Contributions of Propositional Content and Syntactic Category Information in Sentence Processing

Byung-Doh Oh
Department of Linguistics
The Ohio State University
oh.531@osu.edu

William Schuler
Department of Linguistics
The Ohio State University
schuler@ling.osu.edu

Abstract

Expectation-based theories of sentence processing posit that processing difficulty is determined by predictability in context. While predictability quantified via surprisal has gained empirical support, this representation-agnostic measure leaves open the question of how to best approximate the human comprehender's latent probability model. This work presents an incremental left-corner parser that incorporates information about both propositional content and syntactic categories into a single probability model. This parser can be trained to make parsing decisions conditioning on only one source of information, thus allowing a clean ablation of the relative contribution of propositional content and syntactic category information. Regression analyses show that surprisal estimates calculated from the full parser make a significant contribution to predicting self-paced reading times over those from the parser without syntactic category information, as well as a significant contribution to predicting eye-gaze durations over those from the parser without propositional content information. Taken together, these results suggest a role for propositional content and syntactic category information in incremental sentence processing.

1 Introduction

Much work in sentence processing has been dedicated to studying differential patterns of processing difficulty in order to shed light on the latent mechanism behind online processing. As it is now well-established that processing difficulty can be observed in behavioral responses (e.g. reading times, eye movements, and event-related potentials), recent psycholinguistic work has tried to account for these variables by regressing various predictors of interest. Most notably, in support of expectation-based theories of sentence processing (Hale, 2001; Levy, 2008), predictability in context has been quantified through the information-theoretical measure of surprisal (Shannon, 1948). Although there has been empirical support for n-gram, PCFG, and LSTM surprisal in the literature (Goodkind and Bicknell, 2018; Hale, 2001; Levy, 2008; Shain, 2019; Smith and Levy, 2013), as surprisal makes minimal assumptions about linguistic representations that are built during processing, this leaves open the question of how to best estimate the human language comprehender's latent probability model.

One factor related to memory usage that has received less attention in psycholinguistic modeling is the influence of *propositional content*, or meaning that is conveyed by the sentence. Early psycholinguistic experiments have demonstrated that the propositional content of utterances tends to be retained in memory, whereas the exact surface form and syntactic structure are forgotten (Bransford and Franks, 1971; Jarvella, 1971). This suggests that memory costs related to incrementally constructing a representation of propositional content might manifest themselves in behavioral responses during online sentence processing. In addition, there is evidence suggesting that parsing decisions are informed by the ongoing interpretation of the sentence (Brown-Schmidt et al., 2002; Tanenhaus et al., 1995).

Based on this insight, prior cognitive modeling research has sought to incorporate propositional content information into various complexity metrics. A prominent approach in this line of research has been to quantify complexity based on the compatibility between a predicate and its arguments (i.e. *thematic fit*, Baroni and Lenci 2010, Chersoni et al. 2016, Padó et al. 2009). However, these complexity metrics can only be evaluated at a coarse per-sentence level or at critical regions of constructed stimuli where predicates and arguments are revealed, making them less suitable for studying online processing. A more distribu-

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 241–250
Online Event, June 10, 2021. ©2021 Association for Computational Linguistics

