\text{in}(\ell-1)} + \mathbf{b}^{\text{in}(\ell)}\right) \tag{8.11}$$

where $\mathbf{h}_{x,y}^{\text{in}(0)} \overset{\text{def}}{=} \mathbf{h}_{x,y}^{(0)}$. At the final layer, $s_{\text{in}(L_{\text{in}})}$ is identity and $\mathbf{b}^{\text{in}(L_{\text{in}})} = \mathbf{0}$, since $\mathbf{h}_{x,y}$ already has a bias. The final output of the input projection is given by:

$$\mathbf{h}_{x,y}^{\text{in}} \overset{\text{def}}{=} \mathbf{h}_{x,y}^{\text{in}(L_{\text{in}})} \tag{8.12}$$

Note that $\mathbf{h}_{x,y}^{\text{in}}$ is already non-stationary by virtue of its dependence on the event timestamp $\mathbf{t}_{[x]}$, which allows the IRF to differ between timepoints (see e.g. Baayen et al., 2017, for development of a similar idea using generalized additive models). While this model of non-stationarity can be complex and non-linear, it is still limited by context-independence. That is, the change in the IRF over time depends only on the amount of time elapsed since the start of the time series, independently of which events preceded. However, it is possible that the contents of the events in a time series may influence the IRF, above any deterministic change in response over time (for example, if several difficult preceding words have already taxed the processing buffer, additional processing costs may become larger). To account for this possibility, an RNN is built into the CDRNN design.[3] Any variant of RNN can be used (this study uses a long short-term memory network, or LSTM, Hochreiter and Schmidhuber, 1997). The ℓ^{th} RNN hidden state at x, y is designated by $\mathbf{h}_{x,y}^{\text{RNN}(\ell)}$. To account for the possibility of random variation in sensitivity to context, the initial hidden and cell states $\mathbf{h}_{0,y}^{\text{RNN}(\ell)}$, $\mathbf{c}_{0,y}^{\text{RNN}(\ell)}$ depend on the random effects:

$$\mathbf{h}_{0,y}^{\text{RNN}(\ell)} \overset{\text{def}}{=} \mathbf{h}_{0}^{\text{RNN}(\ell)} + \mathbf{W}_{\text{Z}}^{\text{RNN}_{\text{h}}(\ell)}\mathbf{z}_y \tag{8.13}$$

$$\mathbf{c}_{0,y}^{\text{RNN}(\ell)} \overset{\text{def}}{=} \mathbf{c}_{0}^{\text{RNN}(\ell)} + \mathbf{W}_{\text{Z}}^{\text{RNN}_{\text{c}}(\ell)}\mathbf{z}_y \tag{8.14}$$

where $\mathbf{h}_{0}^{\text{RNN}(\ell)}$, $\mathbf{c}_{0}^{\text{RNN}(\ell)}$ are global biases and $\mathbf{W}_{\text{Z}}^{\text{RNN}_{\text{h}}(\ell)}$, $\mathbf{W}_{\text{Z}}^{\text{RNN}_{\text{c}}(\ell)}$ are constrained to have mean 0 within each random grouping factor.

[3] The experiments in this study also consider a variant without the RNN component, which is mathematically equivalent to setting $\mathbf{h}_{x,y}^{\text{RNN}} = \mathbf{0}$.

Non-initial RNN states are computed via a standard LSTM update:

$$\left[\mathbf{h}_{x,y}^{\text{RNN}(\ell)}, \mathbf{c}_{x,y}^{\text{RNN}(\ell)}\right] \overset{\text{def}}{=} \text{LSTM}\left(\mathbf{h}_{x-1,y}^{\text{RNN}(\ell)}, \mathbf{c}_{x-1,y}^{\text{RNN}(\ell)}, \mathbf{h}_{x,y}^{\text{RNN}(\ell-1)}\right) \tag{8.15}$$

The hidden state of the final RNN layer is linearly projected to the dimensionality of the CDRNN hidden state:

$$\mathbf{h}_{x,y}^{\text{RNN}} \overset{\text{def}}{=} \mathbf{W}^{\text{RNNproj}} \mathbf{h}_{x,y}^{\text{RNN}(L_{\text{RNN}})} \tag{8.16}$$

To apply the CDRNN model to data, a mask $\mathbf{F} \in \{0,1\}^{Y \times X}$ admits only those observations in $\mathbf{X}$ that precede each $\mathbf{y}_{[y]}$:

$$\mathbf{F}_{[y,x]} \overset{\text{def}}{=} \begin{cases} 1 & \mathbf{t}_{[x]} \leq \mathbf{t'}_{[y]} \\ 0 & \text{otherwise} \end{cases} \tag{8.17}$$

Letting $\tau_{x,y}$ denote the temporal offset between the predictors at x and the response at y, i.e. $\tau_{x,y} \overset{\text{def}}{=} \mathbf{t'}_{[y]} - \mathbf{t}_{[x]}$. A total of $S(K+1)$ sparse convolution matrices $\mathbf{G}_{s,k} \in \mathbb{R}^{Y \times X}$ are defined to contain the predicted response to each preceding event for the k^{th} dimension of $\mathbf{h}_{x,y}^{(0)}$ and the s^{th} parameter of the predictive distribution, masked by $\mathbf{F}$:

$$\mathbf{G}_{s,k} \overset{\text{def}}{=} \begin{bmatrix} g_{1,1}(\tau_{1,1})_{[s,k]} & \cdots & g_{X,1}(\tau_{X,1})_{[s,k]} \\ \vdots & \ddots & \vdots \\ g_{1,Y}(\tau_{1,Y})_{[s,k]} & \cdots & g_{X,Y}(\tau_{X,Y})_{[s,k]} \end{bmatrix} \odot \mathbf{F} \tag{8.18}$$

The convolved design matrix $\mathbf{X'}^{(s)} \in \mathbb{R}^{Y \times (K+1)}$ for the s^{th} parameter of the predictive distribution is then computed as:

$$\mathbf{X'}_{[*,k]}^{(s)} \overset{\text{def}}{=} \mathbf{G}_{s,k} \left[\mathbf{X},\ \mathbf{t}\right]_{[*,k]} \tag{8.19}$$

Vector $\mathbf{s} \in \mathbb{R}^S$ contains global, population-level estimates of the parameters of the

predictive distribution. Under the univariate normal predictive distribution assumed in this study, $\mathbf{s}$ contains the predictive mean (μ, i.e. the intercept) and variance (σ^2):

$$\mathbf{s} \stackrel{\text{def}}{=} \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \tag{8.20}$$

Matrix $\mathbf{S}^{\text{Z}}$ contains random predictive distribution parameter estimates for the y^{th} response $\mathbf{s}_y^{\text{Z}}$:

$$\mathbf{S}^{\text{Z}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{s}_1^{\text{Z}\top} \\ \vdots \\ \mathbf{s}_Y^{\text{Z}\top} \end{bmatrix} \tag{8.21}$$

The vector of values for each response y for the s^{th} predictive distribution parameter is given by summing the population value, random effects values, and convolved response values:

$$\mathbf{S}_{[*,s]} \stackrel{\text{def}}{=} \mathbf{X}'^{(s)}\mathbf{1} + \mathbf{S}^{\text{Z}}_{[*,s]} + \mathbf{s}_{[s]} \tag{8.22}$$

Given an assumed distributional family $\mathcal{F}$ (here assumed to be univariate normal), the response in the CDRNN model is distributed as:

$$\mathbf{y} \sim \mathcal{F}\left(\mathbf{S}_{[*,1]}, \dots, \mathbf{S}_{[*,S]}\right) \tag{8.23}$$

8.2.3 Asynchronously Measured Predictor Dimensions

As discussed in Chapter 3, CDR applies straightforwardly to time series with asynchronous predictor vectors and response values (i.e. measured at different times, such as word onsets that do not align with fMRI scan times). The CDR implementation of Chapter 3 also supports asynchronously measured dimensions of the predictor matrix, simply by providing each predictor dimension with its own vector of timestamps. This allows e.g. the analyses in Chapter 6 to regress linguistic features (which are word-aligned) and sound power (which

in their definition is measured at regular 100ms intervals) in the same model. Supporting asynchronously measured predictor dimensions is more challenging in CDRNN, especially if the RNN component is used. The solution used in CDR is not available because input dimensions that do not align in time are (1) arbitrarily grouped together and (2) erroneously treated as steps in the RNN input sequence. A more principled solution is to interleave the predictors in time order and pad irrelevant dimensions with zeros. For example, in a model with predictor A and predictor B that are sampled at different times, the values of A and B are temporally sorted together into a single time series, with the B value of A events set to zero and the A value of B events set to zero. This approach carries a computational cost: unlike CDR, the number of inputs to the convolution scales linearly on the number of asynchronously measured sets of predictors in the model.

8.2.4 Objective and Regularization

Given (1) an input configuration $\mathcal{C}$ containing predictors $\mathbf{X}$, input timestamps $\mathbf{t}$, and response timestamps $\mathbf{t}'$, (2) CDRNN parameter vector $\mathbf{w}$, (3) output distribution p, and (4) random effects vector $\mathbf{z}$, the model uses gradient descent to minimize the following objective:

$$\mathcal{L}\left(\mathbf{y} \mid \mathcal{C}; \mathbf{w}, \mathbf{z}\right) \stackrel{\text{def}}{=} -\log p\left(\mathbf{y} \mid \mathcal{C}; \mathbf{w}, \mathbf{z}\right) + \lambda_{\text{w}}||\mathbf{w}||_2^2 + \lambda_{\text{z}}||\mathbf{z}||_2^2 \tag{8.24}$$

In addition to weight decay governed by $\lambda_{\text{w}}, \lambda_{\text{z}}$ above, models are regularized using dropout (Srivastava et al., 2014) with drop rate d_{h} at the outputs of all feedforward hidden layers. Random effects are also dropped at rate d_{z}. Randomly removing access to random effects estimates during training is intended to encourage the model to find population-level estimates that accurately reflect central tendency. Finally, the recurrent contribution to the CDRNN hidden state ($\mathbf{h}^{\text{RNN}}$ above) is dropped at rate d_{r}, which is intended to encourage accurate IRF estimation even when context is unavailable.

8.2.5 Effect Estimation

Because it is a DNN, CDRNN lacks parameters that selectively describe the size and shape of the response to a specific predictor (unlike CDR), and indeed individual parameters (e.g. individual biases or connection weights) are not readily interpretable. Thus, from a scientific perspective, the quantity of general interest is not a distribution over parameters, but rather over the *effect* of a predictor on the response. The current study proposes to accomplish this using perturbation analysis (e.g. Ribeiro et al., 2016; Petsiuk et al., 2018), manipulating the input configuration and quantifying the influence of this manipulation on the predicted response.[4] For example, to obtain an estimate of *rate* effects (i.e. a "base" response estimating effects of stimulus timing, see Chapter 3), a reference stimulus can be constructed, and the response to it can be queried at each timepoint over some interval of interest. To obtain CDR-like estimates of predictor-wise IRFs, the reference stimulus can be increased by 1 in the predictor dimension of interest (e.g. word surprisal) and requeried, taking the difference between the obtained response and the reference response to reveal the influence of an extra unit of the predictor. [5] This study uses the training set mean of $\mathbf{x}$ and t as a reference, since this represents the response of the system to an average stimulus. The model also supports arbitrary additional kinds of queries, including of the curvature of an effect in the IRF over time and of the interaction between two effects at a point in time. Indeed, the IRF can be queried with respect to any combination of values for predictors, t, and τ, yielding an open-ended space of queries that can be constructed as needed by the

[4]Perturbation analyses is one of a growing suite of tools for black box interpretation. It is used here because it straightforwardly links properties of the input to changes in the estimated response, providing a highly general method for querying aspects of the the non-linear, non-stationary, non-additive IRF defined by the CDRNN equations.

[5]Note that 1 is used here to maintain comparability of effect estimates to those generated by methods that assume linearity of effects (especially CDR), but that 1 has no special meaning in the non-linear setting of CDRNN modeling, and effects can be queried at any offset from any reference. Results here show that deflections move relatively smoothly away from the reference, even at smaller steps than 1, and that IRFs queried at 1 are similar to those obtained from (linear) CDR, indicating that this method of effect estimation is reliable. Note finally that because predictors are underlyingly rescaled by their training set standard deviations (though plotted at the original scale for clarity), 1 here corresponds to 1 standard unit, as was the case with the CDR estimates discussed in Chapter 4.

researcher.

Because the estimates of interest all derive from the model's predictive distribution, uncertainty about them can be measured with Monte Carlo techniques as long as training involves a stochastic component, such as dropout (Srivastava et al., 2014) or batch normalization (Ioffe and Szegedy, 2015). This study estimates uncertainty using Monte Carlo dropout (Gal and Ghahramani, 2016), which recasts training neural networks with dropout as variational Bayesian approximation of deep Gaussian process models (Damianou and Lawrence, 2013). At inference time, an empirical distribution over responses to an input is constructed by resampling the model (i.e. sampling different dropout masks).[6]

8.2.6 Implementation

The following implementation details are used in these experiments:[7]

- Python Tensorflow (Abadi et al., 2015) implementation (version 1.13)
- Adam optimizer (Kingma and Ba, 2014) with learning rate 0.01 and default Tensorflow parameters
- Batch size $B = 256$
- History window $T = 128$
- Number of layers $L_{\text{RNN}} \in \{0, 1\}$
- Number of layers $L_{\text{in}} = L_{\text{IRF}} = L_{\epsilon} = 3$
- Dimensions $D_{\text{in}(l)} = D_{\text{RNN}(l)} = D_{\text{h}} = D_{\text{IRF}(l)} = D_{\text{error}(l)} = 32$

[6] Initial experiments also explored uncertainty quantification by implemententing CDRNN as a variational Bayesian DNN. Compared to the methods advocated here, the variational approach was more prone to instability, achieved worse fit, and yielded implausibly narrow credible intervals.

[7] $T = 128$ departs from $T = 256$ of Chapter 4 because it saves substantially on compute time without appreciably harming final fit. While $B = 32$ is a small batch size for a batch-normalized network, the reduction here occurs over both the batch and time dimensions, yielding a large enough sample ($BT = 4096$).

- Dropout levels $d_{\text{h}} = d_{\text{r}} = d_{\text{z}} = 0.01$. Higher levels led to strong train/test asymmetry, including badly underestimated variance and poor likelihood. Nevertheless, even with such a low level of dropout, models generalize well, and Monte Carlo estimated confidence intervals are often broad.

- Regularizer levels $\lambda_{\text{w}} = \lambda_{\text{z}} = 10$, except for fMRI data where $\lambda_{\text{z}} = 100$.

