Coreference information guides human expectations during natural reading

Evan Jaffe **Cory Shain** **William Schuler**
The Ohio State University
{jaffe.59,shain.3,schuler.77}@osu.edu

Abstract

Models of human sentence processing effort tend to focus on costs associated with retrieving structures and discourse referents from memory (memory-based) and/or on costs associated with anticipating upcoming words and structures based on contextual cues (expectation-based) (Levy, 2008). Although evidence suggests that expectation and memory may play separable roles in language comprehension (Levy et al., 2013), theories of coreference processing have largely focused on memory: how comprehenders identify likely referents of linguistic expressions. In this study, we hypothesize that coreference tracking also informs human expectations about upcoming words, and we test this hypothesis by evaluating the degree to which incremental surprisal measures generated by a novel coreference-aware semantic parser explain human response times in a naturalistic self-paced reading experiment. Results indicate (1) that coreference information indeed guides human expectations and (2) that coreference effects on memory retrieval may exist independently of coreference effects on expectations. Together, these findings suggest that the language processing system exploits coreference information both to retrieve referents from memory and to anticipate upcoming material.

1 Introduction

Maintaining a model of discourse referents and their relations is a core function of human language understanding and depends on the ability to recognize when linguistic expressions corefer. Identifying the referent of a linguistic expression (coreference resolution) is a well-established task in natural language processing, and a sizeable psycholinguistic literature has explored the computations that support coreference resolution in humans. Theories of coreference processing generally agree on a critical role played by searching associative memory stores for plausible referents (Greene et al., 1992; Grosz et al., 1995; Gordon, 1998; Almor, 1999; Ariel, 2001), based on findings that coreference is more easily established to more salient or activated referents, as evidenced by human reading latencies (Gordon, 1998; Cunnings et al., 2014). Nonetheless, some coreference-related phenomena are not obviously related to search. Consider the following example:

> Sally frightened her brother because **Sally/she/he**...

Numerous studies indicate that *he* in the boldfaced slot is more difficult to process (i.e. read more slowly) than *she* despite the fact that *he* corefers unambiguously to *her brother*, the most recently activated discourse referent (Garvey and Caramazza, 1974; Garnham et al., 1996; Stewart et al., 2000; Koornneef and Van Berkum, 2006; Hartshorne, 2014, *inter alia*). This is due to the *implicit causality* of the verb *frightened*, which predisposes comprehenders to expect an explanation in terms of the agent of the frightening event, rather than the patient (Garvey and Caramazza, 1974). In addition, the full name *Sally* is more difficult to process than the pronoun *she*, despite the fact that both expressions unambiguously identify their referent. This is due to a phenomenon known as the *repeated name penalty* (Gordon et al., 1993). All of these expressions unambiguously identify their target referents, and for some, the preferred

Proceedings of the 28th International Conference on Computational Linguistics, pages 4587–4599
Barcelona, Spain (Online), December 8-13, 2020

referent is not necessarily the most recent. It is thus not straightforward to account for these effects solely on the basis of retrieval difficulty. Various accounts of these and related coreference phenomena exist, such as the *informational load hypothesis* of Almor (1999), which posits a trade-off between the information content of referring expressions and their linguistic complexity. However, to our knowledge, little prior work has explored the role coreference might play in more general *expectation-based* theories of human sentence processing (Hale, 2001; Levy, 2008; van Schijndel et al., 2013; Rasmussen and Schuler, 2018), according to which comprehension difficulty is driven by mechanisms that incrementally reallocate cognitive resources between competing interpretations of the unfolding sentence. Expectation-based theories have been highly successful in providing broad-coverage explanations of sentence processing difficulty, but have so far been evaluated using local lexical (*n*-gram) or syntactic (parser-based) contexts (Demberg and Keller, 2008; Frank and Bod, 2011; Smith and Levy, 2013, *inter alia*). We show here that they can also be a useful framework for understanding coreference-related phenomena like those mentioned above, since knowledge of available discourse referents and patterns of referential form might guide comprehenders' expectations about upcoming words.

In this study, we implement a novel coreference-aware generative incremental semantic parser and show that coreference awareness significantly improves fit of surprisal estimates from the parser to human reaction times in a self-paced reading experiment. We further show that a previously proposed estimator of coreference resolution difficulty explains additional variance over that explained by the parser, suggesting that the memory phenomenon is at least partially independent of expectation. Together, these results indicate that comprehenders incrementally exploit coreference information both to identify referents of linguistic expressions (a memory effect) and to inform predictions about upcoming words (an expectation effect). Our study thus joins previously reported evidence of the need for both memory-related and expectation-related mechanisms in accounting for the full range of human sentence processing phenomena (Levy et al., 2013).

2 Previous Work

Coreference resolution has been extensively studied, both as a psycholinguistic phenomenon and as a natural language processing task. Psycholinguistic investigations have largely focused on the memory retrieval mechanisms that allow comprehenders to identify the referents of linguistic expressions (Greene et al., 1992; Grosz et al., 1995; Gordon, 1998; Almor, 1999; Ariel, 2001; Kehler and Rohde, 2015; Jaffe et al., 2018). Some computational models have attempted to implement these theories in order to solve the coreference resolution task. For example, Wiemer-Hastings and Iacucci (2001) implement Gordon (1998), Tetreault (2002) is inspired by Centering Theory (Grosz et al., 1995), and Webster and Nothman (2016) draw on the Accessibility Hierarchy (Ariel, 2001) for feature design. Webster and Curran (2014) propose an incremental coreference resolution system inspired by Centering and Accessibility theories by maintaining a ranked stack of entities supposed to reflect the reader's mental state. Kehler and Rohde (2015) offer a Bayesian account of pronominal reference that highlights an asymmetry between pronoun production and interpretation.

