Coreference-aware Surprisal Predicts Brain Response

Evan Jaffe **Byung-Doh Oh** **William Schuler**
Department of Linguistics
The Ohio State University
{jaffe.59,oh.531,schuler.77}@osu.edu

Abstract

Recent evidence supports a role for coreference processing in guiding human expectations about upcoming words during reading, based on covariation between reading times and word surprisal estimated by a coreference-aware semantic processing model (Jaffe et al., 2020). The present study reproduces and elaborates on this finding by (1) enabling the parser to process subword information that might better approximate human morphological knowledge, and (2) extending evaluation of coreference effects from self-paced reading to human brain imaging data. Results show that an expectation-based processing effect of coreference is still evident even in the presence of the stronger psycholinguistic baseline provided by the subword model, and that the coreference effect is observed in both self-paced reading and fMRI data, providing evidence of the effect's robustness.

1 Introduction

Coreference resolution is a core component of language that enables comprehenders to construct a detailed representation of referents throughout a discourse. Extensive prior work has explored various conditions related to coreference resolution that affect language processing (Greene et al., 1992; Grosz et al., 1995; Gordon and Hendrick, 1998; Almor, 1999; Ariel, 2001; Cunnings et al., 2014) and have often relied on constructed stimuli to manipulate variables of interest. Complementary studies using broad-coverage, naturalistic stimuli have observed coreference effects in self-paced reading in two instantiations: (1) as a memory effect based on the count of times an entity has been previously mentioned (Jaffe et al., 2018) and (2) as an expectation-based effect, operationalized by surprisal estimates from a coreference-aware incremental parser (Jaffe et al., 2020). While expectation-based effects have been previously shown for naturalistic stimuli in self-paced reading (SPR) and functional magnetic resonance imaging (fMRI) (Smith and Levy, 2013; Shain et al., 2020), the current study extends these findings by arguing that coreference resolution contributes to predicting human behavioral data over previous implementations (e.g., surprisal) that do not model coreference.

The current study elaborates on Jaffe et al. (2020) by re-examining the expectation-based effect of coreference information using an improved baseline provided by an extension of the coreference-aware incremental parser. First, the probabilities of coreference decisions are modeled using a multilayer perceptron (MLP) model, leading to improved generalizability over the previous system based on maximum-entropy. Additionally, the incremental parser incorporates a character-based word generation model (Oh et al., 2021), which has been shown to yield surprisal estimates that predict human reading times more accurately than surprisal calculated from high-capacity neural language models. Linguistic task accuracy for coreference resolution shows improvements from the extended incremental parser, further motivating its use for psycholinguistic evaluation.

Regression analyses are conducted using surprisal estimates from the parser to determine whether coreference awareness helps explain measures of human sentence comprehension from SPR. In addition, we further evaluate whether these effects generalize to data from fMRI. Results from self-paced reading replicate Jaffe et al. (2020) by showing both (1) that coreference awareness improves the parser's approximation of human subjective surprisal and (2) that this improvement does not fully explain a previously reported facilitation effect from repeated mentions, which is plausibly driven by ease of memory retrieval. Results from fMRI support a contribution of coreference-awareness to human surprisal estimation, but fail to support a dissociable memory effect. Results

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3351–3356
November 7–11, 2021.

from both modalities thus converge in favor of the hypothesis that human linguistic expectations are sensitive to coreferential cues, with possible additional influences of memory retrieval.

2 Background

Jaffe et al. (2018) introduced *MentionCount* as a coreference-related predictor that measures the number of previous mentions for any entity. In this measure, singletons, non-entity mentions, and first mentions have a value of zero, while anaphors are assigned the number of times that entity was previously mentioned. For example, the sentence "Elon Reeve Musk is a business magnate, industrial designer and engineer. He is the founder..." would have *MentionCounts* of zero at "Musk" and one at "He". As such, more central and repeated entities receive higher values for *MentionCount*.