- $s_{\text{in}(l)}$, $s_{\text{RNN}(l)}$, $s_{\text{IRF}(l)}$, and $s_{\text{IRF}(l)}$ are set to identity

- RNN-internal activations follow LSTM defaults (tanh and sigmoid activations, see Hochreiter and Schmidhuber, 1997)

- All other activations $s_{(\cdot)}$ are defined as a computationally efficient approximation to the Gaussian error linear unit (GELU, Hendrycks and Gimpel, 2016) used in current state of the art neural language models (Devlin et al., 2019; Radford et al., 2019):

$$\text{GELU}(v) \stackrel{\text{def}}{=} v\,\text{sigmoid}(1.702v) \tag{8.25}$$

- Models use iterate averaging (Polyak and Juditsky, 1992) with exponential moving average decay rate 0.999 (updates after each minibatch)

- For stability, the norm of the global gradient is clipped at 1.

- To aid training, all predictors and responses are underlyingly standardized (centered at 0 and divided by their standard deviations).[8] These transforms are inverted for evaluation, visualization, and likelihood computation, allowing model estimates to be queried on the original scale.

- Convergence is diagnosed following §3.7.2, which defines convergence as lack of statistical correlation with training time over n epochs. Because CDRNN takes an order of

[8] Although Chapter 3 raises concerns about centering predictors in CDR, they do not apply to CDRNN, where effects are estimated *post hoc* from the predictive distribution using perturbation analysis.

magnitude or more time per epoch than CDR, n was reduced from 500 to 100 *vis-a-vis* Chapter 3. Under this criterion, the feedforward implementation of CDRNN usually converges in about the same amount of clock time as CDR (e.g. ∼10 hrs in a non-GPU implementation for Natural Stories SPR), and the recurrent implementation usually takes about 2 times longer.

- 8 cores of Intel Xeon Platinum 8260 2.4GHz for each training run, no GPU. CDRNN-FF models took around 2-4 hrs to converge on synthetic data, 4-10 hrs to converge on eye-tracking and fMRI data, and 8-16 hrs to converge on self-paced reading data. CDRNN-RNN models took about twice as long as to converge as their feedforward counterparts.

Systematic grid searching was infeasible during model development because of the existence of multiple simultaneous desiderata for scientific applications: not only good generalization performance, but also numerical stability during training, consistency of solutions across runs, convergence speed on realistic compute resources (e.g. laptops without GPUs), plausibility of uncertainty estimates, and robustness of settings to changes of domain. Various parameters were subjectively searched with respect to these considerations, including L_2 regularization weight (1, 10, 100), dropout rate (0.01, 0.05, 0.1, 0.2), and learning rate (0.0001, 0.001, 0.01, 0.1). Larger and deeper networks were briefly explored, but dropped due to both increased training time and a greater propensity to overfit.[9] In the end, only one parameter was adjusted for a specific domain: the random effects in fMRI were more heavily regularized (see §8.2.6). The fMRI dataset is noisy, and rich random effects structures can lead to severe overfitting Chapters 4 and 6. Dev set performance revealed a substantial benefit in fMRI from increasing λ_z, which was therefore done. While it is in principle desirable for a regression method to work in a new domain "out of the box", it is well known that DNNs often benefit from domain-specific tuning. Thus, even if a single

[9]For more complex analyses than those attempted here (e.g. models with dozens/hundreds/+ predictors or of very complex IRFs from high-frequency data), more expressive architectures may be warranted.

parameterization had ended up working well for all domains explored here, this could not rule out poor fit to some other domain. Nonetheless, the fact that this was the only parameter that differed across domains suggests that the settings used here are likely a good starting point for future researchers, even those who lack tuning data.

8.3 Methods

	Natural Stories (SPR)						Dundee					
	ms			log-ms			ms			log-ms		
Model	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
LME	19980†	20471†	20230†	0.0789†	0.0807†	0.0803†	13112†	14162†	14024†	0.1507†	0.1532†	0.1526†
GAM	19873	20349	20109	0.0784	0.0802	0.0799	12882	13948	13771	0.1491	0.1518	0.1508
CDR	**18118**	**18373**	**18212**	0.0646	0.0652	0.0654	13073	14106	13960	0.1505	0.1539	0.1520
CDRNN-FF	18360	18482	18317	**0.0622**	**0.0626**	**0.0625**	12829	13877	**13709**	0.1492	0.1515	**0.1505**
CDRNN-RNN	18310	18537	—	0.0640	0.0645	—	**12738**	**13867**	—	**0.1482**	**0.1508**	—

Table 8.2: **Reading.** Mean squared error by model. Baselines as reported in Chapter 4. Daggers (†) indicate convergence failures.

Following Chapter 4, CDRNN is applied to naturalistic human language processing data from three experimental modalities: the Natural Stories self-paced reading corpus (∼1M instances, Futrell et al., 2020), the Dundee eye-tracking corpus (∼200K instances, Kennedy et al., 2003), and the Natural Stories fMRI corpus (∼200K instances, Chapter 6), using the train/dev/test splits for these corpora defined in Chapter 4. Further details about datasets and preprocessing are given in Chapter 4.

For reading data, CDRNN is compared to CDR as well as lagged LME and GAM baselines equipped with four spillover positions for each predictor (values from the current word, plus three preceding words), since LME and GAM are well established analysis methods in psycholinguistics (e.g. Baayen et al., 2007; Demberg and Keller, 2008; Smith and Levy, 2013; Baayen et al., 2017; Goodkind and Bicknell, 2018, *inter alia*). Because the distribution of reading times is heavy-tailed (Frank et al., 2013), following Chapter 4 models are fitted to both raw and log-transformed reading times. For fMRI data, CDRNN is compared to CDR as well as four existing techniques for analyzing naturalistic fMRI data: pre-convolution with the canonical hemodynamic response function (HRF, Brennan et al., 2012; Willems et al., 2015; Henderson et al., 2015, 2016; Lopopolo et al., 2017), linear interpolation (Chapter 4), binning (Wehbe et al., 2020), and Lanczos interpolation (Huth et al., 2016). Statistical model comparisons use paired permutation tests of test set error (Demšar, 2006).

Models use predictors from Chapters 4 and 6, redefined below for convenience. The following predictors are common to all models presented here:

- **Rate** (CDR/NN only): The deconvolutional intercept, i.e. the base response to a stimulus, independent of its features. In CDR, *rate* is estimated explicitly by fitting an IRF to intercept vector (Chapter 3, i.e., implicitly, the response when all predictors are 0). In CDRNN, *rate* is a reference response, computed by taking the response to an average stimulus (since the zero vector may unlikely for a given input distribution, using it as a reference may not reliably reflect the model's domain knowledge). In

this study, all other IRF queries subtract out *rate* in order to show deviation from the reference.

- **Unigram surprisal**: The negative log of the smoothed context-independent probability of a word according to a unigram KenLM model (Heafield et al., 2013) trained on Gigaword 3 (Graff et al., 2007). While this quantity is typically treated on a frequency or log probability scale in psycholinguistics, it is treated here on a surprisal (negative log prob) scale simply for easy of comparison with *5-gram surprisal* (below), even though it is not a good estimate of the quantity typically targeted by surprisal (contextual predictability), since context is ignored.

- **5-gram surprisal**: The negative log of the smoothed probability of a word given the four preceding words according to a 5-gram KenLM model (Heafield et al., 2013) trained on Gigaword 3 (Graff et al., 2007).

The following predictor is used in all reading models:

- **Word length**: The length of the word in characters.

The following predictors are used in eye-tracking models:

- **Saccade length**: The length in words of the incoming saccade (eye movement), including the current word.

- **Previous was fixated**: Indicator for whether the most recent fixation was to the immediately preceding word.

Replications of Chapter 6 use the following additional predictors:

- **PCFG surprisal:** Lexicalized probabilistic context-free grammar surprisal computed using the incremental left-corner parser of van Schijndel et al. (2013) trained on a generalized categorial grammar (Nguyen et al., 2012) reannotation of Wall Street Journal sections 2 through 21 of the Penn Treebank (Marcus et al., 1993).

- **Sound power:** Stimulus sound power (root mean squared energy), averaged over 250ms intervals. This implementation differs slightly from that of Chapter 6, who sampled the measure every 100ms. The longer interval is designed to provide coverage over the extent of the HRF in this study, which uses a shorter history window for computational reasons (128 timesteps instead of 256). Both for computational reasons, especially under CDRNN-RNN (§8.2.3) and because prior *sound power* estimates in this dataset have been weak (Chapters 4 and 6), *sound power* is omitted from models used in the main comparison.

The deconvolutional intercept term *rate*, an estimate of the general influence of observing a stimulus at a point in time, independently of its properties, is implicit in CDRNN (unlike CDR) and is therefore reported in all results. Reading models include random effects by subject, while fMRI models include random effects by subject and by functional region of interest (fROI). Unlike LME, where random effects capture linear differences in effect size between e.g. subjects, random effects in CDRNN capture differences in overall dynamics between subjects, including differences in size, IRF shape, functional form (e.g. linearity), and relation to other effects.

Two CDRNN variants are considered in all experiments: the full model (CDRNN-RNN) containing an RNN over the predictor sequence, and a feedforward only model (CDRNN-FF) with the RNN ablated (gray arrows removed in Figure 8.1). This manipulation is of interest because CDRNN-FF is both more parsimonious (fewer parameters) and faster to train, and may therefore be preferred in the absence of prior expectation that the IRF is sensitive to context. All plots show means and 95% confidence intervals. Code and documentation are available at **https://github.com/coryshain**.

8.4 Results

As argued in Chapter 4, since CDRNN is designed for scientific modeling, the principal output of interest is the IRF itself and the light it might shed on questions of cognitive dynamics, rather than on performance in some task (predicting reading latencies or fMRI measures are not widely targeted engineering goals). However, predictive performance can help establish the trustworthiness of the IRF estimates. To this end, following the analyses in Chapter 4, this section first performs two sanity checks: (1) evaluating recovery of ground truth IRFs from synthetic data, and (2) evaluating predictive performance on human data relative to existing regression techniques. While results may resemble "bake-off" comparisons familiar from machine learning (and indeed CDRNN does outperform all baselines), their primary purpose is to establish that the CDRNN estimates are trustworthy, since they describe the phenomenon of interest in a way that generalizes accurately to an unseen sample. Baseline models, including CDR, are as reported in Chapter 4.[10]

8.4.1 Model Validation A: Synthetic Evaluation

CDRNN models are fitted to a subset of synthetic datasets from Chapter 4. Figure 8.2 shows IRFs from CDRNN models fitted to a representative subset of the synthetic manipulations in Chapter 4: responses with Gaussian noise with standard deviation 10 ($\sigma_\epsilon = 10$), asynchronous and non-uniform time series with mean interval 500ms (Async), and correlation of 0.75 between each pair of the 20 predictors in the model ($\rho = 0.75$). Respectively, these manipulations assess the model's ability to recover the true impulse response structure in the presence of noise, complex temporal structure, and multicollinearity. The true model in each case, along with (non-neural) CDR estimates, are presented for reference.

Several observations emerge. First, both feedforward (FF) and recurrent (RNN) versions of CDRNN recover much of the underlying model, suggesting that CDRNN is capable

[10] For all datasets, the CDR baseline used here is the model variant that was deployed on the test set in the original study.

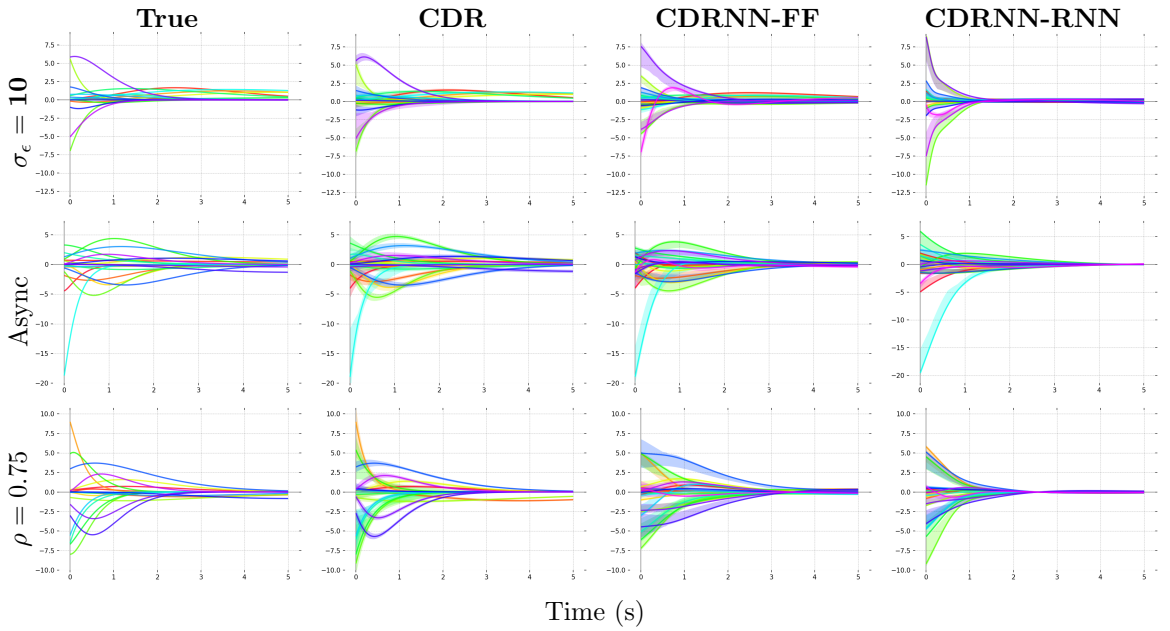


Figure 8.2: **Synthetic Data.** True vs. estimated models under (top) Gaussian noise with standard deviation 10, (middle) asynchronous time series with ∼500ms intervals, and (bottom) multicollinearity of $\rho = 0.75$.

of impulse response identification. Second, while CDRNN-FF closely approximates the ground truth model, it gets less fine-grained detail than CDR. This is because the synthetic data from Chapter 4 match CDR's assumptions: IRFs are stationary and belong to a parametric family that is known in advance, error is constant, effects are linear, and predictors combine additively. CDR gets all of this information for free by virtue of its model specification, while CDRNN must *discover* these properties of the model from data in spite of its highly distributed, interactive, and time-varying forward computation. This problem is even more pronounced in CDRNN-RNN, since the true model is not context-sensitive and the estimates must learn to ignore context when computing the IRF. The RNN component contributes substantially to overfitting on these datasets relative to CDRNN-FF and CDR (hence, poorer IRF recovery), so more data may be necessary in order for CDRNN-RNN to identify the model across synthetic conditions. Although CDRNN's flexibility may be disadvantageous relative to CDR on these simple synthetic datasets, CDRNN still recovers

much of the model, while retaining the ability to adapt to human data where the assumptions of CDR may no longer hold. Finally, CDRNN's uncertainty estimates are generally wider than those of CDR, suggesting empirically that CDRNN may be less susceptible to anticonservative uncertainty intervals Chapter 3.[11] In short, CDRNN provides reasonable approximations to synthetic data that respect the simplifying assumptions of CDR, despite the fact that (unlike CDR) it cannot exploit foreknowledge of these constraints.