Other computational models are less directly inspired by human cognition but achieve strong performance on the coreference resolution task. Wiseman et al. (2016) use LSTMs to track each entity cluster, where the hidden state is updated with each mention that is linked to the cluster, allowing for use of entity-level features. Lee et al. (2017) propose an end-to-end coreference resolution system without external resources using bidirectional LSTMs and an attention mechanism for head finding.

However, none of the above systems integrate explicit coreference tracking into a full generative model of natural language. As a result, while such systems may be amenable to studying coreference retrieval processes in human comprehenders, they cannot be used to test the present hypothesis that coreference information guides comprehenders' overall expectations about upcoming words.

Also of relevance are recent high-performance large-scale language models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and GPT-2 (Radford et al., 2018), which have been argued to learn coreference information (Clark et al., 2019). However, these models rely on powerful deep neural networks that are difficult to interpret, and it is generally not possible to manipulate their access to

coreference information in order to evaluate the impact of this information on human expectations.

Ji et al. (2017) propose an incremental generative model that explicitly tracks coreference but the joint objectives are not depedendent on each other, meaning that they do not model word expectation as dependent on the coreference decision.

3 Background

The models used in this paper obtain surprisal predictors from an incremental processing model based on a left-corner parser, trained on corpora annotated with both syntactic and coreference information (Weischedel et al., 2012), which produces a simple incremental probabilistic account of sentence processing by making a single lexical attachment decision and a single grammatical attachment decision at each word.

3.1 Surprisal from Processing Models

Surprisal can be calculated as the difference between negative log probabilities of prefixes generated by a processing model at consecutive time steps:

$$\mathsf{S}(w_t) \stackrel{\text{def}}{=} -\log \mathsf{P}(w_t|w_{1..t-1}) = -\log \mathsf{P}(w_{1..t}) + \log \mathsf{P}(w_{1..t-1}) \tag{1}$$

These prefix probabilities can be calculated by marginalizing over hidden states in the forward probabilities of an incremental processing model:

$$\mathsf{P}(w_{1..t}) = \sum_{q_t} \mathsf{P}(w_{1..t}\ q_t) \tag{2}$$

These forward probabilities are in turn defined recursively using a transition model:

$$\mathsf{P}(w_{1..t}\ q_t) \stackrel{\text{def}}{=} \sum_{q_{t-1}} \mathsf{P}(w_t\ q_t \mid q_{t-1}) \cdot \mathsf{P}(w_{1..t-1}\ q_{t-1}) \tag{3}$$

3.2 Left-corner Parsing

The transition model described in this paper is based on a probabilistic left-corner parser (Johnson-Laird, 1983; Roark, 2001; van Schijndel et al., 2013; Rasmussen and Schuler, 2018), in part because it requires a bounded amount of working memory at every time step, and because it makes a fixed number of decisions after every word.

The model maintains a distribution over possible working memory store states q_t at every time step t, each of which consist of a bounded number D of nested derivation fragments a_t^d/b_t^d (partial phrase structure trees). Each derivation fragment at a given depth d spans a part of a possible derivation tree from some apex sign a_t^d (syntactic category in the case of a purely syntactic parser) to some base sign b_t^d in its right progeny. Previous work has shown large tree-annotated resources such as the Penn Treebank (Marcus et al., 1993) do not require more than $D = 4$ such fragments (Schuler et al., 2010).

Like other incremental parsers (Dyer et al., 2016; Hale et al., 2018; Jin and Schuler, 2020) this model makes a series of probabilistic decisions which incrementally define a syntactic structure. The left-corner parsing model defines new words w_t and store states q_t by first making a *lexical* decision ℓ_t (related to hypothesizing a preterminal above w_t and possibly attaching it to an existing derivation fragment) and then making a *grammatical* decision g_t (related to hypothesizing some minimal binary branch that subsumes both w_t and w_{t+1} and possibly attaching it to an existing derivation fragment) after encountering each word:[1]

$$\begin{aligned}\mathsf{P}(w_t\ q_t \mid q_{t-1}) = \sum_{\ell_t, g_t} & \mathsf{P}(\ell_t \mid q_{t-1}) \cdot \\ & \mathsf{P}(w_t \mid q_{t-1}\ \ell_t) \cdot \\ & \mathsf{P}(g_t \mid q_{t-1}\ \ell_t\ w_t) \cdot \\ & \mathsf{P}(q_t \mid q_{t-1}\ \ell_t\ w_t\ g_t)\end{aligned} \tag{4}$$

[1] Johnson-Laird (1983) refers to these decisions as 'shift' and 'predict,' respectively.

$$\frac{a^1_{t-1}/b^1_{t-1} \ldots a^d_{t-1}/b^d_{t-1}\ w_t}{a^1_{t-1}/b^1_{t-1} \ldots a^d_{t-1}/b^d_{t-1}\ a^{d+1}_t}\ m_{\ell_t} = 0 \qquad \frac{a^1_{t-1}/b^1_{t-1} \ldots a^{d-1}_{t-1}/b^{d-1}_{t-1}\ a^d_{t-1}}{a^1_{t-1}/b^1_{t-1} \ldots a^{d-1}_{t-1}/b^{d-1}_{t-1}\ a^d_t/b^d_t}\ m_{g_t} = 0$$

$$\frac{a^1_{t-1}/b^1_{t-1} \ldots a^d_{t-1}/b^d_{t-1}\ w_t}{a^1_{t-1}/b^1_{t-1} \ldots a^d_{t-1}}\ m_{\ell_t} = 1 \qquad \frac{a^1_{t-1}/b^1_{t-1} \ldots a^{d-1}_{t-1}/b^{d-1}_{t-1}\ a^d_{t-1}}{a^1_{t-1}/b^1_{t-1} \ldots a^{d-1}_{t-1}/b^{d-1}_t}\ m_{g_t} = 1$$