Jaffe et al. (2018) showed improved fit to self-paced reading times driven by *MentionCount* over surprisal and n-gram baselines, arguing that *MentionCount* could reflect a memory effect that repeated entities are easier to recall and process.

Jaffe et al. (2020) incorporated a coreference decision into a generative, incremental left-corner parser to augment its surprisal estimation with information about discourse-level entities. At its core, this model can generate prefix probabilities by marginalizing over parser states q_t and preterminal decisions $p_{1..t}$:

$$\mathsf{P}(w_{1..t}) = \sum_{p_{1..t}, q_t} \mathsf{P}(w_{1..t}\ p_{1..t}\ q_t) \qquad (1)$$

A transition model captures how these distributions are related over timesteps:

$$\mathsf{P}(w_{1..t}\ p_{1..t}\ q_t) \stackrel{\text{def}}{=} \sum_{q_{t-1}} \mathsf{P}(w_t\ p_t\ q_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \cdot \mathsf{P}(w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \qquad (2)$$

At a given timestep, the full generative process for the parser includes a lexical decision ℓ_t, preterminal decision p_t, word w_t, grammatical decision g_t and parser state q_t:

$$\begin{aligned}\mathsf{P}(w_t\ p_t\ q_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) = \\ \sum_{\ell_t, g_t} \mathsf{P}(\ell_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \cdot \\ \mathsf{P}(p_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t) \cdot \\ \mathsf{P}(w_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t) \cdot \\ \mathsf{P}(g_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t\ w_t) \cdot \\ \mathsf{P}(q_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t\ w_t\ g_t) \end{aligned} \qquad (3)$$

The parser also makes a coreference index decision that chooses an antecedent in a fixed window prior to the current word, or a special null index, which indicates no antecedent. This coreference decision is conditioned on the preterminal sequence up to the current timestep $p_{1..t}$, which includes syntactic category $c_{p_{\ell_{1..t}}}$ and predicate context $\mathbf{h}_{p_{\ell_{1..t}}}$ decisions from earlier timesteps. Syntactic category and predicate context are generated as part of the lexical decision ℓ_t during inference, and are derived from a generalized categorial grammar reannotation (Nguyen et al., 2012) of the Wall Street Journal section of OntoNotes (Weischedel et al., 2012) for training. Predicate contexts consist of a lemmatized predicate name and an argument number, such as *POUR_1*, indicating the first participant in a *pouring* predication. Together, the parser decisions generate word-by-word surprisal estimates that incorporate syntactic structure as well as propositional co-occurrences from the training data.

Recently, Oh et al. (2021) showed improved fit to self-paced reading and eye-tracking data by incorporating a character-based word generation model. Their word generation model is adopted in the current work for an improved surprisal baseline for examining coreference effects. Formally, the word probability from Equation 3 decomposes into probabilities for the lemma x_t, morphological rule r_t, and word w_t with the following conditioned-on variables:

$$\begin{aligned}\mathsf{P}(w_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t) = \\ \sum_{x_t, r_t} \mathsf{P}(x_t \mid q_{t-1}\ \ell_t\ p_t) \cdot \\ \mathsf{P}(r_t \mid q_{t-1}\ \ell_t\ p_t\ x_t) \cdot \\ \mathsf{P}(w_t \mid q_{t-1}\ \ell_t\ p_t\ x_t\ r_t) \end{aligned} \qquad (4)$$

Morphological rules that are part of the generalized categorial grammar reannotation scheme[1] (Nguyen et al., 2012) are used to generate a list of $\langle x_t, r_t \rangle$ pairs that deterministically generate the observed word w_t. The probability of the lemma x_t is modeled as the probability of generating its character sequence one-by-one from a recurrent neural network (RNN) that conditions on the syntactic category and predicate context from the lexical decision, as well as the previous character. Similarly, the probability of the morphological rule r_t is calculated by a softmax classifier that takes as input the last

[1] These rules mostly model affixation through string substitution.

hidden state of a separate RNN that receives the entire character sequence of the lemma, as well as the syntactic category and predicate context from the lexical decision. By allowing the model to posit the word's underlying structure, the parser is better able to handle out-of-vocabulary words.