8.4.2 Model Validation B: Baseline Comparisons

Table 8.2 gives mean squared error by dataset of CDRNN vs. baseline models on reading times from both Natural Stories and Dundee. Both versions of CDRNN outperform all baselines on the dev partition of all datasets except for raw (ms) latencies in Natural Stories (SPR), where CDRNN is edged out by CDR but still substantially outperforms the non-CDR baselines. Nonetheless, results indicate that CDRNN estimates of Natural Stories (ms) are similarly reliable to those of CDR, and, as discussed in Section 8.4.3, CDRNN largely replicates the CDR estimates on Natural Stories while offering advantages for analysis.

Although CDR struggled against GAM baselines on Dundee in Chapter 4, CDRNN has closed the gap. This is noteworthy in light of speculation in Chapter 4 that CDR's poorer performance on Dundee might be due in part to non-linear effects, which GAM can estimate but CDR cannot. CDRNN performance supports this conjecture: once the model can account for non-linearities, it overtakes GAMs.

Perhaps the most striking gains appear in the fMRI data (Table 8.3), where both CDRNN variants yield substantial improvements to training and dev set error. This suggests that the relaxed assumptions afforded by CDR are beneficial for describing the fMRI

[11] CDRNN plots show a magenta curve not present in the true or CDR plots. This is the deconvolutional intercept *rate*, which is present implicitly in all CDRNN models. The underlying model contains no *rate* effect, and indeed most models find little effect. Investigation of the exceptions (e.g. the $\sigma_\epsilon = 10$, condition, where models find substantial *rate* artifacts) is left to future research.

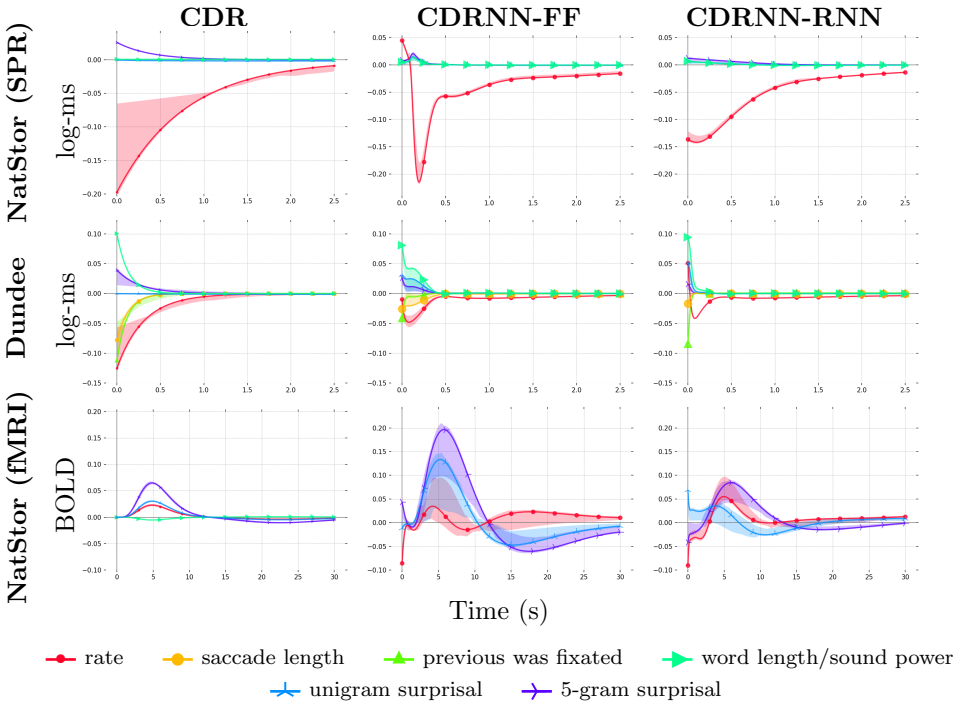


Figure 8.3: CDRNN estimates across datasets, with CDR estimates from Chapter 4 for reference. Sound power omitted from CDRNN fMRI models (see §8.3 for justification).

Model	Train	Expl	Test
Canonical HRF	11.3548†	11.8263†	11.5661†
Interpolated	11.4236†	11.9888†	11.6654†
Averaged	11.3478†	11.9280†	11.6090†
Lanczos	11.3536†	11.9059†	11.5871†
CDR	11.2774	11.6928	11.5369
CDRNN-FF	10.6619	11.4298	**11.4035**
CDRNN-RNN	**10.3311**	**11.4158**	—

Table 8.3: **fMRI.** Mean squared error by model. Baselines as reported in Chapter 4. Daggers (†) indicate convergence failures.

Baseline	Modality	p
LME	Reading	**0.0001*** **
GAM	Reading	**0.0001*** **
Canonical HRF	fMRI	**0.0001*** **
Interpolated	fMRI	**0.0001*** **
Averaged	fMRI	**0.0001*** **
Lanczos	fMRI	**0.0001*** **
CDR	Both	**0.0001*** **

Table 8.4: Permutation test of overall performance improvement from CDRNN-RNN over each baseline.

response, which is known to saturate over time (Friston et al., 2000; Wager et al., 2005; Vazquez et al., 2006; Lindquist et al., 2009).

Following Chapter 4, model error is statistically compared using a paired permutation test that pools across all datasets covered by a given baseline (reading data for LME and GAM, fMRI data for canonical HRF, interpolated, averaged, and Lanczos, and both for CDR).[12] To avoid multiple comparisons, CDRNN-FF alone is evaluated on the test set, selected because it may be preferred in applications: simpler, faster to train, better at recovering synthetic models (§8.4.1), and close in performance to CDRNN-RNN. Results are given in Table 8.4. As shown, CDRNN significantly improves over all baselines. This result supports the reliability of patterns revealed by CDRNN's estimated IRF, which is now used to explore and visualize sentence processing dynamics.

8.4.3 Effect Latencies in CDRNN vs. CDR

CDR-like IRF estimates can be obtained by increasing a predictor by 1 relative to the reference and observing the change in the response over time. Visualizations using this approach are presented in Figure 8.3 alongside CDR estimates from Chapter 4. In general, CDRNN finds similar patterns to CDR. This suggests both (1) that CDRNN is capable of recovering estimates from a preceding state-of-the-art deconvolutional model for these domains, and

[12]The comparison rescales each pair of error vectors by their joint standard deviation in order to enable comparability across datasets with different error variances.

(2) that CDR estimates in these domains are not driven by artifacts introduced by its simplifying assumptions, since a model that lacks those assumptions and has a qualitatively different architecture largely recovers them. Nonetheless there are differences. For example, Dundee estimates decay more quickly over time in CDRNN than in CDR, indicating an even less pronounced influence of temporal diffusion in eye-tracking than CDR had previously suggested. Estimates from CDRNN-FF and CDRNN-RNN generally agree, except that CDRNN-RNN estimates tend to be more attenuated (especially in fMRI). CDR shows little uncertainty in the fMRI domain despite its inherent noise Chapter 6, while CDRNN more plausibly shows more uncertainty in its estimates for the noisier fMRI data.

As noted in Section 8.1, Chapter 4 reports large negative *rate* effects in reading — i.e., a local decrease in subsequent reading time at each word, especially in Natural Stories. This was interpreted as an inertia effect (faster recent reading engenders faster current reading), but it could conceivably also be an artifact of non-linear decreases in latency over time (due to task habituation, e.g. Baayen et al., 2017; Harrington Stack et al., 2018; Prasad and Linzen, 2019) that CDR cannot easily identify. CDRNN estimates suggest that habituation does not fully explain these patterns and thus support the prior interpretation of *rate* effects as inertia, at least in SPR: a model that can flexibly adapt to such trends nonetheless finds SPR rate estimates that are similar in shape and magnitude to those estimated by CDR. In eye-tracking, the negative *rate* effect still emerges, but it is much weaker, suggesting that eye movements may be less prone than SPR to inertia in reading speed.

In addition, note that CDRNN-FF finds a late-peaking response to word surprisal in self-paced reading (at around 200ms) but not in eye-tracking. This result converges with word-discretized timecourses reported in Smith and Levy (2013), who find a peak surprisal response on the current word in an eye-tracking experiment but on the following word in an SPR experiment. While CDR and CDRNN-RNN find relatively slow surprisal effects in SPR, they do not show a late peak. However, CDRNN-FF generalizes better than the other two variants on this dataset, suggesting that its estimates are more reliable. Results thus

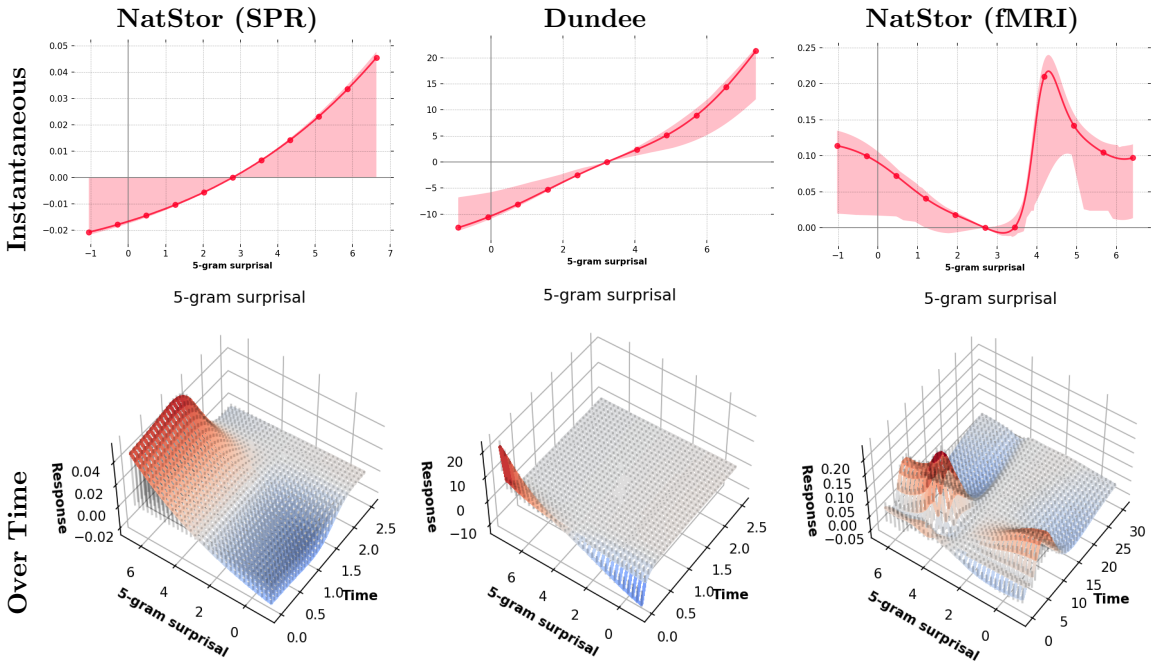


Figure 8.4: Functional curvature of *5-gram surprisal* response

reveal important hidden dynamics in the reading response (inertia effects), continuous-time delays in peak response, and influences of modality on measures of sentence processing, all of which are difficult to estimate using existing regression techniques. Greater response latency and more pronounced inertia effects in self-paced reading may be due to the fact that a gross motor task (paging via button presses) is overlaid on the sentence comprehension task. While the motor task is not generally of interest to psycholinguistic theories, controlling for its effects is crucial when using self-paced reading to study sentence comprehension (Mitchell, 1984).

8.4.4 Linearity of Surprisal Effects

CDRNN also allows the analyst to explore other aspects of the IRF, such as functional curvature at a point in time. For example, in the context of reading, Smith and Levy (2013) argue for a linear increase in processing cost as a function of word surprisal. The present study allows this claim to be assessed across modalities by checking the curvature

of the *5-gram surprisal* response (in raw ms) at a timepoint of interest (0ms for reading and ∼5s for fMRI). As shown in the top row of Figure 8.4, reading estimates are consistent with a linear response (the confidence interval contains a straight line), as predicted, but are highly non-linear in fMRI, with a rapid peak above the mean (center of x-axis) followed by a sharp dip and plateau, and even an estimated increased response at values below the mean (though this component has high uncertainty). This may be due in part to ceiling effects: blood oxygen levels measured by fMRI are bounded, but reading times are not. While this is again a property of experimental modality rather than sentence comprehension itself, understanding such influences is important for drawing scientific conclusions from experimental data. For example, due to the possibility of saturation, fMRI may not be an ideal modality for testing scientific claims about the functional form of effects, and the linearity assumptions of e.g. CDR and LME may be particularly constraining.

The curvature of effects can also be queried over time. If an effect is temporally diffuse but linear, its curvature should be roughly linear at any delay of interest. The second row of Figure 8.4 shows visualizations to this effect. These plots in fact subsume previously queried estimates: univariate IRFs to *5-gram surprisal* like those plotted in Figure 8.3 are simply slices taken at a predictor value (1 unit above the mean), whereas curvature estimates in the first row of Figure 8.4 are simply slices taken at a time value (0s for reading and 5s for fMRI). Plots are consistent with the linearity hypothesis for reading, but again show strong non-linearities in the fMRI domain that are consistent with saturation effects as discussed above.