Figure 1: Natural deduction rules for parsing. The parse is a list of derivation fragments where a lexical match and grammatical match rules modifies the structure at each timestep. Negative lexical match ($m_{\ell_t} = 0$) posits a new left-child sign at increased depth. Positive lexical match ($m_{\ell_t} = 1$) discharges the expected base sign from the previous timestep. Negative grammatical match ($m_{g_t} = 0$) maintains the current derivation fragment and generates new apex and base signs for it, while positive grammatical match ($m_{g_t} = 1$) joins the current derivation fragment with its superordinate fragment and generates a new base sign.

Together, these decisions generate the n unary branches and $n - 1$ binary branches in a Chomsky normal form tree spanning an n-word sentence.

Each lexical decision includes a match decision $m_{\ell_t} \in \{0, 1\}$ about whether to identify the preterminal above the next lexical item as the base sign of the derivation fragment above it, and a decision about the category label c_{ℓ_t} for this preterminal sign. Each grammatical decision includes another match decision $m_{g_t} \in \{0, 1\}$ about whether to identify the lowest nonterminal sign that contains but does not end with the current lexical item as the base sign of the derivation fragment above it, as well as a decision about the category label c_{g_t} of this nonterminal, and a decision about the category label c'_{g_t} of its right child. Derivation fragments above this nonterminal are then carried forward, and derivation fragments below it are set to null ($\perp$):

$$\mathsf{P}(q_t \mid q_{t-1}\ \ell_t\ w_t\ g_t) \stackrel{\text{def}}{=} \prod_{d=1}^{D} \begin{cases} [\![a^d_t, b^d_t = a^d_{t-1}, b^d_{t-1}]\!] & \text{if } d < \delta \\ [\![a^d_t, b^d_t = a_{g_t}, b_{g_t}]\!] & \text{if } d = \delta \\ [\![a^d_t, b^d_t = \perp, \perp]\!] & \text{if } d > \delta \end{cases} \tag{5}$$

where indicator $[\![\varphi]\!] = 1$ if φ is true and 0 otherwise, and $\delta = \text{argmax}_{d'}\{a^{d'}_{t-1} \neq \perp\} + 1 - m_{\ell_t} - m_{g_t}$, and signs a_{g_t} and b_{g_t} are here simply c_{g_t} and c'_{g_t}, respectively. Figure 1 shows the natural deduction rules for the lexical and grammatical match rules, and resulting structures from such decisions.

3.3 Semantic Processing Model

Experiments in this paper use an extension of the left-corner parsing model described above (Oh and Schuler, forthcoming) which incorporates referential semantics by augmenting each preterminal, apex, and base sign to include not only a category label c_{p_t}, $c_{a^d_t}$ or $c_{b^d_t}$, but also a referential context vector $\mathbf{h}_{p_t}$, $\mathbf{h}_{a^d_t}$ or $\mathbf{h}_{b^d_t} \in \{0, 1\}^{K+V \cdot K}$, where K is the dimension of the referential context vector for the predicate and each argument, and V is the maximum valence or number of syntactic arguments of any category. This set of up to V syntactic arguments is well-defined for phrase structure grammars like HPSG (Pollard and Sag, 1994) or categorial grammars like CCG (Steedman, 2000) or GCG (Bach, 1981; Nguyen et al., 2012), and may include non-local arguments such as gap-fillers, relative and interrogative pronoun antecedents, extraposed or heavy-shifted items, and passive subjects.

Vectors for syntactic predicates and arguments may be dense, generated by neural network models in order to maximize prediction accuracy, but experiments described in this paper will adopt a sparse encoding generated by explicit features in order to perform a clean ablation with a coreference model and avoid confounds related to hyperparameter tuning. In this sparse encoding, a vector for a referent signified by a sign contains a set of *referential context* indicators (Levy and Goldberg, 2014), each consisting of a predicate constant and a path of up to three labeled associations that connect that predicate to that referent.

For example, POUR might be a referential context for the eventuality of a pour predicate, and POUR_2 might be a referential context for the second argument of a pour predicate.

The model is then augmented to generate not only match decisions and category labels for each lexical and grammatical decision, but also sparse binary vectors for these referential contexts. Match decisions are then estimated using logistic regression on pairs of referential contexts from these sparse vectors.

Lexical decisions now directly generate not only match m_{ℓ_t} and preterminal category labels c_{ℓ_t} but also referential context vectors $\mathbf{h}_{\ell_t}$ for preterminal signs, which define complete preterminal signs p_{ℓ_t}:

$$p_{\ell_t} \stackrel{\text{def}}{=} \begin{cases} c_{b^{\delta}_{t-1}}, \mathbf{h}_{b^{\delta}_{t-1}} + \mathbf{h}_{\ell_t} & \text{if } m_{\ell_t} = 1 \\ c_{\ell_t}, \mathbf{h}_{\ell_t} & \text{if } m_{\ell_t} = 0 \end{cases} \tag{6}$$

The sparse encoding adopted in these experiments defines each lexical preterminal vector as a Kronecker delta (one-hot vector) concentrated at the lemmatized predicate corresponding to the current word, with no path (in the case of verbs, adjectives, prepositions and relational nouns) or with a path '1' (in the case of non-relational nouns, whose signified referents are usually the first participant of an elementary predication labeled by that noun). For example, POUR is a referential context indicating a 'pouring' predicate, whereas CLOUD_1 is a context indicating the first argument of a 'being a cloud' predicate. Contexts with other paths (e.g., POUR_1 for entities that 'pour' or POUR_2 for entities that are 'poured') are generated by operators in grammatical decisions.

