CDRNN: Discovering Complex Dynamics in Human Language Processing

Cory Shain
The Ohio State University
shain.3@osu.edu

Abstract

The human mind is a dynamical system, yet many analysis techniques used to study it are limited in their ability to capture the complex dynamics that may characterize mental processes. This study proposes the continuous-time deconvolutional regressive neural network (CDRNN), a deep neural extension of continuous-time deconvolutional regression (CDR, Shain and Schuler, 2021) that jointly captures time-varying, non-linear, and delayed influences of predictors (e.g. word surprisal) on the response (e.g. reading time). Despite this flexibility, CDRNN is interpretable and able to illuminate patterns in human cognition that are otherwise difficult to study. Behavioral and fMRI experiments reveal detailed and plausible estimates of human language processing dynamics that generalize better than CDR and other baselines, supporting a potential role for CDRNN in studying human language processing.

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). Regression techniques have grown in sophistication, from ANOVA (e.g. Pickering and Branigan, 1998) to newer linear mixed-effects approaches (LME, Bates et al., 2015) that enable direct word-by-word analysis of effects in naturalistic human language processing (e.g. Demberg and Keller, 2008; Frank and Bod, 2011). However, these methods struggle to account for delayed effects. Because the human mind operates in real time and experiences computational bottlenecks of various kinds (Bouma and De Voogd, 1974; Just and Carpenter, 1980; Ehrlich and Rayner, 1981; Mollica and Piantadosi, 2017), delayed effects may be pervasive, and, if left uncontrolled, can yield misleading results (Shain and Schuler, 2018).

Continuous-time deconvolutional regression (CDR) is a recently proposed technique to address delayed effects in measures of human cognition (Shain and Schuler, 2018, 2021). CDR fits parametric continuous-time impulse response functions (IRFs) that mediate between word features and response measures. An IRF maps the time elapsed between a stimulus and a response to a weight describing the expected influence of the stimulus on the response. CDR models the response as an IRF-weighted sum of preceding stimuli, thus directly accounting for effect latencies. Empirically, CDR reveals fine-grained processing dynamics and generalizes better to human reading and fMRI responses than established alternatives. 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.

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 study 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 (linearity, additivity, homosketasticity, stationarity, and context-independence, see Section 2), resulting in a highly flexible model. Nevertheless, CDRNN is interpretable and can 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.[1]

2 Background

Psycholinguists have 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 (Shain and Schuler, 2018, 2021) addresses this limitation by fitting continuous-time IRFs, but 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. Shain and Schuler (2021) support 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 Shain and Schuler (2021) 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 method-

[1] Because of page constraints, additional replication details and synthetic results are provided in an external supplement, available here: https://osf.io/z89vn/.

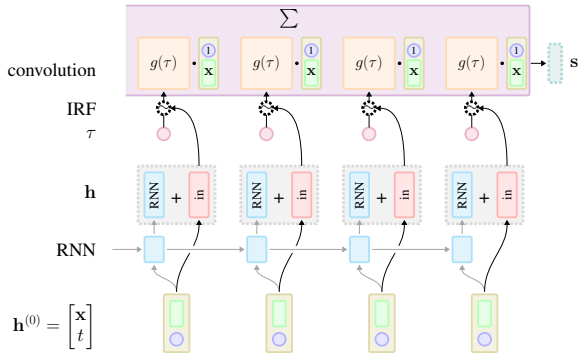

Figure 1: **CDRNN model.** Subscripts omitted to reduce clutter. The IRF $g(\tau)$ at an event computes the expected contribution of each feature of the event vector $\mathbf{h}^{(0)}$ to each element of the parameter vector $\mathbf{s}$ of the predictive distribution for a particular response value. 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 Appendix A.

ological limitations: the GAM models of Smith and Levy (2013) ignore variable event duration, the CDR models of Shain and Schuler (2021) 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.

3 Model

3.1 Architecture

This section presents a high-level description of the model design (for formal definition, see Appendix A). The CDRNN architecture is represented schematically in Figure 1. The primary goal of estimation is to identify the deep neural IRF $g(\tau)$ (top) that computes the influence of a preceding event on the predictive distribution over a subsequent response as a function of their distance in time τ. As shown, the IRF is a feedforward projection of τ into a matrix that defines a weighted sum over the values of input vector $\mathbf{x}$, which is concatenated with a bias to capture general effects of stimulus timing (*rate*). This matrix multiplication determines the contribution of the stimulus event to the parameters of the predictive distribution (e.g. the mean and variance parameters of a Gaussian predictive distribution). Defining the IRF as a function of τ ensures that the model has a continuous-time definition.

To capture non-linear effects of stimulus features, the IRF projection is itself parameterized by a projection of a hidden state $\mathbf{h}$. The dependence on $\mathbf{h}$ permits non-linear influences of the properties of the stimulus sequence on the IRF itself. To generate $\mathbf{h}$, 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 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 contextual influences on the IRF.

The summation over IRF outputs 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. Because the IRF depends on a deep neural projection of the current stimulus as well as (optionally) the entire sequence of preceding stimuli, it implicitly estimates all interactions between these variables in governing the response. Predictors may thus coordinate in a non-linear, non-additive, and time-varying manner.