8.4.5 Effect Interactions

In addition to exploring multivariate relationships of a predictor with time, relationships between predictors can also be studied. Such relationships constitute "interactions" in a CDRNN model, though they are not constrained (cf. interactions in linear models) to be strictly multiplicative — indeed, a major advantage of CDRNN is that interactions come

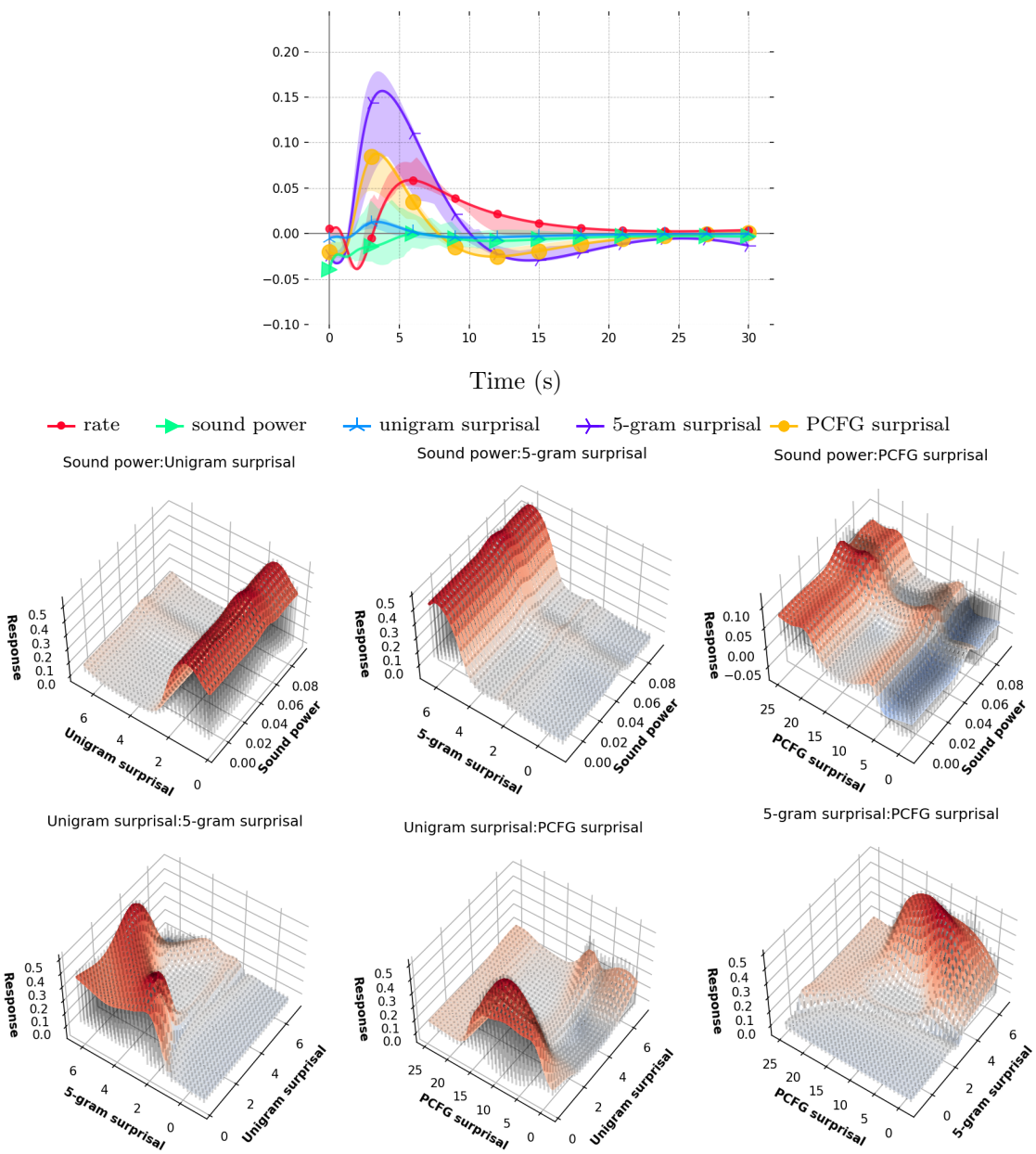


Figure 8.5: Effect interactions in CDRNN replication of Chapter 6.

"for free", along with estimates of their functional form. To explore effect interactions, a CDRNN-FF version of the full model in Chapter 6 is fitted to the fMRI dataset. The model contains more predictors to explore than models considered above, including surprisal computed from a probabilistic context-free grammar (*PCFG surprisal*, see §8.3). Univariate IRFs are shown in the top left panel of Figure 8.5, and pairwise interaction surfaces at a delay of 5s (near the peak response) are shown in the remaining panels. Plots show that the response at any value of the other predictors is roughly flat as a function of *sound power* (middle row). This accords with prior arguments that the cortical language system, whose activity is measured here, does not strongly register low-level perceptual effects (Fedorenko et al., 2010; Braze et al., 2011).

The estimate for *unigram surprisal* (middle left) also shows an unexpected non-linearity: although activity increases with higher surprisal (lower frequency words), it increases much more at lower surprisal (higher frequency words). This spike is primarily restricted to the lower part of the range of *unigram surprisal* values attested in the data (min = 1.29) and should therefore be interpreted with caution. The response starts to rise well before this value, however, suggesting the possible existence of high frequency items that nonetheless engender a large response. The interaction between *PCFG surprisal* and *unigram surprisal* possibly sheds light on this outcome, since it shows a sharp increase in the *PCFG surprisal* response for high frequency items. This may be because the most frequent words in English tend to be function words that play an outsized role in syntactic structure building (e.g. prepositional phrase attachment decisions).

In addition, *5-gram surprisal* interacts with *PCFG surprisal*, showing a large increase in response for words that are high on both measures. This is consistent with a unitary predictive mechanism that sends strong error signals when both string-level (5-gram) and structural (PCFG) cues are poor. The response is estimated to diminish at high values (>20) of *PCFG surprisal* in the bottom center and right plots. High *PCFG surprisal* values may be driven by poor automatic parsing, which might create misalignment with human

subjective surprisal. Again, all these interactions should be interpreted with caution, since the uncertainty interval covers much weaker degrees of interaction.

8.5 Conclusion

This chapter proposed and evaluated CDRNN, a deep neural extension of continuous-time deconvolutional regression that relaxes implausible simplifying assumptions made by widely used regression techniques in psycholinguistics. In so doing, CDRNN provides detailed estimates of human language processing dynamics that are difficult to obtain using other measures. Results showed plausible estimates from human data that generalize better than alternatives and can be brought to bear both for testing existing scientific hypotheses and exploring hitherto understudied properties of the human sentence processing response. This outcome suggests that CDRNN may play a valuable role in analyzing human experimental data.

Part V

Conclusion

Chapter 9

Conclusion

In this thesis, I have advocated a dynamical systems perspective on experimental measures of human language processing and argued that one major challenge emerging from this perspective is the potential influence of effects with delayed timecourses (i.e. effects that are *temporally diffuse*), possibly due to bottlenecks in real-time human information processing (Bouma and De Voogd, 1974; Ehrlich and Rayner, 1981; Mitchell, 1984; Mollica and Piantadosi, 2017). In Chapter 2, I argued (1) that such timecourses are of theoretical importance to many questions in psycholinguistics, (2) that failure to account for temporal diffusion can have a serious impact on the outcome of psycholinguistic hypothesis testing, even for questions that do not directly concern timecourses, and (3) that existing discrete-time approaches to measuring effect latencies — such as "spillover" regression — are ill-suited to many studies of language processing, since stimulus events (words) have variable duration.

In response to these concerns, in Chapter 3 I proposed and implemented a new framework for analyzing experimental measures of human sentence processing: continuous-time deconvolutional regression (CDR). CDR directly estimates the changing influence of stimuli on future responses as a function of continuous time, allowing direct application to data with variably spaced events (e.g. reading time measures) and/or asynchronously measured predictors and responses. CDR thus both controls for and illuminates the structure of the dynamics in measures of human sentence processing. In Chapter 4, I presented an empirical evaluation of the CDR approach and showed that it (1) reliably recovers the ground truth model from synthetic data, even under adverse training conditions like noise, multicollinearity, and impulse response misspecification, and (2) yields plausible, fine-grained estimates

of temporal dynamics in human reading and fMRI data that generalize better than estimates from established tools like linear mixed effects models (LME) and generalized additive models (GAM).

I then deployed CDR to study the influence of theory-driven variables in reading and fMRI measures of naturalistic human sentence processing. In Chapter 5, I reported the results of a CDR analysis of word frequency and predictability effects in three large-scale public reading time corpora, which did not support the existence of distinct lexical prediction and retrieval mechanisms in naturalistic reading, in contrast to much prior work using constructed stimuli (Staub, 2015). In Chapters 6 and 7, I analyzed a large-scale fMRI dataset with respect to the existence and functional specificity of theory-driven measures of language processing difficulty. Chapter 6 used CDR to investigate (1) whether surprisal effects register primarily in a language-specialized cortical network or a multiple demand network that houses domain general executive control and (2) whether surprisal effects are sensitive to syntactic structure independently of string-level co-occurrence patterns. Results showed both string-level and structural sensitivity in surprisal effects in the functional language network, but no sensitivity to either kind of surprisal in the multiple demand network, supporting the existence of predictive mechanisms for language that are structure sensitive and domain-specific. Chapter 7 investigated the role of syntax-related memory retrieval in naturalistic sentence comprehension, finding a substantial effect of integration cost (a measure of retrieval difficulty) over strong controls for word predictability. This result again obtained only in the language (but not the MD) network, supporting a central role for working memory in naturalistic language comprehension, with working memory resources that reside in dedicated, language-specific cortical circuits, rather than in domain-general executive areas.

In Chapter 8, I proposed and evaluated a deep neural extension of CDR — the continuous-time regressive neural network (CDRNN) — that relaxes many of the simplifying assumptions of CDR while retaining its deconvolutional interpretation. Results showed that

CDRNN generalizes better than CDR, LME, and GAM baselines on human data, while supporting novel and theoretically relevant insights about the functional form of effects and effect interactions over time that are difficult to obtain using other methods. CDRNN is thus a promising next step toward detailed modeling of human experimental measures within a dynamical systems framework.

There are several feasible directions in which the research program advanced in this thesis might be extended in future work. First, a major potential untapped application area for CDR research in language processing and cognition is high temporal resolution measures like electroencephalography (EEG), electrocorticography (ECoG), and magnetoencephalography (MEG), all of which reveal fine-grained fluctuations in neuronal activity more directly than the fMRI measures considered here. These measures are well known to exhibit systematic deviations that unfold over time in response to an event (Kutas and Hillyard, 1980): event-related potentials (ERPs) in EEG or event-related fields (ERFs) in MEG (for simplicity, henceforth jointly called ERPs). For example, the well-known n400 ERP component, so named because it is realized as a negative deflection of electrical potential peaking around 400ms after stimulus onset, has been reported in response to a wide range of linguistic manipulations (see Kutas and Federmeier, 2011, for review). ERPs are continuous-time impulse response functions, much like those that initially motivated the CDR design. Because of the inherent latency in some ERP components, responses almost certainly overlap during typical sentence comprehension (Smith and Kutas, 2015). Standard designs in EEG research reduce this overlap experimentally by presenting words one by one at an unnaturally slow rate (Kutas and Hillyard, 1984), including in studies explicitly attempting to improve naturalness (Frank et al., 2015). CDR, by contrast, stands to support ERP identification even in truly naturalistic designs where responses are free to overlap, since it has been shown here to do so accurately from synthetic data that are known to exhibit such overlaps and from human data that likely exhibit such overlaps. CDR may thus enable more widespread use of e.g. free reading or naturalistic auditory presentation

of stimuli in electrophysiological studies of human sentence processing.

A second direction for future research is to implement the CDR equivalent of logistic regression. This would enable the use of CDR to study the lingering influence of stimuli on the probabilities associated with discrete events, such as the probability of executing a regressive eye movement during reading. Doing so simply requires replacing the normal predictive distribution of eq. 3.7 with e.g. a binomial distribution whose parameter is generated via convolution of the predictors over time.

A third direction for future research is to support multidimensional response variables (that is, where each response is a vector rather than a scalar). As in logistic regression, this requires a change of predictive distribution. In this case, one approach would be to require the model to generate the mean of a multivariate normal distribution, with variance-covariance parameters learned from data. Similarly, in the context of logistic regression, multivariate CDR models could be used to learn distributions over many categories (rather than just two).

A final direction for future research is to use CDR for neural decoding (Mitchell et al., 2008), i.e. making inferences about mental representations on the basis of brain signals alone. Standard approaches in this domain use isolated stimulus presentation (e.g. of a single picture or word) along with a form of finite impulse response modeling that fits timestep-specific classifier weights (Mitchell et al., 2008). Thanks to its ability to recover response dynamics from data in which responses overlap, CDR might enable direct decoding of the mental representations associated with words in running speech, thus shedding light on the changes in mental state that are of direct relevance to theories of incremental sentence processing (Gibson, 2000; Lewis and Vasishth, 2005; Levy, 2008). CDRNN-based decoders additionally stand to relax the spatial independence assumptions of standard decoding studies, which regularly use linear regression/classification in order to maintain interpretability as to which brain regions encode the constructs of interest (e.g. Pereira et al., 2018). CDRNN might serve as a deeper, more flexible neural network model that

can account for non-linear interactions between voxels in representing a construct, while still supporting spatial inferences thanks to the perturbation analysis approach discussed in Chapter 8.

In sum, in addition to the merits of the CDR(NN) approach established empirically in this thesis, there are many more potential application areas in which CDR might fruitfully shed light on the timecourses of human mental representations associated with language processing. CDR(NN) therefore stands to further illuminate core questions in the study of language and the mind.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Abney, S. P. and Johnson, M. (1991). Memory Requirements and Local Ambiguities of Parsing Strategies. *J.\ Psycholinguistic Research*, 20(3):233–250.

Aho, A. V. and Ullman, J. D. (1972). *The Theory of Parsing, Translation and Compiling, Vol. 1: Parsing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30(8):2960–2966.

Almor, A. (1999). Noun-Phrase Anaphora and Focus: The Informational Load Hypothesis. *Psychological Review*, 106(4):748–765.

Altmann, G. (2010). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta psychologica*, 137:190–200.

Altmann, G. T. M. (1998). Ambiguity in sentence processing. *Trends in cognitive sciences*, 2(4):146–152.

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

Amalric, M. and Dehaene, S. (2018). Cortical circuits for mathematical knowledge: evidence for a major subdivision within the brain's semantic networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1740):20160515.

Amici, S., Brambati, S. M., Wilkins, D. P., Ogar, J., Dronkers, N. L., Miller, B. L., and Gorno-Tempini, M. L. (2007). Anatomical correlates of sentence comprehension and verbal working memory in neurodegenerative disease. *Journal of Neuroscience*, 27(23):6282–6290.

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B. M., and Zilles, K. (1999). Broca's region revisited: cytoarchitecture and intersubject variability. *Journal of Comparative Neurology*, 412(2):319–341.

Anderson, J. (2004). Modern grammars of case: a personal history. Lectures, Universit{é} de Toulouse-Le Mirail.

Aoshima, S., Phillips, C., and Weinberg, A. (2004). Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language*, 51:23–54.

Armeni, K., Willems, R. M., den Bosch, A., and Schoffelen, J.-M. (2019). Frequency-specific brain dynamics related to prediction during language comprehension. *NeuroImage*, 198:283–295.

Ashby, J., Rayner, K., and Clifton, C. (2005). Eye movements of highly skilled and average readers: Differential effects of frequency and predictability. *The Quarterly Journal of Experimental Psychology Section A*, 58(6):1065–1086.

Assem, M., Blank, I. A., Mineroff, Z., Ademo\uglu, A., and Fedorenko, E. (2020). Activity in the fronto-parietal multiple-demand network is robustly associated with individual differences in working memory and fluid intelligence. *Cortex*, 131:1–16.