Grammatical decisions do not generate referential context vectors directly, but rather generate (in addition to match decisions m_{g_t} and category labels c_{g_t} and c'_{g_t}) triples of operators o_{g_t}, o'_{g_t}, o''_{g_t}, corresponding to unary operations (e.g. to re-order syntactic arguments or to associate non-local syntactic arguments with local syntactic arguments), or operations to derive left or right child signs from parent signs (e.g. to obtain argument or modifier referents from predicate referents or to attach left children as non-local arguments of right children). These operators are then associated with elements of a fixed set of composition matrices $\mathbf{O}_{o_{g_t}}$, $\mathbf{O}_{o'_{g_t}}$, $\mathbf{O}_{o''_{g_t}}$, which can then be used to compose signs a_{g_t} and b_{g_t} in Equation 5. For example, a verb with context POUR might predict a direct object noun phrase sibling with context POUR_2.

In these grammatical decisions, the apex sign a_{g_t} used in Equation 5 is unchanged from the previous time step when there is a match in the grammatical phase, and when there is no match, it is calculated by traversing up from the left child (via $\mathbf{O}_{o'_{g_t}}{}^{\top}$) and then up through the unary operations at the parent (via $\mathbf{O}_{o_{g_t}}{}^{\top}$):

$$a_{g_t} \stackrel{\text{def}}{=} \begin{cases} a^{\delta}_{t-1} & \text{if } m_{\ell_t,g_t} = 1,1 \\ c_{g_t}, \mathbf{O}_{o_{g_t}}{}^{\top}\mathbf{O}_{o'_{g_t}}{}^{\top}\mathbf{h}_{a^{\delta}_{t-1}} & \text{if } m_{\ell_t,g_t} = 1,0 \\ a^{\delta}_{t-1} & \text{if } m_{\ell_t,g_t} = 0,1 \\ c_{g_t}, \mathbf{O}_{o_{g_t}}{}^{\top}\mathbf{O}_{o'_{g_t}}{}^{\top}\mathbf{h}_{p_{\ell_t}} & \text{if } m_{\ell_t,g_t} = 0,0 \end{cases} \tag{7}$$

The base sign b_{g_t} used in Equation 5 is calculated by traversing from the left child to the parent (via $\mathbf{O}_{o'_{g_t}}{}^{\top}$) and then down to the right child (via $\mathbf{O}_{o''_{g_t}}$) when there is no match in the grammatical phase, and when there is a match, it is calculated by combining the result of traversing up from the left child (via $\mathbf{O}_{o'_{g_t}}{}^{\top}$) and down from the parent (via $\mathbf{O}_{o_{g_t}}$) and then traversing down to the right child (via $\mathbf{O}_{o''_{g_t}}$):

$$b_{g_t} \stackrel{\text{def}}{=} \begin{cases} c'_{g_t}, \mathbf{O}_{o''_{g_t}}(\mathbf{O}_{o_{g_t}}\mathbf{h}_{b^{\delta}_{t-1}} + \mathbf{O}_{o'_{g_t}}{}^{\top}\mathbf{h}_{a^{\delta+1}_{t-1}}) & \text{if } m_{\ell_t,g_t} = 1,1 \\ c'_{g_t}, \mathbf{O}_{o''_{g_t}}\mathbf{O}_{o'_{g_t}}{}^{\top}\mathbf{h}_{a^{\delta}_{t-1}} & \text{if } m_{\ell_t,g_t} = 1,0 \\ c'_{g_t}, \mathbf{O}_{o''_{g_t}}(\mathbf{O}_{o_{g_t}}\mathbf{h}_{b^{\delta}_{t-1}} + \mathbf{O}_{o'_{g_t}}{}^{\top}\mathbf{h}_{p_{\ell_t}}) & \text{if } m_{\ell_t,g_t} = 0,1 \\ c'_{g_t}, \mathbf{O}_{o''_{g_t}}\mathbf{O}_{o'_{g_t}}{}^{\top}\mathbf{h}_{p_{\ell_t}} & \text{if } m_{\ell_t,g_t} = 0,0 \end{cases} \tag{8}$$

This model will serve as the non-coreference baseline in experiments described in Sections 5 and 6.