The CDRNN IRF describes the influence over time of predictors on all parameters of the predictive distribution (in these experiments, the mean and variance parameters of a Gaussian predictive distribution). 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).

Despite their flexibility and task performance (Section 5), CDRNN models used in this study have few parameters (Table A1) by current deep learning standards because they are relatively shallow and small (Supplement S1).

3.2 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, (4) random effects vector $\mathbf{z}$, and (5) response vector $\mathbf{y}$, the model uses gradient de-

scent to minimize the following objective:

$$\mathcal{L}\left(\mathbf{y} \mid \mathcal{C}; \mathbf{w}, \mathbf{z}\right) \stackrel{\text{def}}{=} -\log p\left(\mathbf{y} \mid \mathcal{C}; \mathbf{w}, \mathbf{z}\right) + \lambda_{\text{z}}||\mathbf{z}||_2^2 + \mathcal{L}_{\text{reg}} \quad (1)$$

In addition to random effects shrinkage governed by λ_{z} and any arbitrary additional regularization penalties $\mathcal{L}_{\text{reg}}$ (see Supplement S1), 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}, which 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.

3.3 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.[2] For example, to obtain an estimate of *rate* effects (i.e. the base response or "deconvolutional intercept," see Shain and Schuler, 2021), 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.[3] 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 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).[4] As argued by Shain and Schuler (2021) for CDR, in addition to intervals-based tests, common hypothesis tests (e.g. for the presence of an effect) can be performed in a CDRNN framework via bootstrap model comparison on held out data (e.g. of models with and without the effect of interest).

4 Methods

Following Shain and Schuler (2021), 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

[2] 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.

[3] 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 Shain and Schuler (2021).

[4] 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.

	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	18338	18677	18401	0.0644	0.0651	0.0650	**12760**	**13863**	**13678**	0.1479	**0.1507**	0.1498
CDRNN-RNN	18217	18624	18430	**0.0636**	**0.0647**	**0.0642**	12791	13897	13717	**0.1476**	**0.1507**	**0.1495**

Table 1: **Reading.** Mean squared error by model. Baselines as reported in Shain and Schuler (2021). Daggers (†) indicate convergence failures.

et al., 2003), and the Natural Stories fMRI corpus (∼200K instances, Shain et al., 2020), using the train/dev/test splits for these corpora defined in Shain and Schuler (2021). Further details about datasets and preprocessing are given in Supplement S2.

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; Frank and Bod, 2011; 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 Shain and Schuler (2021) 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 (Shain and Schuler, 2021), 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 established by prior psycholinguistic research (e.g. Rayner, 1998; Demberg and Keller, 2008; van Schijndel and Schuler, 2013; Staub, 2015; Shain and Schuler, 2018, *inter alia*): *unigram* and *5-gram surprisal*, *word length* (reading only), *saccade length* (eye-tracking only), and *previous was fixated* (eye-tracking only). Predictor definitions are given in Appendix C. The deconvolutional intercept term *rate* (Shain and Schuler, 2018, 2021), 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), contextual influences on the IRF, and interactions with 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 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% credible intervals. Code and documentation are available at `https://github.com/coryshain/cdr`.

5 Results

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, as a sanity check, this section first evaluates 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

Model	Train	Expl	Test
Canonical HRF	11.3548†	11.8263†	11.5661†
Linearly interpolated	11.4236†	11.9888†	11.6654†
Averaged	11.3478†	11.9280†	11.6090†
Lanczos interpolated	11.3536†	11.9059†	11.5871†
CDR	11.2774	11.6928	11.5369
CDRNN-FF	**10.5648**	**11.3602**	**11.3042**
CDRNN-RNN	10.8736	11.5631	11.3914

Table 2: **fMRI.** Mean squared error by model. Baselines as reported in Shain and Schuler (2021). Daggers (†) indicate convergence failures.

in Shain and Schuler (2021).[5]

5.1 Model Validation: Baseline Comparisons

Table 1 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[6] 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 5.2, CDRNN largely replicates the CDR estimates on Natural Stories while offering advantages for analysis.

Although CDR struggles against GAM baselines on Dundee, CDRNN has closed the gap. This is noteworthy in light of speculation in Shain and Schuler (2021) 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.

Results from fMRI are shown in Table 2, where both CDRNN variants yield substantial improvements to training, dev, and test set error. These results indicate that the relaxed assumptions afforded by CDRNN are beneficial for describing the fMRI 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 Shain and Schuler (2021), 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, linearly interpolated, averaged, and Lanczos interpolated, and both for CDR).[7] Results are given in Table 3. As shown, both variants of CDRNN significantly improve over all baselines, and CDRNN-RNN significantly improves over CDRNN-FF. Notwithstanding, CDRNN-FF may be preferred in applications: simpler, faster to train, better at recovering synthetic models (Supplement S3), more reliable in noisy domains like fMRI, and close in performance to CDRNN-RNN. Results overall support the reliability of patterns revealed by CDRNN's estimated IRF, which is now used to explore and visualize sentence processing dynamics.