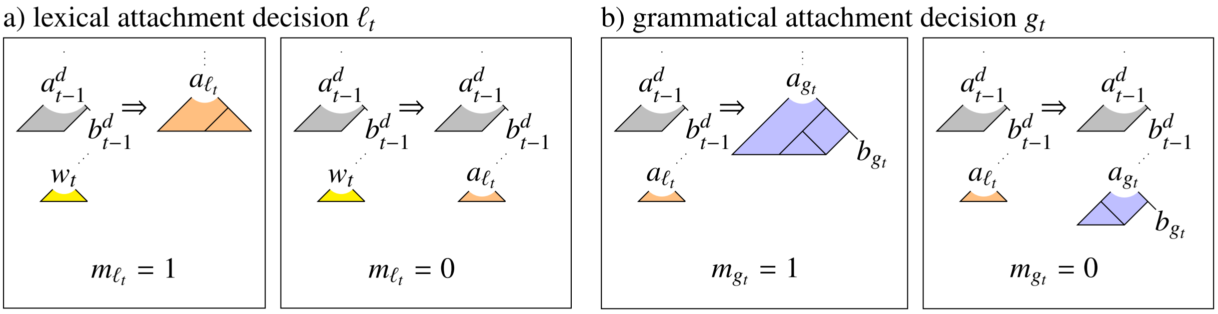

Figure 1: Left-corner parser operations: a) lexical match (m_{ℓ_t}=1) and no-match (m_{ℓ_t}=0) operations, creating new apex a_{ℓ_t}, and b) grammatical match (m_{g_t}=1) and no-match (m_{g_t}=0) operations, creating new apex a_{g_t} and base b_{g_t}.

tional approach has also been explored that relies on word co-occurrence to calculate the *semantic coherence* between each word and its preceding context (Mitchell et al., 2010; Sayeed et al., 2015). Although these models allow more fine-grained per-word metrics to be calculated, their dependence on an aggregate context vector makes it difficult to distinguish 'gist' or topic information from propositional content.

Unlike these models, our approach seeks to incorporate propositional content by augmenting a generative and incremental parser to build an ongoing representation of *predicate context vectors*, which is based on a categorial grammar formalism that captures both local and non-local predicate-argument structure. This processing model can be used to estimate per-word surprisal predictors that capture the influence of propositional content differentially with that of syntactic categories, which are devoid of propositional content.[1] Our experiments demonstrate that the incorporation of both propositional content and syntactic category information into the processing model significantly improves fit to self-paced reading times and eye-gaze durations over corresponding ablated models, suggesting their role in online sentence processing. In addition, we present exploratory work showing how our processing model can be utilized to examine differential effects of propositional content in memory-intensive filler-gap constructions.

2 Background

The experiments presented in this paper use surprisal predictors calculated by an incremental processing model based on a probabilistic left-corner parser (Johnson-Laird, 1983; van Schijndel et al., 2013). This incremental processing model provides a probabilistic account of sentence processing by making a single lexical attachment decision and a single grammatical attachment decision for each input word.[2]

Surprisal can be defined as the negative log of a conditional probability of a word w_t and a state q_t at some time step t given a sequence of preceding words $w_{1..t-1}$, marginalized over these states:

$$\mathsf{S}(w_t) \overset{\text{def}}{=} -\log \sum_{q_t} \mathsf{P}(w_t\ q_t \mid w_{1..t-1}) \qquad (1)$$

These conditional probabilities can in turn be defined recursively using a transition model:

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

A probabilistic left-corner parser defines its transition model over possible working memory store states $q_t = a^1_t/b^1_t, \ldots, a^D_t/b^D_t$, each of which consists of a bounded number D of nested derivation fragments a^d_t/b^d_t. Each derivation fragment spans a part of a derivation tree below some apex node a^d_t, lacking a base node b^d_t yet to come.

At each time step, the parser generates a lexical attachment decision ℓ_t, a word w_t, a grammatical at-

[1] Note that this distinction of propositional content as retained information about the meaning of a sentence and syntactic categories as unretained information about the form of a sentence may differ somewhat from notions of semantics and syntax that are familiar to computational linguists – in particular, predicates corresponding to lemmatized words fall on the content side of this division here because they are retained after processing, even though it may be common in NLP applications to use them in syntactic parsing.