Aurnhammer, C. and Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, 134:107198.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Baayen, H., Vasishth, S., Kliegl, R., and Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94(Supplement C):206–234.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (2007). Mixed effects modelling with crossed random effects for subjects and items. manuscript.

Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Speelman, D., Heylen, K., and Geeraerts, D., editors, *Mixed Effects Regression Models in Linguistics*. Springer, Berlin.

Baddeley, A., Allen, R., and Vargha-Khadem, F. (2010). Is the hippocampus necessary for visual and verbal binding in working memory? *Neuropsychologia*, 48(4):1089–1095.

Baddeley, A. and Warrington, E. K. (1970). Amnesia and the distinction between long-and short-term memory. *Journal of verbal learning and verbal behavior*, 9(2):176–189.

Baddeley, A. D., Thomson, N., and Buchanan, M. (1975). Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior*, 15(6):575–589.

Baggio, G. and Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes*, 26(9):1338–1367.

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

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68:255–278.

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., and Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, 76(4):695–711.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Baumgaertner, A., Weiller, C., and Büchel, C. (2002). Event-related fMRI reveals cortical sites involved in contextual sentence integration. *Neuroimage*, 16(3):736–745.

Beckmann, C. F., Jenkinson, M., and Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage*, 20(2):1052–1063.

Bemis, D. K. and Pylkkänen, L. (2011). Simple composition: a magnetoencephalography investigation into the comprehension of minimal linguistic phrases. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(8):2801–2814.

Bergstrom, A. R. (1984). Continuous time stochastic models and issues of aggregation over time. In Griliches and Intriligator, editors, *Handbook of Econometrics*, volume 2, pages 1145–1212. Elsevier.

Bhattasali, S., Fabre, M., Luh, W.-M., Al Saied, H., Constant, M., Pallier, C., Brennan, J. R., Spreng, R. N., and Hale, J. (2019). Localising memory retrieval and syntactic composition: An fMRI study of naturalistic language comprehension. *Language, Cognition and Neuroscience*, 34(4):491–510.

Bhattasali, S., Hale, J., Pallier, C., Brennan, J., Luh, W.-M., and Spreng, R. N. (2018). Differentiating Phrase Structure Parsing and Memory Retrieval in the Brain. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 74–80.

Binnick, R. I. (1991). *Time and the verb: A guide to tense and aspect*. Oxford University Press.

Blanco-Elorrieta, E. and Pylkkänen, L. (2017). Bilingual language switching in the laboratory versus in the wild: The spatiotemporal dynamics of adaptive language control. *Journal of Neuroscience*, 37(37):9022–9036.

Blank, I., Balewski, Z., Mahowald, K., and Fedorenko, E. (2016). Syntactic processing is distributed across the language system. *Neuroimage*, 127:307–323.

Blank, I. and Fedorenko, E. (2017). Domain-general brain regions do not track linguistic input as closely as language-selective regions. *Journal of Neuroscience*, pages 3616–3642.

Blank, I., Kiran, S., and Fedorenko, E. (2017). Can neuroimaging help aphasia researchers? Addressing generalizability, variability, and interpretability. *Cognitive neuropsychology*, 34(6):377–393.

Blumstein, S. E. (2009). Auditory word recognition: Evidence from aphasia and functional neuroimaging. *Language and Linguistics Compass*, 3(4):824–838.

Blumstein, S. E. and Amso, D. (2013). Dynamic functional organization of language: insights from functional neuroimaging. *Perspectives on Psychological Science*, 8(1):44–48.

Bonhage, C. E., Mueller, J. L., Friederici, A. D., and Fiebach, C. J. (2015). Combined eye tracking and fMRI reveals neural basis of linguistic predictions during sentence comprehension. *Cortex*, 68:33–47.

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., and Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1):1–12.

Botvinick, M. (2007). Multilevel structure in behavior and in the brain: A computational model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences*, 362:1615–1626.

Bouma, H. and De Voogd, A. H. (1974). On the control of eye saccades in reading. *Vision Research*, 14(4):273–284.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221.

Braze, D., Mencl, W. E., Tabor, W., Pugh, K. R., Constable, R. T., Fulbright, R. K., Magnuson, J. S., Van Dyke, J. A., and Shankweiler, D. P. (2011). Unification of sentence processing via ear and eye: An fMRI study. *cortex*, 47(4):416–431.

Breen, M. (2014). Empirical investigations of the role of implicit prosody in sentence processing. *Language and Linguistics Compass*, 8(2):37–50.

Brennan, J. (2016). Naturalistic sentence comprehension in the brain. *Language and Linguistics Compass*, 10(7):299–313.

Brennan, J. and Hale, J. T. (2019). Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741.

Brennan, J., Nir, Y., Hasson, U., Malach, R., Heeger, D. J., and Pylkkänen, L. (2012). Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163–173.

Brennan, J. and Pylkkänen, L. (2012). The time-course and spatial distribution of brain activity associated with sentence processing. *NeuroImage*, pages 1–10.

Brennan, J. and Pylkkänen, L. (2017). MEG evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science*, 41:1515–1531.

Brennan, J., Stabler, E. P., Van Wagenen, S. E., Luh, W.-M., and Hale, J. T. (2016). Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94.

Broca, P. (1861). Remarks on the seat of the faculty of articulated language, following an observation of aphemia (loss of speech). *Bulletin de la Société Anatomique*, 6:330–357.

Brothers, T. and 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.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., and Amodei, D. (2020). Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems 33*.

Bubic, A., Von Cramon, D. Y., and Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, 4:25.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *R Journal*, 10(1).

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

Campbell, K. L. and Tyler, L. K. (2018). Language-related domain-specific and domain-general systems in the human brain. *Current Opinion in Behavioral Sciences*, 21:132–137.

Caplan, D. and Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, 22(1):77–94.

Carpenter, P. A. and Just, M. A. (1983). What your eyes do while your mind is reading. In Rayner, K., editor, *Eye movements in reading: Perceptual and language processes*, pages 275–307. Academic Press.

Carrasco, M. and McElree, B. (2001). Covert attention accelerates the rate of visual information processing. *Proceedings of the National Academy of Sciences of the United States of America*, 98:5363–5367.

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

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

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

Cho, S.-J., Brown-Schmidt, S., and Lee, W.-y. (2018). Autoregressive generalized linear mixed effect models with crossed random effects: an application to intensive binary time series eye-tracking data. *Psychometrika*, 83(3):751–771.

Cole, M. W. and Schneider, W. (2007). The cognitive control network: integrated cortical regions with dissociable functions. *Neuroimage*, 37(1):343–360.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., and Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Cooper, J. C. B. (2005). The poisson and exponential distributions. *Mathematical Spectrum*, 37(3):123–125.

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

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201.

Cristescu, T. C., Devlin, J. T., and Nobre, A. C. (2006). Orienting attention to semantic categories. *Neuroimage*, 33(4):1178–1187.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990? *Cognition*, 10:65–70.

Dagerman, K. S., MacDonald, M. C., and Harm, M. W. (2006). Aging and the use of context in ambiguity resolution: Complex changes from simple slowing. *Cognitive Science*, 30(2):311–345.

Dahan, D. (2010). The Time Course of Interpretation in Speech Comprehension. *Current Directions in Psychological Science - CURR DIRECTIONS PSYCHOL SCI*, 19:121–126.

Damianou, A. and Lawrence, N. D. (2013). Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR.

Dave, S., Brothers, T. A., Traxler, M. J., Ferreira, F., Henderson, J. M., and Swaab, T. Y. (2018). Electrophysiological evidence for preserved primacy of lexical prediction in aging. *Neuropsychologia*, 117:135–147.

Dax, G. (1863). Observations tendant à prouver la co{{\textbackslash}"{\textbackslash}i}ncidence constante des dérangements de la parole avec une lésion de l'hémisphère gauche du cerveau. *CR Acad Sci Hebd Seances Acad Sci*, 61:534.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., and Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, pages 3216–3267.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., and Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284(5416):970–974.

Dehghani, M., Boghrati, R., Man, K., Hoover, J., Gimbel, S. I., Vaswani, A., Zevin, J. D., Immordino-Yang, M. H., Gordon, A. S., Damasio, A., and others (2017). Decoding the neural representation of story meanings across languages. *Human brain mapping*, 38(12):6096–6106.

Delogu, F., Crocker, M. W., and Drenhaus, H. (2017). Teasing apart coercion and surprisal: Evidence from eye-movements and ERPs. *Cognition*, 161:46–59.

Demberg, V. and Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

Desai, R. H., Choi, W., Lai, V. T., and Henderson, J. M. (2016). Toward semantics in the wild: activation to manipulable nouns in naturalistic reading. *Journal of Neuroscience*, 36(14):4050–4055.

D'Esposito, M. and Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual review of psychology*, 66.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL19*.

Diachek, E., Blank, I., Siegelman, M., and Fedorenko, E. (2020). The domain-general multiple demand (MD) network does not support core aspects of language comprehension: A large-scale fMRI investigation. *Journal of Neuroscience*, 40(23):4536–4550.

Dien, J., Franklin, M. S., Michelson, C. A., Lemen, L. C., Adams, C. L., and Kiehl, K. A. (2008). fMRI characterization of the language formulation area. *Brain Research*, 1229:179–192.

Dozat, T. (2016). Incorporating Nesterov momentum into Adam. In *ICLR Workshop*.

Dronkers, N. F., Wilkins, D. P., Van Valin Jr, R. D., Redfern, B. B., and Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, 92(1-2):145–177.

Duffau, H., Moritz-Gasser, S., and Mandonnet, E. (2014). A re-examination of neural basis of language processing: Proposal of a dynamic hodotopical model from data provided by brain stimulation mapping during picture naming. *Brain and language*, 131:1–10.

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

Duncan, J. (2013). The structure of cognition: Attentional episodes in mind and brain. *Neuron*, 80(1):35–50.

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

Egner, T., Monti, J. M. P., Trittschuh, E. H., Wieneke, C. A., Hirsch, J., and Mesulam, M.-M. (2008). Neural integration of top-down spatial and feature-based information in visual search. *Journal of Neuroscience*, 28(24):6141–6151.

Ehrlich, S. F. and Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Embrechts, P., Liniger, T., and Lin, L. (2011). Multivariate Hawkes processes: an application to financial data. *J. Appl. Probab.*, 48A:367–378.

Engbert, R. and Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, 43(9):1035–1045.

Engbert, R., Nuthmann, A., Richter, E. M., and Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, 112:777–813.

Erlich, K. and Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior*, 22:75–87.

Federmeier, K. D. and Kutas, M. (2005). Aging in context: age-related changes in context use during language comprehension. *Psychophysiology*, 42(2):133–141.

Federmeier, K. D., Kutas, M., and Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and language*, 115(3):149–161.

Federmeier, K. D., McLennan, D. B., De Ochoa, E., and Kutas, M. (2002). The impact of semantic memory organization and sentence context information on spoken language processing by younger and older adults: An ERP study. *Psychophysiology*, 39(2):133–146.

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in psychology*, 5:335.

Fedorenko, E., Behr, M. K., and Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*.

Fedorenko, E. and Blank, I. (2020). Brocas Area Is Not a Natural Kind. *Trends in Cognitive Sciences*.

Fedorenko, E., Duncan, J., and Kanwisher, N. (2012). Language-selective and domain-general regions lie side by side within Brocas area. *Current Biology*, 22(21):2059–2062.

Fedorenko, E., Duncan, J., and Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, page 201315235.

Fedorenko, E., Gibson, E., and Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain-specific working memory resources. *Journal of Memory and Language*, 54(4):541–553.

Fedorenko, E., Gibson, E., and Rohde, D. (2007). The nature of working memory in linguistic, arithmetic and spatial integration processes. *Journal of Memory and Language*, 56(2):246–269.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., and Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194.

Fedorenko, E. and Kanwisher, N. (2009). Neuroimaging of language: why hasn't a clearer picture emerged? *Language and Linguistics Compass*, 3(4):839–865.

Fedorenko, E., Mineroff, Z., Siegelman, M., and Blank, I. (2018). Word meanings and sentence structure recruit the same set of fronto-temporal regions during comprehension. *bioRxiv*.

Fedorenko, E. and Thompson-Schill, S. L. (2014). Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126.

Felser, C., Clahsen, H., and Münte, T. F. (2003). Storage and integration in the processing of filler-gap dependencies: An ERP study of topicalization and wh-movement in German. *Brain and Language*, 87(3):345–354.

Ferreira, F. and Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current directions in psychological science*, 27(6):443–448.

Ferreira, F. and Clifton, C. (1986). The independence of syntactic processing. *Journal of memory and language*, 25(3):348.

Fiebach, C. J., Schlesewsky, M., and Friederici, A. D. (2001). Syntactic working memory and the establishment of filler-gap dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research*, 30(3):321–338.

Fiebach, C. J., Schlesewsky, M., Lohmann, G., Von Cramon, D. Y., and Friederici, A. D. (2005). Revisiting the role of Broca's area in sentence processing: Syntactic integration versus syntactic working memory. *Human brain mapping*, 24(2):79–91.

Findelsberger, E., Hutzler, F., and Hawelka, S. (2019). Spill the load: Mixed evidence for a foveal load effect, reliable evidence for a spillover effect in eye-movement control during reading. *Attention Perception and Psychophysics*, 81(5):1442–1453.

Fischl, B., Rajendran, N., Busa, E., Augustinack, J., Hinds, O., Yeo, B. T. T., Mohlberg, H., Amunts, K., and Zilles, K. (2007). Cortical folding patterns and predicting cytoarchitecture. *Cerebral cortex*, 18(8):1973–1980.

Fodor, J. A. (1983). *Modularity of Mind*. MIT Press, Cambridge.

Fossum, V. and Levy, R. (2012). Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of {{CMCL}} 2012*. Association for Computational Linguistics.

Frank, S. and Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*.

Frank, S. L., Bod, R., and Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society B: Biological Sciences*, 279(1747):4522–4531.

Frank, S. L. and Christiansen, M. H. (2018). Hierarchical and sequential processing of language. *Language, Cognition and Neuroscience*, 33(9):1213–1218.

Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., and Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

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

Frazier, L. and Fodor, J. D. (1978). The sausage machine: a new two-stage parsing model. *Cognition*, 6:291–325.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392.

Friederici, A. D. (2012). The cortical language circuit: From auditory perception to sentence comprehension. *Trends in Cognitive Sciences*, 16(5):262–268.

Friederici, A. D., Bahlmann, J., Friedrich, R., and Makuuchi, M. (2011). The neural basis of recursion and complex syntactic hierarchy. *Biolinguistics*, 5(1–2):87–104.