		CDRNN	
		FF	RNN
Baseline	Modality	p	p
LME	Reading	**0.0001*** **	**0.0001*** **
GAM	Reading	**0.0001*** **	**0.0001*** **
Canonical HRF	fMRI	**0.0001*** **	**0.0001*** **
Interpolated	fMRI	**0.0001*** **	**0.0001*** **
Averaged	fMRI	**0.0001*** **	**0.0001*** **
Lanczos	fMRI	**0.0001*** **	**0.0001*** **
CDR	Both	**0.0001*** **	**0.0001*** **
CDRNN-FF	Both	—	**0.0048** **

Table 3: Permutation test of overall test set performance improvement from CDRNN variants over each baseline.

5.2 Effect Latencies in CDRNN vs. CDR

CDR-like IRF estimates can be obtained by increasing a predictor by 1 (standard deviation) relative to the reference and observing the change in the response over time. Visualizations using this approach are presented in Figure 2 alongside CDR estimates from Shain and Schuler (2021). 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 (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

[5] For all datasets, the CDR baseline used here is the variant that was deployed on the test set in Shain and Schuler (2021).

[6] Note that a major advantage of CDRNN is its ability to model dynamics in response variance, which are not reflected in squared error. For example, although CDRNN-FF achieves worse test set error than CDR on the Natural Stories (ms) task, it affords a 31,040 point log likelihood improvement.

[7] The comparison rescales each pair of error vectors by their joint standard deviation in order to enable comparability across datasets with different error variances.

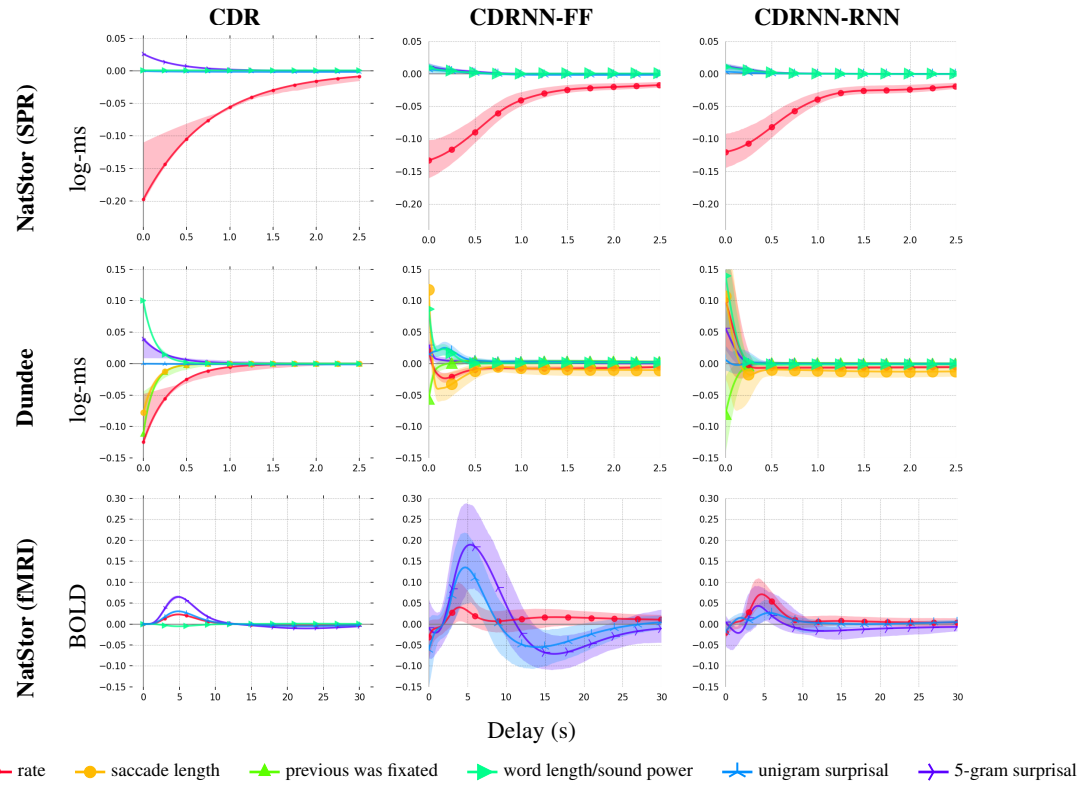

Figure 2: CDRNN-estimated IRFs across datasets, with CDR estimates from Shain and Schuler (2021) for reference. Sound power omitted from CDRNN fMRI models (see Appendix C for justification).

eye-tracking than CDR had previously suggested. Estimates from CDRNN-FF and CDRNN-RNN roughly agree, except that CDRNN-RNN estimates for fMRI are more attenuated. CDR shows little uncertainty in the fMRI domain despite its inherent noise (Shain et al., 2020), while CDRNN more plausibly shows more uncertainty in its estimates for the noisier fMRI data.