[2] Johnson-Laird (1983) refers to lexical and grammatical attachment decisions as 'shift' and 'predict' respectively.

$$\begin{array}{lll}
\mathsf{many}\ (\lambda_{x_1}\ \mathsf{some}\ (\lambda_{e_1}\ \mathsf{person}\ e_1\ x_1) & & \\
\qquad\qquad (\lambda_{e_1}\ \mathsf{true})) & & \\
\quad (\lambda_{x_1}\ \mathsf{some}\ (\lambda_{x_3}\ \mathsf{some} & (\lambda_{e_3}\ \mathsf{pasta}\ e_3\ x_3) & \\
 & (\lambda_{e_3}\ \mathsf{true})) & \\
\qquad\qquad (\lambda_{x_3}\ \mathsf{some} & (\lambda_{e_2}\ \mathsf{eat}\ e_2\ x_1\ x_3) & \\
 & (\lambda_{e_2}\ \mathsf{true}))) &
\end{array}$$

Figure 2: Lambda calculus expression for the propositional content of the sentence *Many people eat pasta,* using generalized quantifiers over discourse entities and eventualities.

tachment decision g_t, and a resulting store state q_t:

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

As shown in Figure 1, the lexical attachment decision ℓ_t generates a new complete node a_{ℓ_t} based on (m_{ℓ_t}) whether the word matches the base of the most recent derivation fragment; and the grammatical attachment decision g_t generates a new derivation fragment a_{g_t}/b_{g_t} based on (m_{g_t}) whether the parent of a grammar rule with this new complete node as a left child matches the base of the most recent remaining derivation fragment.

The semantic processing model described in this paper extends the above left-corner parser to incorporate propositional content by conditioning lexical and grammatical decisions on sparse vectors of predicate contexts $\mathbf{h}_{a_t^d}$ and $\mathbf{h}_{b_t^d}$ in addition to category labels $c_{a_t^d}$ and $c_{b_t^d}$ in apex and base nodes a_t^d and b_t^d. These predicate context vectors for nodes in a derivation tree of a sentence can be defined in terms of argument positions of variables signified by these nodes in predicates of a logical form translation of that sentence. For example, in Figure 2, the variable e_2 (signified by the word *eat*) would have the predicate context EAT_0 because it is the zeroth (initial) participant of the predication (eat $e_2\ x_1\ x_3$).[3] Similarly, the variable x_3 would have both the predicate context PASTA_1, because it is the first participant (counting from zero) of the predication (pasta $e_3\ x_3$), and the predicate context EAT_2, because it is the second participant (counting from zero) of the predication (eat $e_2\ x_1\ x_3$). These predicate contexts are obtained by reannotating the training corpus using a generalized categorial grammar of English (Nguyen et al., 2012), which is sensitive to syntactic valence and non-local dependencies.

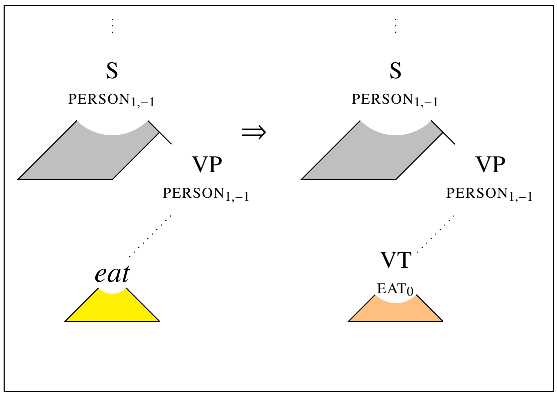

Figure 3: Derivation fragments resulting from example lexical decisions made at the word *eat* in the sentence *People eat pasta*. Note that the *predicate contexts* instead of *predicate context vectors* are displayed here for clarity. The predicate context $\text{PERSON}_{1,-1}$ represents an eventuality that takes the first argument of a PERSON predicate as its first argument.