Friederici, A. D., Fiebach, C. J., Schlesewsky, M., Bornkessel, I. D., and von Cramon, Y. D. (2006). Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex*, 16(12):1709–1717.

Friederici, A. D., Rueschemeyer, S.-A., Hahne, A., and Fiebach, C. J. (2003). The role of left inferior frontal and superior temporal cortex in sentence comprehension: Localizing syntactic and semantic processes. *Cerebral Cortex*, 13(2):170–177.

Frisson, S., Rayner, K., and Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):862.

Friston, K. J., Fletcher, P., Josephs, O., Holmes, A., Rugg, M. D., and Turner, R. (1998a). Event-related fMRI: Characterizing differential responses. *Neuroimage*, 7(1):30–40.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210.

Friston, K. J., Josephs, O., Rees, G., and Turner, R. (1998b). Nonlinear event-related responses in fMRI. *Magn. Reson. Med*, pages 41–52.

Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. (2000). Nonlinear responses in fMRI: The Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477.

Frost, M. A. and Goebel, R. (2012). Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. *Neuroimage*, 59(2):1369–1381.

Futrell, R., Gibson, E., Tily, H. J. ., Blank, I., Vishnevetsky, A., Piantadosi, S., and Fedorenko, E. (2018). The Natural Stories Corpus. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., and Fedorenko, E. (2020). The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, pages 1–15.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Gambi, C., Gorrie, F., Pickering, M. J., and Rabagliati, H. (2018). The development of linguistic prediction: Predictions of sound and meaning in 2- to 5-year-olds. *Journal of Experimental Child Psychology*, 173:351–370.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.

Gibson, E. (2000). The Dependency Locality Theory: A distance-based theory of linguistic complexity. In Marantz, A., Miyashita, Y., and O'Neil, W., editors, *Image, language, brain*, pages 95–106. MIT Press, Cambridge.

Gibson, E. and Ko, K. (1998). An integration-based theory of computational resources in sentence comprehension. In *Fourth Architectures and Mechanisms in Language Processing Conference*.

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., and Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, 59(1):23–59.

Gilmore, R. O., Diaz, M. T., Wyble, B. A., and Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1396(1):5–18.

Gimel'farb, G., Farag, A. A., and El-Baz, A. (2004). Expectation-Maximization for a linear combination of Gaussians. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 422–425. IEEE.

Gitelman, D. R., Penny, W. D., Ashburner, J., and Friston, K. J. (2003). Modeling regional and psychophysiologic interactions in fMRI: the importance of hemodynamic deconvolution. *Neuroimage*, 19(1):200–207.

Gläscher, J., Rudrauf, D., Colom, R., Paul, L. K., Tranel, D., Damasio, H., and Adolphs, R. (2010). Distributed neural system for general intelligence revealed by lesion mapping. *Proceedings of the National Academy of Sciences*, 107(10):4705–4709.

Gloor, P. (1997). *The temporal lobe {\textbackslash}& limbic system*. Oxford University Press, Oxford.

Gold, B. T., Balota, D. A., Jones, S. J., Powell, D. K., Smith, C. D., and Andersen, A. H. (2006). Dissociation of automatic and strategic lexical-semantics: Functional magnetic resonance imaging evidence for differing roles of multiple frontotemporal regions. *Journal of Neuroscience*, 26(24):6523–6532.

Goldberg, I. I., Harel, M., and Malach, R. (2006). When the brain loses its self: prefrontal inactivation during sensorimotor processing. *Neuron*, 50(2):329–339.

Goldman-Rakic, P. S. (1988). Topography of cognition: parallel distributed networks in primate association cortex. *Annual review of neuroscience*.

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., and Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *Journal of Experimental Psychology: General*, 140(2):186.

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

Gordon, P., Hendrick, R., and Johnson, M. (2001). Memory Interference during Language Processing. *Journal of experimental psychology. Learning, memory, and cognition*, 27:1411–1423.

Gorrostieta, C., Ombao, H., Bédard, P., and Sanes, J. N. (2012). Investigating brain connectivity using mixed effects vector autoregressive models. *NeuroImage*, 59(4):3347–3355.

Goshtasby, A. and O'Neill, W. D. (1994). Curve fitting by a sum of Gaussians. *CVGIP: Graphical Models and Image Processing*, 56(4):281–288.

Graff, D. and Cieri, C. (2003). English Gigaword LDC2003T05.

Graff, D., Kong, J., Chen, K., and Maeda, K. (2007). English Gigaword Third Edition LDC2007T07.

Griliches, Z. (1967). Distributed lags: A survey. *Econometrica: journal of the Econometric Society*, pages 16–49.

Grodner, D. J. and Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, 29:261–291.

Gulordava, K., Bojanowski, P., Grave, ., Linzen, T., and Baroni, M. (2018). Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

H. Glover, G. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9:416–429.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in cognitive sciences*, 9(9):416–423.

Hale, J. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the second meeting of the North American chapter of the Association for Computational Linguistics*, pages 159–166, Pittsburgh, PA.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):609–642.

Hale, J. (2014). Automaton theories of human sentence comprehension. CSLI lecture notes, chapter 8. CSLI Publications/Center for the Study of Language & Information.

Hale, J., Lutz, D., Luh, W.-M., and Brennan, J. (2015). Modeling fMRI time courses with linguistic structure at various grain sizes. In *Proceedings of the 6th workshop on cognitive modeling and computational linguistics*, pages 89–97.

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, 18(10):1279–1296.

Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21(4):1639–1651.

Harel, N., Lee, S.-P., Nagaoka, T., Kim, D.-S., and Kim, S.-G. (2002). Origin of negative blood oxygenation leveldependent fMRI signals. *Journal of cerebral blood flow {\textbackslash}& metabolism*, 22(8):908–917.

Harm, M. W. and Seidenberg, M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological review*, 111(3):662.

Harrington Stack, C. M., James, A. N., and Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & cognition*, 46(6):864–877.

Harrison, L., Penny, W. D., and Friston, K. (2003). Multivariate autoregressive modeling of fMRI time series. *Neuroimage*, 19(4):1477–1491.

Hartwigsen, G., Henseler, I., Stockert, A., Wawrzyniak, M., Wendt, C., Klingbeil, J., Baumgaertner, A., and Saur, D. (2017). Integration demands modulate effective connectivity in a fronto-temporal network for contextual sentence integration. *NeuroImage*, 147:812–824.

Hasson, U., Egidi, G., Marelli, M., and Willems, R. M. (2018). Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension. *Cognition*, 180:135–157.

Hasson, U. and Honey, C. J. (2012). Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, 62(2):1272–1278.

Hasson, U., Malach, R., and Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1):40–48.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statist. Sci.*, 1(3):297–310.

Hauk, O., Davis, M. H., Ford, M., Pulvermüller, F., and Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *Neuroimage*, 30(4):1383–1400.

Havron, N., de Carvalho, A., Fiévet, A.-C., and Christophe, A. (2019). Three- to Four-Year-Old Children Rapidly Adapt Their Predictions and Use Them to Learn Novel Word Meanings. *Child Development*, 90(1):82–90.

Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443.

Hawkins, J. A. (1994). *A performance theory of order and constituency*. Cambridge University Press, Cambridge, U.K.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696, Sofia, Bulgaria.

Hein, G. and Knight, R. T. (2008). Superior temporal sulcusit's my area: or is it? *Journal of Cognitive Neuroscience*, 20(12):2125–2136.

Henderson, J. M., Choi, W., Lowder, M. W., and Ferreira, F. (2016). Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*, 132:293–300.

Henderson, J. M., Choi, W., Luke, S. G., and Desai, R. H. (2015). Neural correlates of fixation duration in natural reading: evidence from fixation-related fMRI. *NeuroImage*, 119:390–397.

Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Hsiao, F. and Gibson, E. (2003). Processing relative clauses in {Chinese}. *Cognition*, 90(1):3–27.

Hsu, N. S. and Novick, J. M. (2016). Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. *Psychological science*, 27(4):572–582.

Hu, X. and Yacoub, E. (2012). The story of the initial dip in fMRI. *Neuroimage*, 62(2):1103–1108.

Huettig, F. and Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, 31(1):19–31.

Hugdahl, K., Raichle, M. E., Mitra, A., and Specht, K. (2015). On the existence of a generalized non-specific task-dependent network. *Frontiers in Human Neuroscience*, 9:430.

Humphries, C., Binder, J. R., Medler, D. A., and Liebenthal, E. (2006). Syntactic and semantic modulation of neural activity during auditory sentence comprehension. *Journal of Cognitive Neuroscience*, 18(4):665–679.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453.

Inhoff, A. W. and Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception {\textbackslash}& psychophysics*, 40(6):431–439.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456.

Jacquemot, C. and Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, 10(11):480–486.

Jaffe, E., Shain, C., and Schuler, W. (2018). Coreference and Focus in Reading Times. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 1–9.

January, D., Trueswell, J. C., and Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, 21(12):2434–2444.

Jeneson, A., Mauldin, K. N., and Squire, L. R. (2010). Intact working memory for relational information after medial temporal lobe damage. *Journal of Neuroscience*, 30(41):13624–13629.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.

Jones, E. G. and Powell, T. P. S. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain*, 93(4):793–820.

Jones, M. C. and Pewsey, A. (2009). Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780.

Josephs, O., Turner, R., and Friston, K. (1997). Event-related fMRI. *Human brain mapping*, 5(4):243–248.

Joshi, A. K. (1985). How much context sensitivity is necessary for characterizing structural descriptions: Tree adjoining grammars. In D. Dowty, L. K. and Zwicky, A., editors, *Natural language parsing: Psychological, computational and theoretical perspectives*, pages 206–250. Cambridge University Press, Cambridge, U.K.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Juch, H., Zimine, I., Seghier, M. L., Lazeyras, F., and Fasel, J. H. D. (2005). Anatomical variability of the lateral frontal lobe surface: implication for intersubject variability in language neuroimaging. *Neuroimage*, 24(2):504–514.

Just, M. A. and Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, 1(6):571–591.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, 4(2):257–282.

Kaan, E., Harris, A., Gibson, E., and Holcomb, P. (2000). The P600 as an index of syntactic integration difficulty. *Language and cognitive processes*, 15(2):159–201.

Kaan, E. and Swaab, T. Y. (2002). The brain circuitry of syntactic comprehension. *Trends in cognitive sciences*, 6(8):350–356.

Kannurpatti, S. S. and Biswal, B. B. (2004). Negative functional response to sensory stimulation and its origins. *Journal of Cerebral Blood Flow {\textbackslash}& Metabolism*, 24(6):703–712.

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, 27(2):117–143.

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

Kennedy, A., Pynte, J., and Hill, R. (2003). The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Kiehl, K. A., Laurens, K. R., and Liddle, P. F. (2002). Reading anomalous sentences: An event-related fMRI study of semantic processing. *Neuroimage*, 17(2):842–850.

Kimberg, D. Y. and Farah, M. J. (1993). A unified account of cognitive impairments following frontal lobe damage: The role of working memory in complex, organized behavior. *Journal of Experimental Psychology: General*, 122(4):411.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6.

Kliegl, R., Nuthmann, A., and Engbert, R. (2006). Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of experimental psychology: General*, 135(1):12.

Kluender, R. and Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, 5(2):196–214.

Koechlin, E. and Jubault, T. (2006). Broca's Area and the Hierarchical Organization of Human Behavior. *Neuron*, 50(6):963–974.

Koelsch, S., Gunter, T. C., Cramon, D. Y. v., Zysset, S., Lohmann, G., and Friederici, A. D. (2002). Bach speaks: a cortical "language-network" serves the processing of music. *Neuroimage*, 17(2):956–966.

Kolers, P. A. (1976). Buswells discoveries. *Eye movements and psychological processes*, pages 371–395.

Konieczny, S. (2000). On the Difference between Merging Knowledge Bases and Combining them. *KR*, pages 135–144.

Koyck, L. M. (1954). *Distributed lags and investment analysis*, volume 4. North-Holland Publishing Company.

Kruggel, F. and von Cramon, D. Y. (1999). Temporal properties of the hemodynamic response in functional MRI. *Human brain mapping*, 8(4):259–271.

Kruggel, F., Wiggins, C. J., Herrmann, C. S., and von Cramon, D. Y. (2000). Recording of the event-related potentials during functional MRI at 3.0 Tesla field strength. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 44(2):277–282.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Kuperberg, G. R., Holcomb, P. J., Sitnikova, T., Greve, D., Dale, A. M., and Caplan, D. (2003). Distinct patterns of neural modulation during the processing of conceptual and syntactic anomalies. *Journal of Cognitive Neuroscience*, 15(2):272–293.

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

Kuperberg, G. R., McGuire, P. K., Bullmore, E. T., andS. Rabehesketh, M. J. B., Wright, I. C., Lythgoe, D. J., and David, S. C. R. W. A. S. (2000). Common and Distinct Neural Substrates for Pragmatic, Semantic, and Syntactic Processing of Spoken Sentences: An fMRI Study. *Journal of Cognitive Neuroscience*, 12(2):321–341.

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

Kutas, M. and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62:621–647.

Kutas, M. and Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205.

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

Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2003). *Applied Linear Regression Models*. McGraw-Hill Higher Education, Boston.

Lambon Ralph, M. A., Jefferies, E., Patterson, K., and Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42.

Lapham, B. M. (2014). *Hawkes processes and some financial applications*. PhD thesis, University of Cape Town.

Lau, E. F., Weber, K., Gramfort, A., Hämäläinen, M. S., and Kuperberg, G. R. (2016). Spatiotemporal signatures of lexical–semantic prediction. *Cerebral Cortex*, 26(4):1377–1387.

Lawrence, D. H. (1971). Two studies of visual search for word targets with controlled rates of presentation. *Perception {\textbackslash}& Psychophysics*, 10(2):85–89.

Lee, A. T., Glover, G. H., and Meyer, C. H. (1995). Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging. *Magnetic Resonance in Medicine*, 33(6):745–754.

Lesage, E., Hansen, P. C., and Miall, R. C. (2017). Right lateral cerebellum represents linguistic predictability. *Journal of Neuroscience*, 37(26):6231–6241.

Leszczynski, M. (2011). How does hippocampus contribute to working memory processing? *Frontiers in Human Neuroscience*, 5:168.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press.

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

Levy, R., Fedorenko, E., and Gibson, E. (2013). The syntactic complexity of {Russian} relative clauses. *Journal of Memory and Language*, 69:461–495.

Levy, R. and Gibson, E. (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology*, 4(229).

Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Lewis, S. and Phillips, C. (2015). Aligning grammatical theories and language processing models. *Journal of Psycholinguistic Research*, 44(1):27–46.

Linck, J. A., Osthus, P., Koeth, J. T., and Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic bulletin {\textbackslash}& review*, 21(4):861–883.