3 Methods

The current work attempts to replicate Jaffe et al. (2020) but reimplements portions of their model using a multilayer perceptron for the coreference decision and a character-based word generation model for a stronger surprisal baseline. Furthermore, in addition to the SPR data analyzed in Jaffe et al. (2020), the influence of coreference information is also evaluated on fMRI data.[2] For SPR experiments, this work uses linear mixed-effects regression (LMER; Bates et al., 2015) with spillover predictors (Erlich and Rayner, 1983) and likelihood ratio tests between full and ablated models, following prior work for comparability.

fMRI studies of naturalistic language comprehension must contend with a slow hemodynamic response function (HRF) that causes effects on the response to spread out over several seconds (Boynton et al., 1996). This low temporal resolution of response data must be reconciled with relatively faster word-level predictors in our models. To accomplish this, the current study follows Shain et al. (2020) by using continuous-time deconvolutional regression (CDR; Shain and Schuler, 2018, 2021) to identify the HRF from fMRI data.

CDR models individual predictor response functions and convolves them to generate a continuous prediction of blood oxygenation level-dependent (BOLD) signals as the combination of previous events. Since the effect of a predictor on the response variable is modeled as an impulse function, predictors can have varying amplitude and decay over time. This approach therefore allows predictor and response variables to have different temporal granularity. For model details, see Appendix A.

For each fMRI experiment, two models are fit which differ minimally by the addition of a fixed effect for the predictor of interest (all models include by-subject random effects for all predictors), and correlation coefficients are calculated between each model's predictions and the fMRI observations. The difference between correlation coefficients across models provides the test statistic that is probed for significance by running a permutation test, where 10,000 permuted runs are generated to find the likelihood of the differences being at least as extreme as the observed difference.

3.1 Response Data

SPR data comes from the Natural Stories corpus (Futrell et al., 2018) and consists of reading times from 181 participants that read 10 short narratives on Amazon's Mechanical Turk platform. Filtering observations of $<$100ms and $>$3000ms, sentence-initial and sentence-final words, and participants who answered fewer than four comprehension questions correctly resulted in 768,584 observations, which were split into fit and held-out partitions (50/50). Because likelihood ratio tests with LMER (Bates et al., 2015) require the same data for fitting and evaluation, this work fits a single regression model on the held-out partition for all SPR results.

The fMRI analyses use publicly available data from Shain et al. (2020), consisting of mean responses in the most language-responsive voxels of six individually-localized regions of a left-hemisphere fronto-temporal language network, selected for analysis in light of prior evidence that this network is selective for language processing (Fedorenko et al., 2010).

This data contains BOLD measures from 78 subjects recruited from the Boston area who listened to the Natural Stories narratives for an average of 13.5 minutes during a passive comprehension task. The audio narratives consist of two audio recordings (one male, one female) presented at a normal speaking rate. This data is also split into fit and held-out partitions (50/50) by assigning alternate 60-second intervals for each subject into the two partitions. All fMRI results are fit using the 'fit' partition and evaluated on the held-out partition.

fMRI and reading time responses could be correlated based on other results using these corpora (Shain et al., 2020), but evidence also exists that they can be capturing different aspects of language processing (Oh et al., 2021).