As noted in Section 2, Shain and Schuler (2021) report negative *rate* effects in reading — i.e., a local decrease in subsequent reading time at each word, especially in SPR. This was interpreted as an inertia effect (faster recent reading engenders faster current reading), but it might 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 model. CDRNN estimates nonetheless thus support the prior interpretation of *rate* effects as inertia, at least in SPR: a model that can flexibly adapt to non-linear habituation trends finds SPR *rate* estimates that are similar in shape and magnitude to those estimated by CDR.

In addition, CDRNN finds a slower response to word surprisal in self-paced reading than in eye-tracking. This result converges with word-discretized timecourses reported in Smith and Levy (2013), who find more extensive spillover of surprisal effects in SPR than in eye-tracking. Results thus reveal important hidden dynamics in the reading response (inertia effects), continuous-time delays in processing effects, and influences of modality the continuous dynamics 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).

5.3 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 curva-

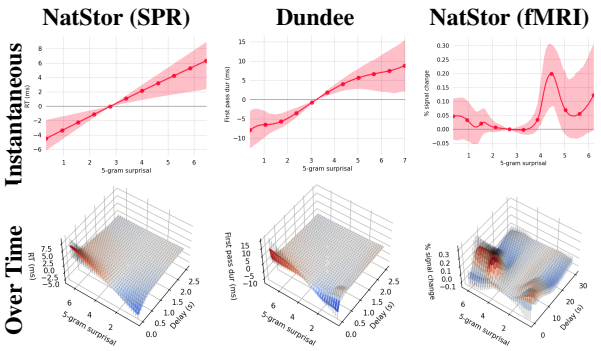

Figure 3: CDRNN-FF-estimated functional curvature of the *5-gram surprisal* response. In 3D plots, 95% credible intervals shown as vertical gray bars.

ture 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 3, reading estimates are consistent with a linear response (the credible interval contains a straight line), as predicted, but are highly non-linear in fMRI, with a rapid peak above the mean (zero-crossing) followed by a sharp dip and plateau, and even an estimated increased response at values below the mean (though estimates at the extremes have 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 3 shows visualizations to this effect. These plots in fact subsume the kinds of univariate plots shown above: univariate IRFs to *5-gram surprisal* like those plotted in Figure 2 are simply slices taken at a predictor value (1 sample standard deviation above the mean), whereas curvature estimates in the first row of Figure 3 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

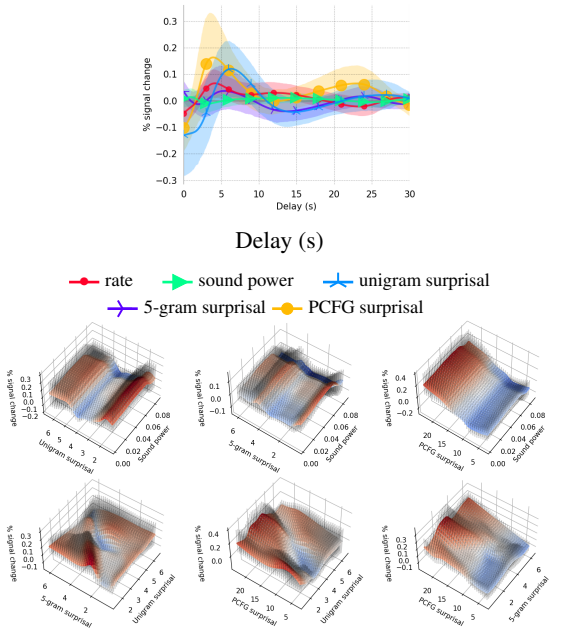

Figure 4: Effect interactions in a CDRNN-FF replication of Shain et al. (2020). 95% credible intervals shown as vertical gray bars.

as discussed above.

5.4 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 "for free", along with estimates of their functional form. To explore effect interactions, a CDRNN-FF version of the full model in Shain et al. (2020) 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 Appendix C for details). Univariate IRFs are shown in the top left panel of Figure 4, 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* (i.e. signal power of the auditory stimulus, 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).

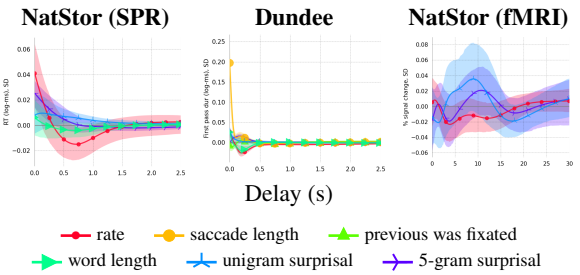

Figure 5: CDRNN-FF-estimated IRFs of the *variance* of the response by dataset.

The estimate for *unigram surprisal* (middle left) shows an unexpected non-linearity: although activity increases with higher surprisal (lower frequency words), it also increases at lower surprisal (higher frequency words), suggesting the 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 sharper increase in the *PCFG surprisal* response in higher frequency (lower unigram surprisal) regions. 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 non-linear increase in response for words that are high on both measures. This is consistent with a unitary predictive mechanism that experiences strong error signals when both string-level (5-gram) and structural (PCFG) cues are poor. All these interactions should be interpreted with caution, since the uncertainty interval covers much weaker degrees of interaction.