Aqua_i	thought	he_i	was	lucky	to	see	a	cloud_j	pouring	its_j	rain	onto	him_i
0	0	1	0	0	0	0	0	0	0	1	0	0	2

Figure 2: An example sentence with predictor values for MentionCount (Jaffe et al., 2018), the number of previous mentions to the same referent. Words sharing a subscript value are coreferential.

$$\frac{p_{1\ldots t-1}\ a^1_{t-1}/b^1_{t-1}\ldots a^d_{t-1}/b^d_{t-1}\ w_t}{p_{1\ldots t}\ a^1_{t-1}/b^1_{t-1}\ldots a^d_{t-1}/b^d_{t-1}\ a^{d+1}_t}\ m_{\ell_t}=0 \qquad \frac{p_{1\ldots t}\ a^1_{t-1}/b^1_{t-1}\ldots a^{d-1}_{t-1}/b^{d-1}_{t-1}\ a^d_{t-1}}{p_{1\ldots t}\ a^1_{t-1}/b^1_{t-1}\ldots a^{d-1}_{t-1}/b^{d-1}_{t-1}\ a^d_t/b^d_t}\ m_{g_t}=0$$

$$\frac{p_{1\ldots t-1}\ a^1_{t-1}/b^1_{t-1}\ldots a^d_{t-1}/b^d_{t-1}\ w_t}{p_{1\ldots t}\ a^1_{t-1}/b^1_{t-1}\ldots a^d_{t-1}}\ m_{\ell_t}=1 \qquad \frac{p_{1\ldots t}\ a^1_{t-1}/b^1_{t-1}\ldots a^{d-1}_{t-1}/b^{d-1}_{t-1}\ a^d_{t-1}}{p_{1\ldots t}\ a^1_{t-1}/b^1_{t-1}\ldots a^{d-1}_{t-1}/b^{d-1}_t}\ m_{g_t}=1$$

Figure 3: Natural deduction rules for parsing with coreference. Note the presence of the additional preterminal sequence allows semantic contexts to be inherited from antecedents to anaphors. Otherwise, the rules are identical.

4 Coreference Model

Like the semantic parser described above, the coreference-aware parser[2] is trained with explicit supervision from annotated corpora to maximize the probability of all variable values. However, a number of changes and additions are necessary to add coreference information to the parser described above.

The coreference-aware model marginalizes prefix probabilities off of entire sequences of durable preterminal signs $p_{1..t}$ (for mathematical convenience, we henceforth split the lexical decision ℓ_t above into a lexical decision ℓ_t and a preterminal decision p_t) as well as the current store state q_t:

$$\mathsf{P}(w_{1..t}) = \sum_{p_{1..t}, q_t} \mathsf{P}(w_{1..t}\ p_{1..t}\ q_t) \tag{9}$$

These sequence probabilities are again defined recursively using a transition model with a larger sequential context:

$$\mathsf{P}(w_{1..t}\ p_{1..t}\ q_t) \stackrel{\text{def}}{=} \sum_{q_{t-1}} \mathsf{P}(w_t\ p_t\ q_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \cdot \mathsf{P}(w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \tag{10}$$

Since the size of this distribution increases exponentially with the length of the sentence, only hidden state sequences with optimal $p_{1..t-1}$ are maintained on a beam. This larger context is propagated to the lexical model:

$$\begin{aligned}
&\mathsf{P}(w_t\ p_t\ q_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) = \\
&\quad \sum_{\ell_t, g_t} \mathsf{P}(\ell_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \cdot \\
&\qquad \mathsf{P}(p_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t) \cdot \\
&\qquad \mathsf{P}(w_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t) \cdot \\
&\qquad \mathsf{P}(g_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t\ w_t) \cdot \\
&\qquad \mathsf{P}(q_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t\ p_t\ w_t\ g_t)
\end{aligned} \tag{11}$$

which is augmented with a factor over coreference indices i_{ℓ_t} (highlighted):

$$\begin{aligned}
\mathsf{P}(\ell_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \stackrel{\text{def}}{=} &\ \mathsf{P}(i_{\ell_t} \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \\
&\cdot \mathsf{P}(m_{\ell_t}\ \mathbf{h}_{\ell_t}\ c_{\ell_t} \mid i_{\ell_t}\ q_{t-1})
\end{aligned} \tag{12}$$

[2]All code is publicly available at `https://github.com/modelblocks/modelblocks-release`

Referential contexts for preterminals are then augmented with contexts of antecedents $\mathbf{h}_{p_{t-i_{\ell_t}}}$ (highlighted), where this term is a zero vector in case there is no coreference:

$$\mathsf{P}(p_t \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}\ \ell_t) \stackrel{\text{def}}{=} \begin{cases} [\![p_t = c_{b_t^\delta}, \mathbf{h}_{b_t^\delta} + \mathbf{h}_{\ell_t} + \mathbf{h}_{p_{t-i_{\ell_t}}}]\!] & \text{if } m_{\ell_t} = 1 \\ [\![p_t = c_{\ell_t}, \mathbf{h}_{\ell_t} + \mathbf{h}_{p_{t-i_{\ell_t}}}]\!] & \text{if } m_{\ell_t} = 0 \end{cases} \tag{13}$$

The grammatical decisions retain the original context but depend on the lexical decision. Adding referential context $\mathbf{h}_{\ell_t}$ conditioned on coreference index i_{ℓ_t} allows the probabability of lexical decision ℓ_t to be influenced by the referential contexts of prior coreferent material. Since grammatical decisions depend on the lexical decision in a given timestep, parsing and coreference are fully interactive.

For example, at the word *its* in Figure 2, the model would consider coreference antecedent offsets including those corresponding to *he* (ten words back: $i_{\ell_t} = 10$) and *cloud* (two words back: $i_{\ell_t} = 2$). With the correct antecedent, the referential context of *its* would inherit the CLOUD_1 context, making the following word *rain* more likely than if it had inherited the incorrect antecedent context HE_1.