Lexical attachment probabilities. The probability of each lexical decision ℓ_t in this parser is therefore decomposed into one term for generating a match decision m_{ℓ_t} and a predicate context vector $\mathbf{h}_{\ell_t}$, and another term for generating a syntactic category label c_{ℓ_t} for the new complete node a_{ℓ_t}:

$$\begin{aligned}
&\mathsf{P}(\ell_t \mid q_{t-1}) = \\
&\qquad \mathsf{P}(m_{\ell_t}\ \mathbf{h}_{\ell_t} \mid q_{t-1}) \cdot \mathsf{P}(c_{\ell_t} \mid q_{t-1}\ m_{\ell_t}\ \mathbf{h}_{\ell_t}) \quad (4)
\end{aligned}$$

The probability of generating the match decision and the predicate context vector depends on the base node b_{t-1}^d of the previous derivation fragment:

$$\begin{aligned}
&\mathsf{P}(m_{\ell_t}\ \mathbf{h}_{\ell_t} \mid q_{t-1}) = \\
&\qquad \underset{m_{\ell_t}\mathbf{h}_{\ell_t}}{\textsc{SoftMax}}(\,\mathrm{FF}_{\theta_\mathrm{L}}[\delta_d{}^\top, [\delta_{c_{b_{t-1}^d}}^\top, \mathbf{h}_{b_{t-1}^d}^\top]\,\mathbf{E}_\mathrm{L}]\,) \quad (5)
\end{aligned}$$

where FF is a feedforward neural network, δ_i is a Kronecker delta vector consisting of a one at element i and zeros elsewhere, depth $d = \operatorname{argmax}_{d'}\{a_{t-1}^{d'} \neq \bot\}$ is the number of non-null derivation fragments at the previous time step, and $\mathbf{E}_\mathrm{L}$ is a matrix of jointly trained dense embeddings for each syntactic category and predicate context. The probabilities of category labels are calculated using relative frequency estimation on training data based on the base node of the previous derivation

[3] Participants of predications are numbered starting with zero so as to align loosely with syntactic arguments in canonical form.

fragment. The new complete node a_{ℓ_t} then depends on the match decision m_{ℓ_t} (see Figure 3):

$$a_{\ell_t} \overset{\text{def}}{=} \begin{cases} a^d_{t-1} & \text{if } m_{\ell_t} = 1 \\ c_{\ell_t}, \mathbf{h}_{\ell_t} & \text{if } m_{\ell_t} = 0 \end{cases} \tag{6}$$

Word probabilities. Probabilities for generating words are estimated as the probability of generating their character sequence using a recurrent neural network implementation of a character model.

Grammatical attachment probabilities. The probability of each grammatical decision g_t in this parser is similarly decomposed into a term for generating a match decision m_{g_t} and a composition operator for a grammar rule o_{g_t},[4] and terms for category labels c_{g_t} and c'_{g_t} at the apex and base nodes of the new derivation fragment:

$$\begin{aligned} \mathsf{P}(g_t \mid q_{t-1}\ \ell_t\ w_t) = {} & \mathsf{P}(m_{g_t}\ o_{g_t} \mid q_{t-1}\ \ell_t\ w_t) \cdot \\ & \mathsf{P}(c_{g_t} \mid q_{t-1}\ \ell_t\ w_t\ m_{g_t}\ o_{g_t}) \cdot \\ & \mathsf{P}(c'_{g_t} \mid q_{t-1}\ \ell_t\ w_t\ m_{g_t}\ o_{g_t}\ c_{g_t}) \end{aligned} \tag{7}$$

The probability of generating the match decision and the composition operator depends on the base node of the previous derivation fragment and the new complete node a_{ℓ_t}:

$$\begin{aligned} & \mathsf{P}(m_{g_t}\ o_{g_t} \mid q_{t-1}\ \ell_t\ w_t) = \\ & \underset{m_{g_t} o_{g_t}}{\textsc{SoftMax}}(\,\mathrm{FF}_{\theta_{\mathrm{G}}}[\delta_d{}^\top, [\delta^\top_{c_{b^{d-m_{\ell_t}}_{t-1}}}, \mathbf{h}^\top_{b^{d-m_{\ell_t}}_{t-1}}, \delta^\top_{c_{a_{\ell_t}}}, \mathbf{h}^\top_{a_{\ell_t}}]\,\mathbf{E}_{\mathrm{G}}]\,) \end{aligned} \tag{8}$$