3.2 Predictors

As in Jaffe et al. (2020), coreference-aware and coreference-unaware surprisal predictors are generated from an incremental left-corner parser described in Section 2 trained on the coreference-annotated OntoNotes corpus (Weischedel et al.,

[2] All code used in this work is available at: `github.com/modelblocks/modelblocks-release`

Coreference Model	Pro P/R/F1	Pro Acc	All P/R/F1	All Acc	Weighted Acc
MaxEnt (Jaffe et al., 2020)	87.5/80.8/84.1	41.8	**72.0**/34.1/46.3	36.2	1676.1609
MLP (this work)	87.7/**86.3/87.0**	**53.1**	70.4/**46.0/55.6**	**41.0**	**2279.8963**

Table 1: Reimplementing the coreference decision with dense feature embeddings in an MLP, together with the character-based word generation model, slightly improves coreference performance. Precision, recall, and F1 is shown for mention detection for both pronouns and all mention types. Linking accuracy is reported as the correct antecedent choice within correctly recalled mentions. Weighted linking accuracy is the product of mention F1 and mention linking accuracy. Evaluation data is the dev sections of the Wall Street Journal portion of OntoNotes.

Paradigm	Main Effect	Baseline	BaselineLogLik	FullLogLik	p-value
SPR	Δ*coref-5gramsurp*	*5gramsurp*	-2431803	-2431760	1.33e-20***
SPR	Δ*coref-nocorefsurp*	*nocorefsurp*	-2431843	-2431822	1.48e-10***
SPR	*MentionCount*	*corefsurp*	-340579	-340559	1.51e-10***
fMRI	Δ*coref-5gramsurp*	*5gramsurp*	-249797	-249763	10.00e-05***
fMRI	Δ*coref-nocorefsurp*	*nocorefsurp*	-249781	-249770	3.00e-04***
fMRI	*MentionCount*	*corefsurp*	-249761	-249757	1

Table 2: Individual model fits and significance of main effects as measured by full vs. baseline model comparisons using likelihood ratio tests for LMER (SPR) and permutation of correlation coefficients for CDR (fMRI). Baseline predictors in SPR include *word length* for all models; in fMRI, they are *end-of-sentence*, *pause duration*, and *rate*.

2012). However, this work differs in that the coreference decision is implemented with a two-layer MLP, in contrast to the maximum-entropy model used originally. The MLP uses dense embeddings for the syntactic category and predicate-context features that contribute to the coreference decision, but otherwise follows the original model. As seen in Table 1, the character-based word model and coreference MLP implementation demonstrate some improvement in coreference resolution, primarily in the recall of pronominal anaphors. While improved coreference resolution may indicate more human-like processing, it remains to be seen whether the surprisal estimates from the model will better predict SPR and fMRI data during language processing.

In order to avoid collinearity, this study uses the difference between the surprisals from the coreference-aware and coreference-unaware versions of the same parser (Δ*coref-nocorefsurp*) as a predictor that captures the contribution of coreference information. 5-gram surprisal is estimated using KenLM (Heafield et al., 2013) on the same training sections of OntoNotes as for the parser-based surprisal estimates. *Word length* is measured in characters.

CDR models (fMRI only) include the deconvolutional intercept *rate*, which estimates the base response of the system to a stimulus (Shain and Schuler, 2021). Wrap-up effects are controlled using an indicator for *end-of-sentence*, and prosodic effects are controlled using *pause duration*, the time elapsed (in ms) for any pauses in speech during the audio recording. To avoid wrap-up effects at the end of the scanning session, all images following the end of the audio stimulus are dropped.

4 Results

4.1 Self-paced Reading Data

The results in Table 2 show that Δ*coref-nocorefsurp* significantly improves fit to SPR data over a coreference-unaware surprisal baseline by a likelihood ratio test ($p < .0001$). Additionally, the delta predictor between coreference-aware surprisal and 5-gram surprisal (Δ*coref-5gramsurp*) significantly improves fit to SPR data over the 5-gram surprisal baseline by a likelihood ratio test ($p < .0001$). *MentionCount* significantly improves fit to SPR data over the coreference-aware surprisal baseline by likelihood ratio test ($p < .0001$). These results on SPR data are consistent with prior results reported by Jaffe et al. (2020).