5.5 IRFs of the Response Variance

As discussed in Section 3, CDRNN implements *distributional regression* and thus also contains an IRF describing the influence of predictors on the *variance* of the predictive distribution as a function of time. IRFs of the variance can be visualized identically to IRFs of the mean.

For example, Figure 5 shows the estimated change in the standard deviation of the predictive distribution over time from observing a stimulus.[8] Estimates show stimulus-dependent changes in variance across datasets whose shapes are not straightforwardly related to that of the IRFs of the mean (Figure 2). For example, both reading datasets (left and center) generally show mean and standard deviation traveling together, with increases in the mean corresponding to increases in standard deviation. In Dundee, the shapes of these changes resemble each other strongly, whereas in Natural Stories the IRFs of the standard deviation (especially *rate*) differ substantially from the IRFs of the mean. By contrast, in fMRI (right), the IRFs of the standard deviation look roughly like inverted HRFs (especially for *rate* and *5-gram surprisal*), indicating that BOLD variance tends to *decrease* with larger values of the predictors. While detailed interpretation of these patterns is left to future work, these results demonstrate the utility of CDRNN for analyzing a range of links between predictors and response that are otherwise difficult to study.

[8] Because standard deviation is a bounded variable and the IRF applies before the constraint function (softplus), the relationship between the standard deviation and the y axis of the plots is not straightforward. Estimates nonetheless clearly indicate the shape and relative contribution to the response variance of the stimulus features.

6 Conclusion

This study 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 illuminate hitherto understudied properties of the human sentence processing response. This outcome suggests that CDRNN may play a valuable role in analyzing human experimental data.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.

Steven P Abney and Mark Johnson. 1991. Memory Requirements and Local Ambiguities of Parsing Strategies. *J.\ Psycholinguistic Research*, 20(3):233–250.

Amit Almor. 1999. Noun-Phrase Anaphora and Focus: The Informational Load Hypothesis. *Psychological Review*, 106(4):748–765.

Harald Baayen, Shravan Vasishth, Reinhold Kliegl, and Douglas Bates. 2017. The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 94(Supplement C):206–234.

R Harald Baayen, Doug J Davidson, and Douglas M Bates. 2007. Mixed effects modelling with crossed random effects for subjects and items. manuscript.

R Harald Baayen, Jacolien van Rij, Cecile de Cat, and Simon Wood. 2018. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. In Dirk Speelman, Kris Heylen, and Dirk Geeraerts, editors, *Mixed Effects Regression Models in Linguistics*. Springer, Berlin.

Alan D Baddeley, Neil Thomson, and Mary Buchanan. 1975. Word length and the structure of short term memory. *Journal of Verbal Learning and Verbal Behavior*, 15(6):575–589.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

H Bouma and A H De Voogd. 1974. On the control of eye saccades in reading. *Vision Research*, 14(4):273–284.

David Braze, W Einar Mencl, Whitney Tabor, Kenneth R Pugh, R Todd Constable, Robert K Fulbright, James S Magnuson, Julie A Van Dyke, and Donald P Shankweiler. 2011. Unification of sentence processing via ear and eye: An fMRI study. *cortex*, 47(4):416–431.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J Heeger, and Liina Pylkkänen. 2012. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and Language*, 120(2):163–173.

Trevor Brothers and Gina R Kuperberg. 2021. Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, 116:104174.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems 33*.

Paul-Christian Bürkner. 2018. Advanced Bayesian Multilevel Modeling with the R Package brms. *R Journal*, 10(1).

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological review*, 108(1):204.

Andreas Damianou and Neil D Lawrence. 2013. Deep gaussian processes. In *Artificial intelligence and statistics*, pages 207–215. PMLR.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL19*.

Susan F Ehrlich and Keith Rayner. 1981. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655.

Kate Erlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning & Verbal Behavior*, 22:75–87.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194.

Victoria Fossum and Roger Levy. 2012. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of {{CMCL}} 2012*. Association for Computational Linguistics.

Stefan Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of English sentence processing. *Behavior Research Methods*, 45(4):1182–1190.

Lyn Frazier and Jerry D Fodor. 1978. The sausage machine: a new two-stage parsing model. *Cognition*, 6:291–325.

Karl J Friston, Andrea Mechelli, Robert Turner, and Cathy J Price. 2000. Nonlinear responses in fMRI: The Balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. 2020. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, pages 1–15.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Edward Gibson. 2000. The Dependency Locality Theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain*, pages 95–106. MIT Press, Cambridge.

Adam Goodkind and Klinton Bicknell. 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.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2007. English Gigaword Third Edition LDC2007T07.

Daniel J Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input. *Cognitive Science*, 29:261–291.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 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.

John Hale. 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.

Caoimhe M Harrington Stack, Ariel N James, and Duane G Watson. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & cognition*, 46(6):864–877.

Trevor Hastie and Robert Tibshirani. 1986. Generalized additive models. *Statist. Sci.*, 1(3):297–310.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. 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.