where $\mathbf{E}_{\mathrm{G}}$ is a matrix of jointly trained dense embeddings for each syntactic category and predicate context. The probabilities of category labels c_{g_t} and c'_{g_t} in Equation 7 are calculated using relative frequency estimation on training data based on the base node of the previous derivation fragment. The composition operator o_{g_t} in Equations 7 and 8 is associated with sparse composition matrices $\mathbf{A}_{o_{g_t}}$, which can be used to compose predicate context vectors associated with the apex node a_{g_t} of the new derivation fragment,

$$a_{g_t} \overset{\text{def}}{=} \begin{cases} a^{d-m_{g_t}}_{t-1} & \text{if } m_{g_t} = 1 \\ c_{g_t}, \mathbf{A}_{o_{g_t}} \mathbf{h}_{a_{\ell_t}} & \text{if } m_{g_t} = 0 \end{cases} \tag{9}$$

and sparse composition matrices $\mathbf{B}_{o_{g_t}}$, which can be used to compose predicate context vectors associated with the base node b_{g_t} of the new derivation fragment (see Figure 4):

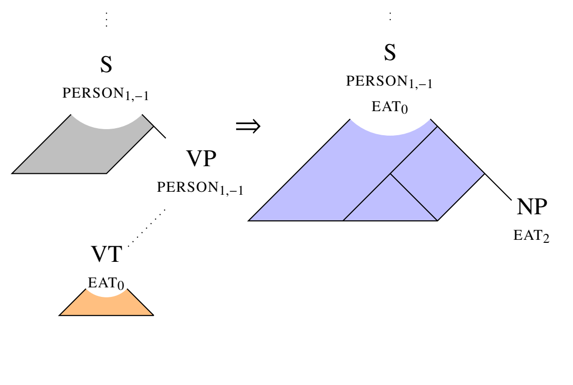

Figure 4: Derivation fragments resulting from example grammatical decisions made at the word *eat* in the sentence *People eat pasta*.

$$b_{g_t} \overset{\text{def}}{=} \begin{cases} c'_{g_t}, \mathbf{B}_{o_{g_t}}[\mathbf{h}_{b^{d-m_{\ell_t}}_{t-1}}{}^\top, \mathbf{h}_{a_{\ell_t}}{}^\top]^\top & \text{if } m_{g_t}{=}1 \\ c'_{g_t}, \mathbf{B}_{o_{g_t}}[\mathbf{0}^\top, \mathbf{h}_{a_{\ell_t}}{}^\top]^\top & \text{if } m_{g_t}{=}0 \end{cases} \tag{10}$$

These composition matrices allow predicate contexts to propagate appropriately through the tree to allow parsing decisions to depend on predicates that may be several words away.

Resulting store state probabilities. In order to update the store state based on the lexical and grammatical decisions, derivation fragments above the most recent nonterminal node are carried forward, and derivation fragments below it are set to null ($\perp$),

$$\mathsf{P}(q_t \mid \ldots) \overset{\text{def}}{=} \prod_{d=1}^{D} \begin{cases} [\![a^d_t, b^d_t = a^d_{t-1}, b^d_{t-1}]\!] & \text{if } d < d' \\ [\![a^d_t, b^d_t = a_{g_t}, b_{g_t}]\!] & \text{if } d = d' \\ [\![a^d_t, b^d_t = \perp, \perp]\!] & \text{if } d > d' \end{cases} \tag{11}$$

where the indicator function $[\![\varphi]\!] = 1$ if φ is true and 0 otherwise, and $d' = \operatorname{argmax}_d\{a^d_{t-1} {\neq} \perp\} + 1 - m_{\ell_t} - m_{g_t}$. Together, these probabilistic decisions generate the n unary branches and $n - 1$ binary branches of a parse tree in Chomsky normal form for an n-word sentence.