Coreference decisions are modeled as a conditional probability distribution of an offset index i_{ℓ_t} at time t, given the referential context sets of the base sign and of all previous preterminal signs:

$$\mathsf{P}(i_{\ell_t} \mid w_{1..t-1}\ p_{1..t-1}\ q_{t-1}) \stackrel{\text{def}}{=} \frac{\mathsf{P}(n_{t,i_{\ell_t}}\ \forall_{i' \neq i_{\ell_t}}\ \neg n_{t,i'} \mid p_1 \ldots p_{t-1}\ b_{t-1}^\delta)}{\sum_i \mathsf{P}(n_{t,i}\ \forall_{i' \neq i}\ \neg n_{t,i'} \mid p_1 \ldots p_{t-1}\ b_{t-1}^\delta)} \tag{14}$$

where non-anaphors are represented with an offset index of zero. This distribution is defined as the probability that a binary indicator model chooses the word at time step $t - i$ as being immediately coreferential with (i.e. a direct antecedent of) the word at the current time step t and also not immediately coreferential with any other offset i' prior to time t. This positive coreference choice is normalized by all other positive coreference choices of other offsets i. Probabilities for these indicator variables are then estimated independently (using logistic regression on pairs of referential contexts) and multiplied together:

$$\mathsf{P}(n_{t,i}\ \forall_{i' \neq i}\ \neg n_{t,i'} \mid p_1 \ldots p_{t-1}\ b_{t-1}^\delta) \stackrel{\text{def}}{=} \mathsf{P}(n_{t,i} \mid p_{t-i}\ b_{t-1}^\delta) \cdot \prod_{i' \neq i} \mathsf{P}(\neg n_{t,i'} \mid p_{t-i'}\ b_{t-1}^\delta) \tag{15}$$

Natural deduction rules for the coreference-aware parser are shown in Figure 3; lexical and grammatical match rules follow the same structural outcomes but the presence of an additional preterminal list now allows for coreference information to be inherited as described in Equation 13.

5 Experiment 1

To test the hypothesis that comprehenders use coreference information to guide linguistic expectations, the non-coreference parsing model described in Section 3.3 and the joint parsing and coreference model described in Section 4 are trained on version 5 of the coreference-annotated subset of the OntoNotes corpus (Weischedel et al., 2012), totaling 12,256 sentences, reannotated into a generalized categorial grammar (Nguyen et al., 2012) to provide valences and operators for referential contexts. In a baseline model, a variety of control predictors including the non-coreference-aware parser surprisal are then regressed to self-paced reading times from the Natural Stories (Futrell et al., 2018) corpus using the R LMER (Bates et al., 2008) package. Then, a full model augments the baseline model with a fixed effect for the coreference-aware surprisal estimate and a likelihood ratio test (LRT) is performed. By-subject random effects are included for all predictors in both models.

5.1 Data

Natural Stories consists of 10 naturalistic stories with data from 181 participants who took part in a self-paced reading experiment.The data are filtered to exclude sentence initial and final words to control for edge effects, removing sentences for which respondents answered fewer than 4 comprehension questions correctly, and reading time durations less than 100ms or greater than 3000ms. This results in 768,584 reading time observations which are partitioned into an exploratory set of 383,906 observations and a held-out set of 384,678 observations. The partition allows for parameter selection on exploratory data and a single hypothesis test on held-out data, eliminating the need for multiple trials correction.

Effect	β (z)	β (ms)
Word Length	7.487	3.22
Story Position	-15.47	-41.26
NgramSurp	9.958	5.40
NoCorefSurpS1	0.676	0.118
CorefSurpS1	3.198	0.554

Table 1: Experiment 1 fixed effect estimates for full model on held-out data partition. Positive effects correspond to increases in reading time duration while negative effects correspond to a decrease in reading time duration.

5.2 Predictors and Spillover

Low-level control predictors include: *Word Length*, the length of a word measured in characters, and *Story Position*, a measure of the proportion of the story completed, ranging from 0 to 1 and calculated as current sentence index divided by total sentence count. *Story Position* is meant to control for order effects and task learning and habituation (Baayen et al., 2017; Jaffe et al., 2018).

Three surprisal predictors are also included: *NgramSurp* is the incremental surprisal estimate based on a 5-gram language model trained on the Gigaword corpus (Graff and Cieri, 2003) using the KenLM toolkit (Heafield et al., 2013). It is meant to account for word predictability based on local lexical context alone. *NoCorefSurp* is the baseline incremental surprisal estimate that comes from the parser described above in Section 3.3. Surprisal is calculated as the difference in log probability from the previous timestep to the current timestep as allocated on the beam (see Equation 1). *CorefSurp* is the incremental surprisal estimate that comes from augmenting the parser with coreference resolution, as described in Section 4.

Because the effect of interest may be delayed, it is standard psycholinguistic practice to consider *spillover* values of predictors from preceding words (Rayner et al., 1983). However, including all plausible spillover variants for every predictor often leads to identifiability problems in mixed effects modeling (Shain and Schuler, 2019). For this reason, we select the single spillover position that maximizes effect magnitude in the exploratory set for each predictor in this study. To optimize, we use a stepwise selection procedure on the exploratory set starting with control variables and moving on to critical (surprisal) variables. At each step (predictor), we select the spillover position up to three words away that maximizes the effect estimate for the predictor, then add the selected spillover variant to the model. This procedure makes the model maximally parsimonious while enabling us to account for delayed effects. Selection based on effect size in this mixed model setting focuses the selection procedure on effects that generalize most strongly across participants. Using this procedure, NgramSurp had the largest fixed effect magnitude in situ and was added first. Once NgramSurp was in the baseline, the NoCorefSurp fixed effect was found to be strongest when spilled-over by one (S1). Since CorefSurp should be compared fairly with NoCorefSurp, it was also spilled over by one position. NoCorefSurp and CorefSurp are consequently spilled over one word (S1), while all other predictors are not spilled over. Results for both experiments are shown for held-out data that was not optimized directly for spillover position. All predictors were z-scored prior to fitting.