John M Henderson, Wonil Choi, Matthew W Lowder, and Fernanda Ferreira. 2016. Language structure in the brain: A fixation-related fMRI study of syntactic surprisal in reading. *Neuroimage*, 132:293–300.

John M Henderson, Wonil Choi, Steven G Luke, and Rutvik H Desai. 2015. Neural correlates of fixation duration in natural reading: evidence from fixation-related fMRI. *NeuroImage*, 119:390–397.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*.

John Hewitt and Christopher D Manning. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453.

Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning*, pages 448–456.

Marcel Adam Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354.

Alan Kennedy, James Pynte, and Robin Hill. 2003. The Dundee corpus. In *Proceedings of the 12th European conference on eye movement*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6.

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

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29(3):375–419.

Martin A Lindquist, Ji Meng Loh, Lauren Y Atlas, and Tor D Wager. 2009. Modeling the hemodynamic response function in fMRI: Efficiency, bias and mismodeling. *NeuroImage*, 45(1, Supplement 1):S187 – S198.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Alessandro Lopopolo, Stefan L Frank, Antal den Bosch, and Roel M Willems. 2017. Using stochastic language models (SLM) to map lexical, syntactic, and phonological information processing in the brain. *PloS one*, 12(5):e0177794.

Mitchell P Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Don C Mitchell. 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.

Francis Mollica and Steve Piantadosi. 2017. An incremental information-theoretic buffer supports sentence processing. In *Proceedings of the 39th Annual Cognitive Science Society Meeting*.

John Morton. 1964. The effects of context upon speed of reading, eye movements and eye-voice span. *Quarterly Journal of Experimental Psychology*, 16(4):340–354.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate Unbounded Dependency Recovery using Generalized Categorial Grammars. In *Proceedings of COLING 2012*.

Dennis Norris. 2006. The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological review*, 113(2):327.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Boris T Polyak and Anatoli B Juditsky. 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Grusha Prasad and Tal Linzen. 2019. Rapid syntactic adaptation in self-paced reading: detectable, but requires many participants.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Keith Rayner. 1977. Visual attention in reading: Eye movements reflect cognitive processes. *Memory \& Cognition*, 5(4):443–448.

Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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.

Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.

Marten van Schijndel and William Schuler. 2013. An Analysis of Frequency- and Memory-Based Processing Costs. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics.

Marten van Schijndel and William Schuler. 2015. Hierarchic syntax improves reading time prediction. In *Proceedings of NAACL-HLT 2015*. Association for Computational Linguistics.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2020. Artificial Neural Networks Accurately Predict Language Processing in the Brain. *BioRxiv*.

Cory Shain, Idan Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138.

Cory Shain and William Schuler. 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*.

Cory Shain and William Schuler. 2021. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *Cognition*.

Christopher A Sims. 1971. Discrete approximations to continuous time distributed lags in econometrics. *Econometrica: Journal of the Econometric Society*, pages 545–563.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128:302–319.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C E Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *ACL19*.

Shravan Vasishth and Richard L Lewis. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.

Alberto L Vazquez, Eric R Cohen, Vikas Gulani, Luis Hernandez-Garcia, Ying Zheng, Gregory R Lee, Seong-Gi Kim, James B Grotberg, and Douglas C Noll. 2006. Vascular dynamics and BOLD fMRI: CBF level effects and analysis considerations. *Neuroimage*, 32(4):1642–1655.

Tor D Wager, Alberto Vazquez, Luis Hernandez, and Douglas C Noll. 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.

Leila Wehbe, Idan A Blank, Cory Shain, Richard Futrell, Roger Levy, Titus von der Malsburg, Nathaniel Smith, Edward Gibson, and Evelina Fedorenko. 2020. Incremental language comprehension difficulty predicts activity in the language network but not the multiple demand network. *bioRxiv*.

Ethan Wilcox, Roger Levy, and Richard Futrell. 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.

Roel M Willems, Stefan L Frank, Annabel D Nijhof, Peter Hagoort, and Antal den Bosch. 2015. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

Dataset	CDR	CDRNN-FF	CDRNN-RNN
Synth	662	7,330	17,058
NatStor (SPR)	21,845	22,546	40,408
Dundee	2,080	6,870	14,838
NatStor (fMRI)	331	13,834	26,058

Table A1: Number of trainable parameters by model and dataset.

Simon N Wood. 2006. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton.

A Mathematical Definition

This appendix formally defines the CDRNN model. CDRNN assumes the following quantities as input:[9]

- $X \in \mathbb{N}$: 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).

[9] 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.