3 Isolating Content and Category Contributions

In order to examine the contribution of propositional content on the content-sensitive processing model, the model is modified to allow it to be trained to make lexical and grammatical decisions without conditioning on the predicate context vec-

[4] Examples of composition operators include using the predicate context of the left child as a modifier or an argument, as well as introducing or discharging filler-gap dependencies.

tors,

$$\mathsf{P}(m_{\ell_t}\ \mathbf{h}_{\ell_t} \mid q_{t-1}) = \underset{m_{\ell_t}\mathbf{h}_{\ell_t}}{\textsc{SoftMax}}(\ \mathrm{FF}_{\theta_\mathrm{L}}[\delta_d{}^\top, [\delta^\top_{c_{b^d_{t-1}}}, \mathbf{0}^\top]\,\mathbf{E}_\mathrm{L}]\ ) \quad (12)$$

$$\mathsf{P}(m_{g_t}\ o_{g_t} \mid q_{t-1}\ \ell_t\ w_t) = \underset{m_{g_t}o_{g_t}}{\textsc{SoftMax}}(\ \mathrm{FF}_{\theta_\mathrm{G}}[\delta_d{}^\top, [\delta^\top_{c_{b^d_{t-1}}}, \mathbf{0}^\top, \delta^\top_{c_{p_t}}, \mathbf{0}^\top]\,\mathbf{E}_\mathrm{G}]\ ) \quad (13)$$

where $\mathbf{0}$ is a vector of 0s.

Likewise, to examine the contribution of syntactic category information on the content-sensitive processing model, the model is modified to allow it to be trained to make decisions without conditioning on the syntactic category labels:

$$\mathsf{P}(m_{\ell_t}\ \mathbf{h}_{\ell_t} \mid q_{t-1}) = \underset{m_{\ell_t}\mathbf{h}_{\ell_t}}{\textsc{SoftMax}}(\ \mathrm{FF}_{\theta_\mathrm{L}}[\delta_d{}^\top, [\mathbf{0}^\top, \mathbf{h}^\top_{b^d_{t-1}}]\,\mathbf{E}_\mathrm{L}]\ ) \quad (14)$$

$$\mathsf{P}(m_{g_t}\ o_{g_t} \mid q_{t-1}\ \ell_t\ w_t) = \underset{m_{g_t}o_{g_t}}{\textsc{SoftMax}}(\ \mathrm{FF}_{\theta_\mathrm{G}}[\delta_d{}^\top, [\mathbf{0}^\top, \mathbf{h}^\top_{b^d_{t-1}}, \mathbf{0}^\top, \mathbf{h}^\top_{p_t}]\,\mathbf{E}_\mathrm{G}]\ ) \quad (15)$$

These two ablated models will respectively be referred to as the *content-* and *category-ablated* models in the following experiments.

4 Experiment 1: Linguistic Accuracy

4.1 In-domain Linguistic Accuracy

In order to assess the parsing performance of the content-sensitive processing model outlined in Section 2, a linguistic accuracy evaluation was conducted on the development set and test set (i.e. sections 22 and 23 respectively) of the Wall Street Journal (WSJ) corpus of the English Penn Treebank (Marcus et al., 1993). The performance of the content-sensitive processing model is compared to the incremental left-corner parser of van Schijndel et al. (2013), which is based on a PCFG with subcategorized syntactic categories from the Berkeley latent variable inducer (Petrov et al., 2006).

The content-sensitive processing model was trained on a generalized categorial grammar (Nguyen et al., 2012) reannotation of sections 02 to 21 of the WSJ corpus. Choices regarding hyperparameters were made based on the parsing performance on the development set of the WSJ corpus. In order to account for sensitivity to initial parameters, the average performance of the content-sensitive processing model trained using three different random seeds is reported. Likewise, the left-corner parser of van Schijndel et al. (2013) was trained on the same generalized categorial grammar reannotation of sections 02 to 21 of the WSJ corpus, using four iterations of the split-merge-smooth algorithm (Petrov et al., 2006). Both parsers used beam search decoding with a beam width of 5,000 to return the most likely sequence of parsing decisions.