4.2 fMRI Data

As with SPR data, Δ*coref-nocorefsurp* significantly improves fit to fMRI data over a coreference-unaware surprisal baseline by a paired permutation test evaluating the improvement in correlation between the predicted and true responses on the held-out partition ($p < .0001$). Δ*coref-5gramsurp* also

significantly improves fit to fMRI data over the 5-gram surprisal baseline by the same permutation test ($p < .0001$). However, *MentionCount* does not significantly improve fit over a coreference-aware surprisal baseline predictor ($p = 1$).

Taken together, coreference-aware surprisal is robustly attested in both SPR and fMRI as a strong predictor of psycholinguistic data. The mixed results showing an effect of *MentionCount* in SPR but not in fMRI suggest that memory might be variably recruited in SPR vs. passive listening tasks, where SPR requires more memory resources. Similarly, it may be that fMRI as a dependent variable with language-specific localization is tracking different language processes than those evident in reading time latencies (Oh et al., 2021).

5 Conclusion

This study reproduces previously reported coreference effects in self-paced reading using an improved surprisal estimator baseline, finding evidence for a coreference-driven expectation effect during naturalistic reading. Additionally, a new analysis using fMRI data shows that coreference-aware surprisal contributes to significantly better fit, further supporting the overall claim that expectation-based language processing utilizes coreferential cues. However, a memory retrieval effect for coreference is observed in SPR but not in fMRI, highlighting the complex nature of human coreference processing and offering potential future directions for investigation.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Amit Almor. 1999. Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological review*, 106(4):748–765.

Mira Ariel. 2001. Accessibility theory: An overview. In *Text representation: Linguistic and psycholinguistic aspects*, pages 29–87. John Benjamins Publishing Company, Amsterdam, The Netherlands.

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.

Geoffrey M. Boynton, Stephen A. Engel, Gary H. Glover, and David J. Heeger. 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16(13):4207–4221.

Ian Cunnings, Clare Patterson, and Claudia Felser. 2014. Variable binding and coreference in sentence comprehension: Evidence from eye movements. *Journal of Memory and Language*, 71(1):39–56.

Kate Erlich and Keith Rayner. 1983. Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of Verbal Learning and Verbal Behavior*, 22(1):75–87.

Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. 2010. ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104:1177–1194.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2018. The Natural Stories Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 76–82.

Peter C. Gordon and Randall Hendrick. 1998. The representation and processing of coreference in discourse. *Cognitive Science*, 22:389–424.

Steven B. Greene, Gail McKoon, and Roger Ratcliff. 1992. Pronoun resolution and discourse models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(2):266–283.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

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.

Evan Jaffe, Cory Shain, and William Schuler. 2018. Coreference and focus in reading times. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–9.

Evan Jaffe, Cory Shain, and William Schuler. 2020. Coreference information guides human expectations during natural reading. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4587–4599.

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2125–2140.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3746–4757.

Cory Shain, Idan Asher 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*, pages 2679–2689.

Cory Shain and William Schuler. 2021. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *Cognition*, 215.

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

Ralph Weischedel, Sameer S. Pradhan, Lance Ramshaw, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Nianwen Xue, Martha Palmer, Jena Hwang, Claire Bonial, Jinho Choi, Aous Mansouri, Maha Foster, Abdel-aati Hawwary, Mitchell Marcus, Ann Taylor, Craig Greeberg, Eduard Hovy, Robert Blevin, and Ann Houston. 2012. OntoNotes. Technical report, Linguistics Data Consortium.

A CDR Implementation

Following Shain et al. (2020), CDR models used in this study tie the parameters of the HRF across predictors, thereby allowing predictors to vary only the the scale of their influence on the response. Models include random HRF estimates by fROI and random intercepts by subject, and are fitted with black box variational inference using default priors as described in Shain and Schuler (2021). The CDR codebase is available at `https://github.com/coryshain/cdr`, and complete model configuration files are available at `https://github.com/modelblocks/modelblocks-release`.