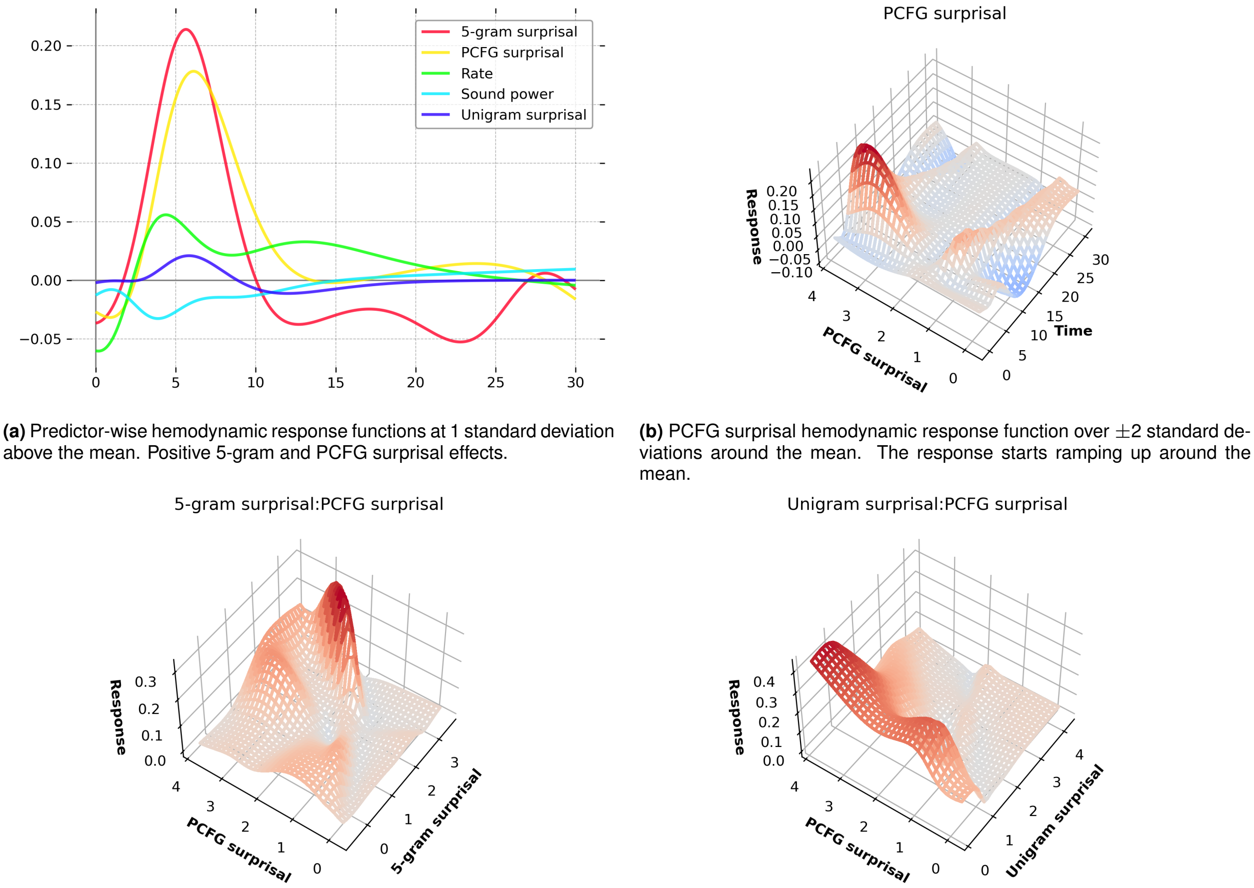
Analyzing complex human sentence processing dynamics with CDRNNs

Cory Shain and William Schuler, Ohio State

The empirical predictions of theories of incremental sentence processing are often cached out in word-level features [7, 5, 9, 8, 12], but experimental measures from human participants reflect the dynamics of real-time cognition, which may be complex, non-linear, and time-varying [1]. Thus, a central challenge in psycholinguistic theory evaluation is to specify a sufficiently expressive linking function between theory-driven word features and measures of human sentence processing, and prior work has argued that widely-used linear time series models may be inadequate, especially for naturalistic designs [1, 2, 14]. The recently proposed technique of continuous-time deconvolutional regression (CDR) relaxes assumed *instantaneity* of effects [14], instead directly estimating continuous-time *impulse response functions* (IRFs) that describe the temporal extent of a predictor's influence on the response (cf. e.g. spillover regressors [10], which ignore clock time). Empirically, CDR learns plausible effect estimates that generalize significantly better than linear models across experimental modalities [15]. However, CDR retains simplifying assumptions that may not hold of human cognition: the parametric form of the IRF must be specified in advance, the IRF is fixed over time (stationary), effects are strictly linear and additive, and the response variance is assumed to be constant (homoskedastic). Any of these constraints may be violated in practice when analyzing the outputs of a complex system like the human mind, with potential impacts on the resulting model.