Lindquist, M. and Wager, T. (2007). Validity and Power in Hemodynamic Response Modeling: A Comparison Study and a New Approach. *Human brain mapping*, 28:764–784.

Lindquist, M. A., Loh, J. M., Atlas, L. Y., and Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage*, 45(1, Supplement 1):S187 – S198.

Linzen, T., Dupoux, E., and Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Ljung, L. and Glad, T. (1994). *Modeling of dynamic systems*. Prentice-Hall.

Logothetis, N. K. (2003). The underpinnings of the BOLD functional magnetic resonance imaging signal. *Journal of Neuroscience*, 23(10):3963–3971.

Lopopolo, A., Frank, S. L., den Bosch, A., and Willems, R. M. (2017). Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PloS one*, 12(5):e0177794.

Lopopolo, A., van den Bosch, A., Petersson, K.-M., and Willems, R. M. (2020). Distinguishing syntactic operations in the brain: Dependency and phrase-structure parsing. *Neurobiology of Language*, (Just Accepted):1–64.

Lowe, J. J. (2019). The Syntax and Semantics of Nonfinite Forms. *Annual Review of Linguistics*, 5:309–328.

MacDonald, M. C., Just, M. A., and Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive psychology*, 24(1):56–98.

MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Madisetti, V. (1997). *The digital signal processing handbook*. CRC press.

Maess, B., Koelsch, S., Gunter, T. C., and Friederici, A. D. (2001). Musical syntax is processed in Broca's area: an MEG study. *Nature Neuroscience*, 4:540–545.

Mahowald, K. and Fedorenko, E. (2016). Reliable individual-level neural markers of high-level language processing: A necessary precursor for relating neural variability to behavioral and genetic variability. *Neuroimage*, 139:74–93.

Makel, M. C. and Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher*, 43(6):304–316.

Makuuchi, M., Bahlmann, J., Anwander, A., and Friederici, A. D. (2009). Segregating the core computational faculty of human language from working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 106:8362–8367.

Mani, N. and Huettig, F. (2012). Prediction during language processing is a piece of cakeBut only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance*, 38(4):843.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge.

Marslen-Wilson, W. D. (1975). Sentence Perception as an Interactive Parallel Process. *Science*, 189(4198):226–228.

Martin, A., Peperkamp, S., and Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, 37(1):103–124.

Matchin, W., Brodbeck, C., Hammerly, C., and Lau, E. (2018). The temporal dynamics of structure and content in sentence comprehension: Evidence from fMRI-constrained MEG. *Human brain mapping*, 40(2):663–678.

Matchin, W., Hammerly, C., and Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *Cortex*, 88:106–123.

Matchin, W. and Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, 30(3):1481–1498.

Matchin, W., Sprouse, J., and Hickok, G. (2014). A structural distance effect for backward anaphora in Brocas area: An fMRI study. *Brain and language*, 138:1–11.

Mather, M., Cacioppo, J. T., and Kanwisher, N. (2013). How fMRI can inform cognitive theories. *Perspectives on Psychological Science*, 8(1):108–113.

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., and Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience*, 5(4):467–479.

McDonough, K. and Trofimovich, P. (2012). How to use Psycholinguistic Methodologies for Comprehension and Production. In *Research Methods in Second Language Acquisition*, chapter 7, pages 117–138. John Wiley & Sons, Ltd.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of psycholinguistic research*, 29(2):111–123.

McElree, B., Foraker, S., and Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48:67–91.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

McKoon, G. and Ratcliff, R. (1979). Priming in episodic and semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 18(4):463–480.

McMillan, C. T., Clark, R., Gunawardena, D., Ryant, N., and Grossman, M. (2012). fMRI evidence for strategic decision-making during resolution of pronoun reference. *Neuropsychologia*, 50(5):674–687.

McMillan, C. T., Coleman, D., Clark, R., Liang, T.-W., Gross, R. G., and Grossman, M. (2013). Converging evidence for the processing costs associated with ambiguous quantifier comprehension. *Frontiers in psychology*, 4:153.

Mei, H. and Eisner, J. (2017). The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS17, page 67576767, Red Hook, NY, USA. Curran Associates Inc.

Mesulam, M.-M. (1998). From sensation to cognition. *Brain: A Journal of Neurology*, 121(6):1013–1052.

Meyer, L., Obleser, J., and Friederici, A. D. (2013). Left parietal alpha enhancement during working memory-intensive sentence processing. *Cortex*, 49(3):711–721.

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., and Buckner, R. L. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *Neuroimage*, 11(6):735–759.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202.

Miller, G. and Chomsky, N. (1963). Finitary models of language users. In Luce, R., Bush, R., and Galanter, E., editors, *Handbook of Mathematical Psychology*, volume 2, pages 419–491. John Wiley.

Mineroff, Z., Blank, I. A., Mahowald, K., and Fedorenko, E. (2018). A robust dissociation among the language, multiple demand, and default mode networks: evidence from inter-region correlations in effect size. *Neuropsychologia*, 119:501–511.

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading. *New methods in reading comprehension research*, pages 69–89.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science*, 320:1191–1195.

Mitsugi, S. and MacWhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism: Language and Cognition*, 19(1):19–35.

Mollica, F. and Piantadosi, S. (2017). An incremental information-theoretic buffer supports sentence processing. In *Proceedings of the 39th Annual Cognitive Science Society Meeting*.

Montague, P. R., Dayan, P., and Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of neuroscience*, 16(5):1936–1947.

Monti, M. M., Parsons, L. M., and Osherson, D. N. (2012). Thought beyond language: Neural dissociation of algebra and natural language. *Psychological science*, 23(8):914–922.

Morton, J. (1964). The effects of context upon speed of reading, eye movements and eye-voice span. *Quarterly Journal of Experimental Psychology*, 16(4):340–354.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R. J., Luuk, A., and others (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature*, 385(6615):432–434.

Näätänen, R., Paavilainen, P., Rinne, T., and Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical neurophysiology*, 118(12):2544–2590.

Nadel, L. and MacDonald, L. (1980). Hippocampus: Cognitive map or working memory? *Behavioral and neural biology*, 29(3):405–409.

Nakatani, K. and Gibson, E. (2010). An on-line study of Japanese nesting complexity. *Cognitive Science*, 34(1):94–112.

Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., and others (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate ${O}(1/kˆ2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

Neter, J., Wasserman, W., and Kutner, M. H. (1989). Applied linear regression models.

Neuvo, Y., Cheng-Yu, D., and Mitra, S. (1984). Interpolated finite impulse response filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(3):563–570.

Newman, A. J., Pancheva, R., and Ozawa, K. (2001). An Event-Related {fMRI} Study of Syntactic and Semantic Violations. *Journal of Psycholinguistic Research*, 30(3):339–364.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Nguyen, L., van Schijndel, M., and Schuler, W. (2012). Accurate Unbounded Dependency Recovery using Generalized Categorial Grammars. In *Proceedings of COLING 2012*.

Nieto-Castañón, A. and Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, 63(3):1646–1669.

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience*, 14(9):1105.

Nieuwland, M. S., Martin, A. E., and Carreiras, M. (2012). Brain regions that process case: evidence from Basque. *Human brain mapping*, 33(11):2509–2520.

Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological review*, 113(2):327.

Novais-Santos, S., Gee, J., Shah, M., Troiani, V., Work, M., and Grossman, M. (2007). Resolving sentence ambiguity with planning and working memory resources: Evidence from fMRI. *Neuroimage*, 37(1):361–378.

Novick, J. M., Trueswell, J. C., and Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Brocas area in sentence comprehension. *Cognitive, Affective, {\textbackslash}& Behavioral Neuroscience*, 5(3):263–281.

Obleser, J. and Kotz, S. A. (2009). Expectancy constraints in degraded speech modulate the language comprehension network. *Cerebral Cortex*, 20(3):633–640.

Obleser, J., Wise, R. J. S., Dresner, M. A., and Scott, S. K. (2007). Functional integration across brain regions improves speech perception under adverse listening conditions. *Journal of Neuroscience*, 27(9):2283–2289.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

Olson, I. R., Page, K., Moore, K. S., Chatterjee, A., and Verfaellie, M. (2006). Working memory for conjunctions relies on the medial temporal lobe. *Journal of Neuroscience*, 26(17):4596–4601.

Olton, D. S., Becker, J. T., and Handelmann, G. E. (1979). Hippocampus, space, and memory. *Behavioral and Brain sciences*, 2(3):313–322.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Osterhout, L. and Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of memory and language*, 31(6):785–806.

Owen, A. M., Downes, J. J., Sahakian, B. J., Polkey, C. E., Robbins, T. W., and others (1990). Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia*, 28(10):1021–1034.

Ozaki, T. (1979). Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(1):145–155.

Pallier, C., Devauchelle, A.-D., and Dehaene, S. (2011). Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527.

Pashler, H. and Wagenmakers, E.-J. (2012). Editors introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7(6):528–530.

Payne, B. R. and Federmeier, K. D. (2018). Contextual constraints on lexico-semantic processing in aging: Evidence from single-word event-related brain potentials. *Brain Research*, 1687:117–128.

Pearlmutter, N. J., Garnsey, S. M., and Bock, K. (1999). Agreement Processes in Sentence Comprehension. *Journal of Memory and Language*, 41(3):427–456.

Pedregosa, F., Eickenberg, M., Ciuciu, P., Gramfort, A., and Thirion, B. (2014). Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, 104.

Peelle, J. E., Troiani, V., Wingfield, A., and Grossman, M. (2009). Neural processing during older adults' comprehension of spoken sentences: Age differences in resource allocation and connectivity. *Cerebral Cortex*, 20(4):773–782.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., and Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):963.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Petersen, S. E., Fox, P. T., Posner, M. I., Mintun, M., and Raichle, M. E. (1988). Positron emission tomographic studies of the cortical anatomy of single-word processing. *Nature*, 331(6157):585.

Petsiuk, V., Das, A., and Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Pickering, M. J. and Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological bulletin*, 144(10):1002.

Pinker, S. (1994). *The Language Instinct: How the Mind Creates Language*. HarperCollins, New York.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in cognitive sciences*, 10(2):59–63.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692–697.

Pollatsek, A., Juhasz, B., Reichle, E., Machacek, D., and Rayner, K. (2008). Immediate and Delayed Effects of Word Frequency and Word Length on Eye Movements in Reading: A Reversed Delayed Effect of Word Length. *Journal of experimental psychology. Human perception and performance*, 34:726–750.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Posner, M., Walker, J., Friedrich, F., and Rafal, R. (1987). How do parietal lobes direct covert attention? *Neuropsychologia*, 25:135–145.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1):3–25.

Potter, M. C. (1984). Rapid serial visual presentation (RSVP): A method for studying language processing. In Kieras, D. E. and Just, M. A., editors, *New methods in reading comprehension research*, pages 91–118. Lawrence Erlbaum Associates, Inc.

Prabhakaran, V., Narayanan, K., Zhao, Z., and Gabrieli, J. D. E. (2000). Integration of diverse information in working memory within the frontal lobe. *Nature neuroscience*, 3(1):85–90.

Prasad, G. and Linzen, T. (2019). Rapid syntactic adaptation in self-paced reading: detectable, but requires many participants.

Pratt, H. (2011). Sensory ERP components. *The Oxford handbook of event-related potential components*, pages 89–114.

Pynte, J., New, B., and Kennedy, A. (2008). On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision research*, 48(21):2172–2183.

Radach, R. and Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *Quarterly Journal of Experimental Psychology*, 66(3):429–452.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

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

Raskutti, G., Wainwright, M. J., and Yu, B. (2014). Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366.

Rasmussen, N. E. and Schuler, W. (2018). Left-Corner Parsing With Distributed Associative Memory Produces Surprisal and Locality Effects. *Cognitive Science*, 42:1009–1042.

Rayner, K. (1977). Visual attention in reading: Eye movements reflect cognitive processes. *Memory \& Cognition*, 5(4):443–448.

Rayner, K. (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422.

Rayner, K., Ashby, J., Pollatsek, A., and Reichle, E. D. (2004a). The effects of frequency and predictability on eye fixations in reading: Implications for the EZ Reader model. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):720.

Rayner, K., Carlson, M., and Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, 22(3):358–374.

Rayner, K. and Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, 14(3):191–201.

Rayner, K., Sereno, S. C., Morris, R. K., Schmauder, A. R., and Clifton Jr, C. (1989). Eye movements and on-line language comprehension processes. *Language and Cognitive Processes*, 4(3-4):SI21–SI49.

Rayner, K., Slattery, T. J., Drieghe, D., and Liversedge, S. P. (2011). Eye movements and word skipping during reading: effects of word length and predictability. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):514.

Rayner, K., Warren, T., Juhasz, B. J., and Liversedge, S. P. (2004b). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(6):1290.

Reichle, E. and Drieghe, D. (2014). Using E-Z Reader to Examine the Consequences of Fixation-Location Measurement Error. *Journal of experimental psychology. Learning, memory, and cognition*, 41.

Reichle, E., Liversedge, S., Pollatsek, A., and Rayner, K. (2009). Encoding multiple words simultaneously in reading is implausible. *Trends in cognitive sciences*, 13:115–119.

Reichle, E., Pollatsek, A., Fisher, D., and Rayner, K. (1998). Toward a Model of Eye Movement Control in Reading. *Psychological review*, 105:125–157.

Reichle, E. D., Rayner, K., and Pollatsek, A. (2003). The EZ Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and brain sciences*, 26(4):445–476.

Remington, R. W., Burt, J. S., and Becker, S. I. (2018). The curious case of spillover: Does it tell us much about saccade timing in reading? *Attention, Perception, {\textbackslash}& Psychophysics*, 80(7):1683–1690.

Resnik, P. (1992). Left-Corner Parsing and Psychological Plausibility. In *Proceedings of {COLING}*, pages 191–197, Nantes, France.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Richlan, F., Gagl, B., Hawelka, S., Braun, M., Schurz, M., Kronbichler, M., and Hutzler, F. (2013). Fixation-related fMRI analysis in the domain of reading research: using self-paced eye movements as markers for hemodynamic brain responses during visual letter string processing. *Cerebral Cortex*, 24(10):2647–2656.

Roark, B., Bachrach, A., Cardenas, C., and Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down

parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Langauge Processing*, pages 324–333.

Robinson, P. (1976). Fourier estimation of continuous time models. *Statistical Inference in Continuous Time Economic Models.*, pages 215–266.

Robinson, P. M. (1975). Continuous time regressions with discrete data. *The Annals of Statistics*, pages 688–697.

Rodd, J. M., Davis, M. H., and Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, 15(8):1261–1269.

Rogalsky, C., Almeida, D., Sprouse, J., and Hickok, G. (2015). Sentence processing selectivity in Broca's area: Evident for structure but not syntactic movement. *Language, cognition and neuroscience*, 30(10):1326–1338.