The definition below can be 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. Regressing all parameters of the predictive distribution in this way has previously been called *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:

- $L_{\text{in}}, L_{\text{RNN}}, L_{\text{IRF}} \in \mathbb{N}$: Number of layers in the input projection, RNN, and IRF, respectively
- $D_{\text{in}(\ell)}, D_{\text{RNN}(\ell)}, D_{\text{h}}, D_{\text{IRF}(\ell)} \in \mathbb{N}$: Number of output dimensions in the ℓ^{th} layer of the input projection, RNN, hidden state, and IRF, 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 $\mathbf{W}^{\text{h,Z}} \in \mathbb{R}^{D_{\text{h}} \times Z}$, $\mathbf{W}^{\text{IRF(1),Z}} \in \mathbb{R}^{2D_{\text{IRF(1)}} \times Z}$, 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 ($\mathbf{w}_y^{\text{IRF(1),Z}}$, $\mathbf{b}_y^{\text{IRF(1),Z}}$), and the parameters of the predictive distribution ($\mathbf{e}_y^{\text{Z}}$, i.e. random intercepts and variance parameters) are generated as follows:

$$\mathbf{h}_y^{\text{Z}} \overset{\text{def}}{=} \mathbf{W}^{\text{h,Z}} \mathbf{z}_y \tag{2}$$

$$\begin{bmatrix} \mathbf{w}_y^{\text{IRF(1),Z}} \\ \mathbf{b}_y^{\text{IRF(1),Z}} \end{bmatrix} \overset{\text{def}}{=} \mathbf{W}^{\text{IRF(1),Z}} \mathbf{z}_y \tag{3}$$

$$\mathbf{s}_y^{\text{Z}} \overset{\text{def}}{=} \mathbf{W}^{\text{s,Z}} \mathbf{z}_y \tag{4}$$

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 $\mathbf{W}^{\text{h,Z}}$, $\mathbf{W}^{\text{IRF(1),Z}}$, 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:

$$g_{x,y}^{(1)}(\tau) \overset{\text{def}}{=} s_{\text{IRF(1)}}\left(\mathbf{w}_{x,y}^{\text{IRF(1)}} \tau + \mathbf{b}_{x,y}^{\text{IRF(1)}}\right) \tag{5}$$

$$g_{x,y}^{(\ell)}(\tau) \overset{\text{def}}{=} s_{\text{IRF}(\ell)}\left(\mathbf{W}^{\text{IRF}(\ell)} g_{x,y}^{(\ell-1)}(\tau) + \mathbf{b}^{\text{IRF}(\ell)}\right), \quad \ell > 1 \tag{6}$$

$$\mathbf{w}_{x,y}^{\text{IRF(1)}} \overset{\text{def}}{=} \mathbf{w}^{\text{IRF(1)}} + \mathbf{w}_y^{\text{IRF(1),Z}} + \mathbf{W}_{\Delta}^{\text{IRF(1)}} \mathbf{h}_{x,y} \tag{7}$$

$$\mathbf{b}_{x,y}^{\text{IRF(1)}} \overset{\text{def}}{=} \mathbf{b}^{\text{IRF(1)}} + \mathbf{b}_y^{\text{IRF(1),Z}} + \mathbf{B}_{\Delta}^{\text{IRF(1)}} \mathbf{h}_{x,y} \tag{8}$$

$\mathbf{W}_{x,y}^{\text{IRF}(\ell)}$, $\mathbf{b}_{x,y}^{\text{IRF}(\ell)}$, 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:

$$g_{x,y}(\tau) \overset{\text{def}}{=} \text{reshape}\left(g_{x,y}^{(L_{\text{IRF}})}(\tau), (S, K+1)\right) \tag{9}$$

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 $\mathbf{h}^{\text{in}}_{x,y}$ of the inputs at x, y, and a neural projection $\mathbf{h}^{\text{RNN}}_{x,y}$ of the hidden state of an RNN over the sequence of predictors up to and including timestep x:

$$\mathbf{h}_{x,y} \overset{\text{def}}{=} s_{\text{h}}\left(\mathbf{h}^{\text{bias}} + \mathbf{h}^{\text{Z}}_{y} + \mathbf{h}^{\text{in}}_{x,y} + \mathbf{h}^{\text{RNN}}_{x,y}\right) \quad (10)$$

The IRF $g_{x,y}$ is therefore feature-dependent via the neural projection $\mathbf{h}^{\text{in}}_{x,y}$ of the input at x, y and context-dependent via the neural projection $\mathbf{h}^{\text{RNN}}_{x,y}$ 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 $\mathbf{h}^{\text{in}}_{x,y}$ and $\mathbf{h}^{\text{RNN}}_{x,y}$ 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 $\mathbf{h}^{(0)}_{x,y}$ to the CDRNN model are defined as the vertical concatenation of the predictors $\mathbf{x}_x$ and the event timestamp t_x:

$$\mathbf{h}^{(0)}_{x,y} \overset{\text{def}}{=} \begin{bmatrix}\mathbf{x}_x \\ t_x\end{bmatrix} \quad (11)$$

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

$$\mathbf{h}^{\text{in}(\ell)}_{x,y} \overset{\text{def}}{=} s_{\text{in}(\ell)}\left(\mathbf{W}^{\text{in}(\ell)}\mathbf{h}^{\text{in}(\ell-1)}_{x,y} + \mathbf{b}^{\text{in}(\ell)}\right) \quad (12)$$

where $\mathbf{h}^{\text{in}(0)}_{x,y} \overset{\text{def}}{=} \mathbf{h}^{(0)}_{x,y}$. 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:

$$\mathbf{h}^{\text{in}}_{x,y} \overset{\text{def}}{=} \mathbf{h}^{\text{in}(L_{\text{in}})}_{x,y} \quad (13)$$