In this study, we reimplement the CDR impulse response and error distribution using a time-varying deep neural network applied to the predictor sequence (CDRNN). The IRF in CDRNN is therefore a joint (potentially non-linear and interactive) function of all the predictors and time, which can be arbitrarily queried with respect to any collection of feature values, yielding a highly flexible model that relaxes all of the aforementioned simplifying assumptions. We use CDRNN both to (1) reanalyze prior claims based on CDR analysis and (2) shed new exploratory light on the dynamics of human sentence processing. In particular, we focus on the CDR-based claim from [13] that participants' word predictions are syntax-sensitive (significant effects over 5-gram surprisal of probabilistic context-free grammar or PCFG surprisal) based on activity in language-selective voxels during naturalistic listening in an fMRI experiment. To obtain this result, the authors assumed a parametric stationary *hemodynamic response function* (HRF) based on the double-gamma canonical HRF [3] and tied the parameters of the HRF across predictors within each brain region. This design improves on the standard approach of assuming the canonical HRF, instead discovering the HRF from the data [11] and allowing it to vary parametrically by region [6]. However, the HRF is known to be non-stationary, since the vascular response saturates over time [4]. Furthermore, processing effects may coordinate non-linearly [16] and non-additively, especially correlated measures such as different variants of word surprisal.

Nonetheless, CDRNN also shows that PCFG surprisal significantly improves generalization error against a 5-gram baseline ($p < 0.0001$***), validating the prior result obtained using (non-neural) CDR. *Post hoc* analyses show an estimated HRF that independently replicates known features of the HRF, including initial dip, peak response, and undershoot components (Fig 1a), despite the fact that (unlike [13]), no such *a priori* knowledge was provided. The exploratory insights afforded by CDRNN go beyond those obtainable using CDR. For example, similar *post hoc* visualizations (Fig 1b) show a PCFG surprisal response that only ramps up at values larger than its mean of 1.45 standard units (cf. [16]), suggesting that the system may be calibrated to the expected information gain per word. In addition, we find a coordination between 5-gram and PCFG surprisal near the HRF peak (Fig 1c): substantial increases in activity occur only when both variables are large, suggesting a unitary predictive mechanism that exploits both string-level and structural features (and thus incurs high error when both predictive cues are poor), rather than separate mechanisms that track each information source independently. By contrast, PCFG surprisal and unigram surprisal show a different kind of interplay (Fig 1d): there is a large jump in response to PCFG surprisal for highly frequent words (low unigram surprisal) compared to highly infrequent words (high unigram surprisal), possibly 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). Development of statistical methods for testing non-linearities in the CDRNN-estimated IRF is ongoing, but even in their absence, these results show CDRNN to be a valuable tool for exploratory data analysis, providing detailed insights about how predictors coordinate (potentially non-linearly) over time in order to determine the response. These insights can support both theoretical innovation and the design and testing of standard statistical models.



(a) Predictor-wise hemodynamic response functions at 1 standard deviation above the mean. Positive 5-gram and PCFG surprisal effects.

(b) PCFG surprisal hemodynamic response function over ±2 standard deviations around the mean. The response starts ramping up around the mean.

(c) Interaction between 5-gram and PCFG surprisal, 5s after stimulus onset. Ramp up depends on both measures being large.

(d) Interaction between unigram and 5-gram surprisal, 5s after stimulus onset. Largest PCFG surprisal effects for highly frequent (low unigram surprisal) words.

Figure 1: CDRNN estimates from naturalistic fMRI. All predictor values are in standard units.

References

[1] Baayen, H., Vasishth, S., Kliegl, R., and Bates, D. *Journal of Memory and Language*, 2017.

[2] Baayen, R. H., van Rij, J., de Cat, C., and Wood, S. Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. 2018.

[3] Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. *Journal of Neuroscience*, 1996.

[4] Friston, K. J., Mechelli, A., Turner, R., and Price, C. J. *NeuroImage*, 2000.

[5] Gibson, E. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 2000.

[6] Handwerker, D. A., Ollinger, J. M., and D'Esposito, M. *NeuroImage*, 2004.

[7] Hawkins, J. A. *A performance theory of order and constituency*, 1994.

[8] Levy, R. *Cognition*, 2008.

[9] Lewis, R. L. and Vasishth, S. *Cognitive Science*, 2005.

[10] Mitchell, D. C. *New methods in reading comprehension research*, 1984.

[11] Pedregosa, F., Eickenberg, M., Ciuciu, P., Gramfort, A., and Thirion, B. *NeuroImage*, 2014.

[12] Rasmussen, N. E. and Schuler, W. *Cognitive Science*, 2018.

[13] Shain, C., Blank, I. A., van Schijndel, M., Schuler, W., and Fedorenko, E. *Neuropsychologia*, 2020.

[14] Shain, C. and Schuler, W. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[15] Shain, C. and Schuler, W. *PsyArXiv*, 2019.

[16] Smith, N. J. and Levy, R. *Cognition*, 2013.