5.3 Results

Experiment 1 results show an improvement in fit to reading time data when using a coreference-aware incremental surprisal estimate. The predictor CorefSurpS1 is statistically significant (p = 5.6e-5) according to a likelihood ratio test (LRT) between the ablated and full model, where the full model differs minimally to include a fixed effect for CorefSurpS1. A positive effect estimate means that increased surprisal correlates with increased reading time. Table 1 shows the fixed effect estimates for the full model and Table 2 shows the p-value for the LRT.

One concern is that jointly training coreference with parsing could create noise for basic parsing decisions if incorrect referential context is used to condition parsing decisions. However, categorial

Comparison	p
CorefSurpS1 over NoCorefSurpS1	5.6e-5
MentionCount over CorefSurpS1	3.4e-5

Table 2: Main result. Hypothesis tests for critical coreference predictors show that CorefSurpS1 is significant over a baseline model that includes low-level predictors and NoCorefSurpS1. MentionCount is significant over a baseline model that includes Story Position, NgramSurp, and CorefSurpS1.

Model	Training Corpus	F-score
NoCoref	OntoNotes	77.2
Coref	OntoNotes	76.9

Table 3: Categorial grammar parsing performance is comparable between coreference and non-coreference versions of the parser.

grammar labeled bracketing F scores for the parser[3] trained with and without coreference actually show comparable performance, as seen in Table 3. Additionally, the incremental coreference submodel does indeed specifically learn coreference information.[4]

Analysis of residuals further supports the hypothesis that improved fit is driven in part by better control of coreference effects. Binning reading time residuals by word position, we find that the average mean residual for word positions following a pronoun is significantly smaller in magnitude in the full coreference-aware model (-7.67) than in the ablated model (-9.13) (p = 3.79e-48 by paired t-test).[5]

6 Experiment 2

It is conceivable that expectation alone is sufficient to account for coreference-related processing costs that have previously been attributed to memory. To test this possibility, we conduct a follow-up analysis in which we test whether the number of previous mentions of a discourse referent (MentionCount) facilitates its retrieval from memory, as has been argued by a previous study that did not directly control for the effects of coreference on expectation (Jaffe et al., 2018). See Table 2 for example values of MentionCount.

Two models are regressed to self-paced reading time data from the NaturalStories corpus, and an LRT assesses the difference in fit resulting from adding a fixed effect for MentionCount. Following Jaffe et al. (2018), we restrict the analysis to head words of anaphoric expressions, leaving 54,026 observations in the held-out data. Using CorefSurpS1 in the baseline means that a fair comparison would also spill over MentionCount by one, resulting in MentionCountS1.[6] *Word Length* was removed as the weakest baseline predictor in order to achieve model convergence.

6.1 Results

MentionCount shows a significant facilitation effect over the novel coreference-aware surprisal baseline (Tables 2 and 4), strengthening evidence for MentionCount as a possible memory effect in self-paced reading, separate from effects of coreference information on expectation.

[3]These results are not directly comparable to PTB parsing results due to the finer-grained label set of the categorial grammars used to define valences and operators for referential contexts. Additionally, when trained on the coreference-annotated subset of sentences in OntoNotes, the parser is undertrained compared to when training on the entire Treebank (roughly 4x larger).

[4]For pronouns, anaphoricity detection recall is 81% and the accuracy of correct entity choices when restricted to correctly recalled pronominal mentions is 42%. Correct entity choice here means choosing any correct antecedent in the chain. For all words, anaphoricity detection shows recall: 34%, precision: 72%, F1: 40%, and accuracy of correct entity choices when restricted to correctly recalled mentions is 36%; this corresponds to non-incremental scores of: MUC:23.47, B^3:17.66, CEAFm:24.6, CEAFe:17.8, and BLANC:10.28. Although these are relatively low compared to state-of-the-art coreference models, this model is incremental and generative (in order to provide surprisal estimates) and is computing joint syntactic probabilities, and the coreference performance is sufficient to show significant effects.

[5]By contrast, the average mean residual over all word positions is near zero for both models (full: 0.003, ablated 0.004) and not significantly different. This is an expected outcome given the definition of linear mixed effects models.

[6]Using in situ CorefSurp and MentionCount would likely result in a larger effect based on spillover optimization results; the spilled over results presented in experiment 2 is a more conservative and still significant result.

Effect	β(z)	β(ms)
Story Position	-11.26	-30.03
NgramSurp	13.11	7.11
CorefSurpS1	3.32	0.58
MentionCountS1	-3.92	-0.20

Table 4: Fixed effects from Experiment 2 full model. MentionCount is significant when added to a strong baseline including lexical, syntactic **and** coreference controls. Word Length is omitted for convergence reasons as the weakest predictor. There are 54,026 observations from 180 subjects. With a range of 1 to 90 in this corpus, MentionCountS1 accounts for approximately 20ms of reading time facilitation at its max value.