Rogalsky, C. and Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, 19(4):786–796.

Rogalsky, C. and Hickok, G. (2011). The role of Broca's area in sentence comprehension. *Journal of Cognitive Neuroscience*, 23(7):1664–1680.

Röther, J., Knab, R., Hamzei, F., Fiehler, J., Reichenbach, J. R., Büchel, C., and Weiller, C. (2002). Negative dip in BOLD fMRI is caused by blood flowoxygen consumption uncoupling in humans. *Neuroimage*, 15(1):98–102.

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., Fox, P. T., and Eickhoff, S. B. (2012). Modelling neural correlates of working memory: a coordinate-based meta-analysis. *Neuroimage*, 60(1):830–846.

Ryskin, R., Levy, R. P., and Fedorenko, E. (2020). Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia*, 136:107258.

Saad, Z. S., Ropella, K. M., Cox, R. W., and DeYoe, E. A. (2001). Analysis and use of FMRI response delays. *Human brain mapping*, 13(2):74–93.

Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909.

Saramaeki, T., Mitra, S. K., and Kaiser, J. F. (1993). Finite impulse response filter design. *Handbook for digital signal processing*, 4:155–277.

Saxe, R., Brett, M., and Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *Neuroimage*, 30(4):1088–1096.

Scheperjans, F., Eickhoff, S. B., Hömke, L., Mohlberg, H., Hermann, K., Amunts, K., and Zilles, K. (2008). Probabilistic maps, morphometry, and variability of cytoarchitectonic areas in the human superior parietal cortex. *Cerebral cortex*, 18(9):2141–2157.

Schotter, E. R., Leinenger, M., and von der Malsburg, T. (2018). When your mind skips what your eyes fixate: How forced fixations lead to comprehension illusions in reading. *Psychonomic Bulletin & Review*, 25(5):1884–1890.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N. G., Tenenbaum, J. B., and Fedorenko, E. (2020). Artificial Neural Networks Accurately Predict Language Processing in the Brain. *BioRxiv*.

Schuster, S., Hawelka, S., Himmelstoss, N. A., Richlan, F., and Hutzler, F. (2019). The neural correlates of word position and lexical predictability during sentence reading: Evidence from fixation-related fMRI. *Language, Cognition and Neuroscience*, pages 1–12.

Schuster, S., Hawelka, S., Hutzler, F., Kronbichler, M., and Richlan, F. (2016). Words in context: The effects of length, frequency, and predictability on brain responses during natural reading. *Cerebral Cortex*, 26(10):3889–3904.

Scott, T. L., Gallée, J., and Fedorenko, E. (2017). A new fun and robust version of an fMRI localizer for the frontotemporal language system. *Cognitive neuroscience*, 8(3):167–176.

Seidenberg, M. S. and McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4086–4094.

Shain, C. and Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Shain, C., van Schijndel, M., Futrell, R., Gibson, E., and Schuler, W. (2016). Memory access during incremental sentence processing causes reading time latency. In *COLING 2016, workshop on Computational Linguistics for Linguistic Complexity*.

Shain, C., van Schijndel, M., and Schuler, W. (2018). Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources (LREC 2018)*.

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

Sharkey, N. E. and Sharkey, A. J. C. (1987). What is the point of integration? The loci of knowledge-based facilitation in sentence processing. *Journal of Memory and Language*, 26(3):255–276.

Shmuel, A., Augath, M., Oeltermann, A., and Logothetis, N. K. (2006). Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. *Nature neuroscience*, 9(4):569.

Shmuel, A., Yacoub, E., Pfeuffer, J., de Moortele, P.-F., Adriany, G., Hu, X., and Ugurbil, K. (2002). Sustained negative BOLD, blood flow and oxygen consumption response and its coupling to the positive response in the human brain. *Neuron*, 36(6):1195–1210.

Shrager, Y., Levy, D. A., Hopkins, R. O., and Squire, L. R. (2008). Working memory and the organization of brain systems. *Journal of Neuroscience*, 28(18):4818–4822.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9(1):76–80.

Sims, C. A. (1971). Discrete approximations to continuous time distributed lags in econometrics. *Econometrica: Journal of the Econometric Society*, pages 545–563.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

Singer, Y., Teramoto, Y., Willmore, B. D. B., Schnupp, J. W. H., King, A. J., and Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *eLife*, 7:e31557.

Smith, A. T., Singh, K. D., and Greenlee, M. W. (2000). Attentional suppression of activity in the human visual cortex. *Neuroreport*, 11(2):271–278.

Smith, N. J. and Kutas, M. (2015). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology*, 52(2):157–168.

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

Smith, N. J. and Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the 33rd CogSci Conference*.

Smith, N. J. and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Sood, M. R. and Sereno, M. I. (2016). Areas activated during naturalistic reading comprehension overlap topological visual, auditory, and somatotomotor maps. *Human brain mapping*, 37(8):2784–2810.

Speer, N. K., Reynolds, J. R., Swallow, K. M., and Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological science*, 20(8):989–999.

Speer, N. K., Zacks, J. M., and Reynolds, J. R. (2007). Human brain activity time-locked to narrative event boundaries. *Psychological Science*, 18(5):449–455.

Sreenivasan, K. K., Curtis, C. E., and DEsposito, M. (2014). Revisiting the role of persistent neural activity during working memory. *Trends in cognitive sciences*, 18(2):82–89.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Stanescu-Cosson, R., Pinel, P., van de Moortele, P.-F., Le Bihan, D., Cohen, L., and Dehaene, S. (2000). Understanding dissociations in dyscalculia: A brain imaging study of the impact of number size on the cerebral networks for exact and approximate calculation. *Brain*, 123(11):2240–2255.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86.

Staub, A. (2011). Word recognition and syntactic attachment in reading: Evidence for a staged architecture. *Journal of experimental psychology. General*, 140:407–433.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Staub, A. and Benatar, A. (2013). Individual differences in fixation duration distributions in reading. *Psychonomic Bulletin & Review*, 20(6):1304–1311.

Staub, A. and Clifton, C. (2006). Syntactic Prediction in Language Comprehension: Evidence From Either...or. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(2):425–436.

Steedman, M. (2000). *The syntactic process*. MIT Press/Bradford Books, Cambridge, MA.

Stowe, L. A., Broere, C. A. J., Paans, A. M. J., Wijers, A. A., Mulder, G., Vaalburg, W., and Zwarts, F. (1998). Localizing components of a complex task: Sentence processing and working memory. *Neuroreport*, 9(13):2995–2999.

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Networks*, 18(3):225–230.

Strijkers, K., Chanoine, V., Munding, D., Dubarry, A.-S., Trébuchon, A., Badier, J.-M., and Alario, F.-X. (2019). Grammatical class modulates the (left) inferior frontal gyrus within 100 milliseconds when syntactic context is predictive. *Scientific reports*, 9(1):4830.

Swets, B., Desmet, T., Clifton, C., and Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory and Cognition*, 36(1):201–216.

Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of Merely Local Syntactic Coherence on Sentence Processing. *Journal of Memory and Language*, 50(4):355–370.

Tahmasebi, A. M., Artiges, E., Banaschewski, T., Barker, G. J., Bruehl, R., Büchel, C., Conrod, P. J., Flor, H., Garavan, H., Gallinat, J., and others (2012). Creating probabilistic maps of the face network in the adolescent brain: a multicentre functional MRI study. *Human brain mapping*, 33(4):938–957.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Tauroza, S. and Allison, D. (1990). Speech rates in British english. *Applied linguistics*, 11(1):90–105.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. *ACL19*.

Theeuwes, J., Kramer, A. F., Hahn, S., and Irwin, D. E. (1998). Our Eyes Do Not Always Go Where We Want Them to Go: Capture of the Eyes by New Objects. *Psychological Science*, 9(5):379–385.

Thompson-Schill, S. L., Bedny, M., and Goldberg, R. F. (2005). The frontal lobes and the regulation of mental activity. *Current opinion in neurobiology*, 15(2):219–224.

Tomaiuolo, F., MacDonald, J. D., Caramanos, Z., Posner, G., Chiavaras, M., Evans, A. C., and Petrides, M. (1999). Morphology, morphometry and probability mapping of the pars opercularis of the inferior frontal gyrus: an in vivo MRI analysis. *European Journal of Neuroscience*, 11(9):3033–3046.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

Travis, R. (1936). The latency and velocity of the eye in saccadic movements. *Psychological Monographs*, 47:242–249.

Trueswell, J. C., Tanenhaus, M. K., and Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*, 33(3):285–318.

Tyler, L. K., Marslen-Wilson, W. D., Randall, B., Wright, P., Devereux, B. J., Zhuang, J., Papoutsi, M., and Stamatakis, E. A. (2011). Left inferior frontal cortex and syntax: Function, structure and behaviour in patients with left hemisphere damage. *Brain*, 134(2):415–431.

Ullman, M. T. (2016). The declarative/procedural model: a neurobiological model of language learning, knowledge, and use. In *Neurobiology of language*, pages 953–968. Elsevier.

Vagharchakian, L., Dehaene-Lambertz, G., Pallier, C., and Dehaene, S. (2012). A temporal bottleneck in the language comprehension network. *Journal of Neuroscience*, 32(26):9089–9102.

Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):407.

Van Dyke, J. A. and 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*, 49:285–316.

van Schijndel, M., Exley, A., and Schuler, W. (2013). A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

van Schijndel, M. and Linzen, T. (2018). A Neural Model of Adaptation in Reading. In *EMNLP 2018*, pages 4704–4710.

van Schijndel, M. and Schuler, W. (2013). An Analysis of Frequency- and Memory-Based Processing Costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

van Schijndel, M. and Schuler, W. (2015). Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics.

Vandenberghe, R., Nobre, A. C., and Price, C. J. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience*, 14(4):550–560.

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center embeddings*. Routledge.

Vasishth, S. (2006). On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In *Proceedings of the International Conference on Linguistic Evidence*, pages 96–100.

Vasishth, S. and Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.

Vasishth, S., Mertzen, D., Jäger, L. A., and Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 103:151–175.

Vazquez, A. L., Cohen, E. R., Gulani, V., Hernandez-Garcia, L., Zheng, Y., Lee, G. R., Kim, S.-G., Grotberg, J. B., and Noll, D. C. (2006). Vascular dynamics and BOLD fMRI: CBF level effects and analysis considerations. *Neuroimage*, 32(4):1642–1655.

Visser, M., Jefferies, E., and Lambon Ralph, M. A. (2010). Semantic processing in the anterior temporal lobes: A meta-analysis of the functional neuroimaging literature. *Journal of cognitive neuroscience*, 22(6):1083–1094.

Wacongne, C., Changeux, J.-P., and Dehaene, S. (2012). A neuronal model of predictive coding accounting for the mismatch negativity. *Journal of Neuroscience*, 32(11):3665–3678.

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

Wager, T. D., Vazquez, A., Hernandez, L., and Noll, D. C. (2005). Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies. *NeuroImage*, 25(1):206–218.

Wagers, M. W., Lau, E. F., and Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, 61(2):206–237.

Ward, B. D. (2006). Deconvolution Analysis of {fMRI} Time Series Data.

Warren, T. and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85:79–112.

Wehbe, L., Blank, I. A., Shain, C., Futrell, R., Levy, R., von der Malsburg, T., Smith, N., Gibson, E., and Fedorenko, E. (2020). Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *bioRxiv*.

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., and Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575.

Wernicke, C. (1874). *Der aphasische Symptomencomplex: eine psychologische Studie auf anatomischer Basis*. Cohn.

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

Wilcox, E., Levy, R., and Futrell, R. (2019). Hierarchical Representation in Neural Language Models: Suppression and Recovery of Expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 181–190.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (2012). Effortful listening: the processing of degraded speech depends critically on attention. *Journal of Neuroscience*, 32(40):14010–14021.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

Willems, R. M., der Haegen, L., Fisher, S. E., and Francks, C. (2014). On the other hand: Including left-handers in cognitive neuroscience and neurogenetics. *Nature Reviews Neuroscience*, 15(3):193.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., and den Bosch, A. (2015). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Wilson, S. M., DeMarco, A. T., Henry, M. L., Gesierich, B., Babiak, M., Mandelli, M. L., Miller, B. L., and Gorno-Tempini, M. L. (2014). What role does the anterior temporal lobe play in sentence-level processing? Neural correlates of syntactic processing in semantic variant primary progressive aphasia. *Journal of Cognitive Neuroscience*, 26(5):970–985.

Wilson, S. M., Galantucci, S., Tartaglia, M. C., and Gorno-Tempini, M. L. (2012). The neural basis of syntactic deficits in primary progressive aphasia. *Brain and language*, 122(3):190–198.

Wise, R. J. S., Scott, S. K., Blank, S. C., Mummery, C. J., Murphy, K., and Warburton, E. A. (2001). Separate neural subsystems within Wernicke's area. *Brain*, 124(1):83–95.

Wlotko, E. W. and Federmeier, K. D. (2012). Age-related changes in the impact of contextual strength on multiple aspects of sentence comprehension. *Psychophysiology*, 49(6):770–785.

Wood, S. N. (2006). *Generalized Additive Models: {An} Introduction with {R}*. Chapman and Hall/CRC, Boca Raton.

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, {\textbackslash}& Psychophysics*, 72(8):2031–2046.

Wu, S., Bachrach, A., Cardenas, C., and Schuler, W. (2010). Complexity Metrics in an Incremental Right-corner Parser. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ({ACL}'10)*, pages 1189–1198.

Yacoub, E., Shmuel, A., Pfeuffer, J., Van De Moortele, P.-F., Adriany, G., Ugurbil, K., and Hu, X. (2001). Investigation of the initial dip in fMRI at 7 Tesla. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 14(7-8):408–412.

Yang, Q., Bucci, M., and Kapoula, Z. (2002). The latency of saccades, vergence, and combined eye movements in children and in adults. *Investigative ophthalmology & visual science*, 43:2939–2949.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Yarkoni, T., Speer, N. K., and Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory. *Neuroimage*, 41(4):1408–1425.

Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

Yonelinas, A. P. (2013). The hippocampus supports high-resolution binding in the service of perception, working memory and long-term memory. *Behavioural brain research*, 254:34–44.

Yoon, T., Okada, J., Jung, M. W., and Kim, J. J. (2008). Prefrontal cortex and hippocampus subserve different components of working memory in rats. *Learning & memory*, 15(3):97–105.

Zaccarella, E., Schell, M., and Friederici, A. D. (2017). Reviewing the functional basis of the syntactic Merge mechanism for language: A coordinate-based activation likelihood estimation meta-analysis. *Neuroscience & Biobehavioral Reviews*, 80:646–656.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional Hawkes processes. *30th International Conference on Machine Learning, ICML 2013*, pages 2338–2346.