Note that $\mathbf{h}^{\text{in}}_{x,y}$ 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.[10] 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 $\mathbf{h}^{\text{RNN}(\ell)}_{x,y}$. To account for the possibility of random variation in sensitivity to context, the initial hidden and cell states $\mathbf{h}^{\text{RNN}(\ell)}_{0,y}$, $\mathbf{c}^{\text{RNN}(\ell)}_{0,y}$ depend on the random effects:

$$\mathbf{h}^{\text{RNN}(\ell)}_{0,y} \overset{\text{def}}{=} \mathbf{h}^{\text{RNN}(\ell)}_{0} + \mathbf{W}^{\text{RNN}_{\text{h}}(\ell)}_{\text{Z}}\mathbf{z}_y \quad (14)$$

$$\mathbf{c}^{\text{RNN}(\ell)}_{0,y} \overset{\text{def}}{=} \mathbf{c}^{\text{RNN}(\ell)}_{0} + \mathbf{W}^{\text{RNN}_{\text{c}}(\ell)}_{\text{Z}}\mathbf{z}_y \quad (15)$$

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

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

$$\left[\mathbf{h}^{\text{RNN}(\ell)}_{x,y}, \mathbf{c}^{\text{RNN}(\ell)}_{x,y}\right] \overset{\text{def}}{=} \text{LSTM}\left(\mathbf{h}^{\text{RNN}(\ell)}_{x-1,y}, \mathbf{c}^{\text{RNN}(\ell)}_{x-1,y}, \mathbf{h}^{\text{RNN}(\ell-1)}_{x,y}\right) \quad (16)$$

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

$$\mathbf{h}^{\text{RNN}}_{x,y} \overset{\text{def}}{=} \mathbf{W}^{\text{RNNproj}}\mathbf{h}^{\text{RNN}(L_{\text{RNN}})}_{x,y} \quad (17)$$

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 $\mathbf{y}_{[y]}$:

$$\mathbf{F}_{[y,x]} \overset{\text{def}}{=} \begin{cases} 1 & \mathbf{t}_{[x]} \leq \mathbf{t}'_{[y]} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

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 $\mathbf{h}^{(0)}_{x,y}$ and the s^{th} parameter of the predictive distribution, masked by $\mathbf{F}$:

$$\mathbf{G}_{s,k} \overset{\text{def}}{=} \begin{bmatrix} g_{1,1}(\tau_{1,1})_{[s,k]} & \cdots & g_{X,1}(\tau_{X,1})_{[s,k]} \\ \vdots & \ddots & \vdots \\ g_{1,Y}(\tau_{1,Y})_{[s,k]} & \cdots & g_{X,Y}(\tau_{X,Y})_{[s,k]} \end{bmatrix} \odot \mathbf{F} \quad (19)$$

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:

$$\mathbf{X}'^{(s)}_{[*,k]} \overset{\text{def}}{=} \mathbf{G}_{s,k}\left[\mathbf{X},\ \mathbf{t}\right]_{[*,k]} \quad (20)$$

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

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):

$$\mathbf{s} \stackrel{\text{def}}{=} \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} \tag{21}$$

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

$$\mathbf{S}^{\text{Z}} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{s}_1^{\text{Z}\top} \\ \vdots \\ \mathbf{s}_Y^{\text{Z}\top} \end{bmatrix} \tag{22}$$

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:

$$\mathbf{S}_{[*,s]} \stackrel{\text{def}}{=} f_{\text{constraint}(s)} \left(\mathbf{X}'^{(s)} \mathbf{1} + \mathbf{S}^{\text{Z}}_{[*,s]} + \mathbf{s}_{[s]} \right) \tag{23}$$

where $f_{\text{constraint}(s)}$ enforces any required constraints on the s^{th} parameter of the predictive distribution. In the Gaussian predictive distribution assumed here, $f_{\text{constraint}(1)}$ (the constraint function for the mean) is identity and $f_{\text{constraint}(2)}$ (the constraint function for the variance) is the softplus bijection:

$$\text{softplus}(x) \stackrel{\text{def}}{=} \ln(e^x + 1) \tag{24}$$

Given an assumed distributional family $\mathcal{F}$ (here assumed to be univariate normal), the response in the CDRNN model is distributed as:

$$\mathbf{y} \sim \mathcal{F} \left(\mathbf{S}_{[*,1]}, \dots, \mathbf{S}_{[*,S]} \right) \tag{25}$$

B Asynchronously Measured Predictor Dimensions

As discussed in Shain and Schuler (2018, 2021), 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 Shain and Schuler (2021) 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. Shain et al. (2020) 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.

C Predictors

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 (Shain and Schuler, 2021) (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 Shain et al. (2020) 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 Shain et al. (2020), 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 (Appendix B) and because prior *sound power* estimates in this dataset have been weak (Shain et al., 2020), *sound power* is omitted from models used in the main comparison.