7 Conclusion

In this study, we have argued that expectation-based theories of human sentence processing (Hale, 2001; Levy, 2008; van Schijndel et al., 2013; Rasmussen and Schuler, 2018) can be augmented to account for the influence of coreference information on human comprehenders' expectations about upcoming words. We proposed a novel coreference-aware generative incremental semantic parser and showed empirically that coreference awareness explains significant variance in human reading behavior. We nevertheless showed that a previously proposed effect of coreference information on memory retrieval survives in the presence of the coreference-aware predictability estimate, indicating that this effect cannot be explained away by expectation as currently implemented.[7] Our results suggest that coreference information may be used by multiple subroutines in the language comprehension system, both to retrieve likely antecedents from memory and to inform expectations about upcoming words. This proposal is in line with prior arguments that both memory and expectation are necessary to account for the full range of human sentence processing phenomena.

8 Acknowledgements

This material is based upon work supported by the National Science Foundation under grant no. 1551313 and 1816891. Any opinions, findings, and conclusions or recommendations expressed in this material 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, October.

Mira Ariel. 2001. Accessibility theory: An overview. In *Text Representation: Linguistic and Psycholinguistic Aspects*, pages 29–87.

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.

Emmon Bach. 1981. Discontinuous constituents in generalized categorial grammars. *Proceedings of the Annual Meeting of the Northeast Linguistic Society (NELS)*, 11:1–12.

Douglas Bates, Martin Maechler, and Bin Dai. 2008. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-31.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. In *BlackBoxNLP 2019*.

[7] It is possible that MentionCount, as a measure of repetition, could be interpreted as information that affects expectation (repeated entities are more likely to be mentioned again) or reflects stronger activation in memory, or both. Futrell et al. (2020) offer an account that incorporates memory and expectation, but their proposal more directly covers distance-sensitive locality effects rather than the repetition effect here.

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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019*, pages 4171–4186.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent Neural Network Grammars. In *NAACL-HLT 2016*.

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

Lynn Friedman and Melanie Wall. 2005. Graphical views of suppression and multicollinearity in multiple linear regression. *American Statistician*, 59(2):127–136.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2018. The natural stories corpus. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 76–82.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science*, 44(3), 3.

Alan Garnham, Matthew Traxler, Jane Oakhill, and Morton Ann Gernsbacher. 1996. The locus of implicit causality effects in comprehension. *Journal of memory and language*, 35(4):517–543.

Catherine Garvey and Alfonso Caramazza. 1974. Implicit causality in verbs. *Linguistic inquiry*, 5(3):459–464.

N. J. Gordon, D. J. Salmond, and A. F. M. Smith. 1993. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113.

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

David Graff and Christopher Cieri, 2003. *English Gigaword LDC2003T05*.

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

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

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R. Brennan. 2018. Finding syntax in human encephalography with beam search. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1(2014):2727–2736.

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.

Joshua K Hartshorne. 2014. What is implicit causality? *Language, Cognition and Neuroscience*, 29(7):804–824.

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, August.

Evan Jaffe, Cory Shain, and William Schuler. 2018. Coreference and Focus in Reading Times. In *Cognitive Modeling and Computational Linguistics (CMCL)*, pages 1–9.

Yangfeng Ji, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith. 2017. Dynamic entity representations in neural language models. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1830–1839.

Lifeng Jin and William Schuler. 2020. Memory-bounded Neural Incremental Parsing for Psycholinguistic Prediction. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task*, pages 48–61.

Philip N. Johnson-Laird. 1983. *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.

Andrew Kehler and Hannah Rohde. 2015. Pronominal Reference and Pragmatic Enrichment: A Bayesian Analysis. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, (2008):1063–1068.

Arnout W Koornneef and Jos JA Van Berkum. 2006. On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54(4):445–465.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 188–197.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 302–308.

Roger Levy, Evalina Fedorenko, and Edward Gibson. 2013. The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, 69:461–495.

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

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.

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 (COLING '12)*, pages 2125–2140, Mumbai, India.

Byung-Doh Oh and William Schuler. 2020. Contributions of propositional content and syntactic categories in sentence processing. *manuscript submitted for publication*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL HLT 2018*, pages 2227–2237.

Carl Pollard and Ivan Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. Technical report.

Nathan E Rasmussen and William Schuler. 2018. Left-Corner Parsing With Distributed Associative Memory Produces Surprisal and Locality Effects. *Cognitive Science*, 42:1009–1042.

Keith Rayner, Marcia Carlson, and Lyn Frazier. 1983. The interaction of syntax and semantics during sentence processing: Eye movements in the analysis of semantically biased sentences. *Journal of verbal learning and verbal behavior*, 22(3):358–374.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage incremental parsing using human-like memory constraints. *Computational Linguistics*, 36(1):1–30.

Cory Shain and William Schuler. 2019. Continuous-Time Deconvolutional Regression for Psycholinguistic Modeling. *PsyArXiv*.

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

Mark Steedman. 2000. *The syntactic process*. MIT Press/Bradford Books, Cambridge, MA.

Andrew J Stewart, Martin J Pickering, and Anthony J Sanford. 2000. The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language*, 42(3):423–443.

Joel R. Tetreault. 2002. A corpus-based evaluation of centering theory. *NLE Technical Note TN-02-*, (Section 2).

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.

Kellie Webster and JR Curran. 2014. Limited memory incremental coreference resolution. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*.

Kellie Webster and Joel Nothman. 2016. Using mention accessibility to improve coreference resolution. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437, Berlin, Germany, August. Association for Computational Linguistics.

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.

Peter Wiemer-Hastings and Carlo Iacucci. 2001. A computational model of human coreference judgements. *Proceedings of the First Workshop on Cognitively Plausible Models of Semantic Processings (SEMPRO 2001)*.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning Global Features for Coreference Resolution.

Lee H. Wurm and Sebastiano A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*.