Parsing model	WSJ22	WSJ23	NS
vS et al. (2013)	85.20	84.08	69.60
Full model (avg.)	84.60	82.45	71.64
Con-ablated (avg.)	81.64	79.86	69.88
Cat-ablated (avg.)	75.63	74.45	64.19

Table 1: Bracketing F1 scores on sentences with 40 or fewer words for the incremental parsing models. WSJ: Wall Street Journal, NS: Natural Stories.

The unlabeled WSJ bracketing F1 scores from both parsers are presented in the WSJ22 and WSJ23 columns of the vS et al. and Full model rows of Table 1.[5] The results show that the two parsers achieve comparable performance on WSJ22 and WSJ23, indicating that the current processing model is a reasonable model of syntactic parsing.

4.2 Cross-Domain Linguistic Accuracy

The two parsers were also evaluated on the Natural Stories Corpus (Futrell et al., 2018). This corpus consists of 10 naturalistic stories (10,245 tokens) adapted from existing texts such as fairy tales and short stories. As can be seen in the NS column of the vS et al. and Full model rows of Table 1, parsing accuracy on this corpus is substantially lower. This is likely due to the "deceptively naturalistic" nature of the Natural Stories Corpus; this corpus was designed to over-represent rare words and syntactic constructions, therefore representing a different "syntactic domain" from the WSJ corpus. Interestingly, the content-sensitive processing model seems to generalize better to the Natural Stories domain than the model based on the Berkeley latent

[5] It should be noted that the performance of the van Schijndel et al. (2013) parser here is lower than their reported performance because they trained their parser on data with PTB-style annotation, which has substantially fewer syntactic categories than the GCG annotation scheme.

variable inducer. This could be the result of the latent-variable subcategorized syntactic categories overfitting to the WSJ domain.

4.3 Linguistic Accuracy of Ablated Models

To determine the differential effect of propositional content and syntactic categories, models with each of the propositional content and syntactic category components ablated (i.e. the content- and category-ablated models) were evaluated against the full processing model.[6] As with the full model, the ablated models were trained using three different random seeds to account for sensitivity to initial parameters. The results in the Con-ablated and Cat-ablated rows of Table 1 show substantial contributions of both components to parsing accuracy in all domains. On Natural Stories, bootstrap significance tests revealed that seven out of nine (3×3) pairwise comparisons between the full model and the content-ablated model, and all nine pairwise comparisons between the full model and the category-ablated model were statistically significant at the $p < 0.05$ level, which are both highly significant overall by a binomial test.

5 Experiment 2: Self-paced Reading

In order to evaluate the contribution of propositional content and syntactic categories to predicting behavioral responses, surprisal predictors were calculated from the content-sensitive processing model and its two ablated versions, which are outlined in Section 3. Subsequently, linear mixed-effects models containing common baseline predictors and one or more surprisal predictors were fitted to self-paced reading times. Finally, a series of likelihood ratio tests (LRTs) were conducted in order to evaluate the contribution of the surprisal predictor from the full processing model to regression model fit.

5.1 Response Data

Experiments described in this paper used the Natural Stories Corpus (Futrell et al., 2018), which contains self-paced reading times from 181 subjects that read 10 naturalistic stories consisting of 10,245 tokens. The data were filtered to exclude observations corresponding to sentence-initial and sentence-final words, observations from subjects who answered fewer than four comprehension questions correctly, and observations with durations shorter than 100 ms or longer than 3000 ms. This resulted in a total of 768,584 observations, which were subsequently partitioned into an exploratory set of 383,906 observations and a held-out set of 384,678 observations. The partitioning allows model selection to be conducted on the exploratory set and a single hypothesis test to be conducted on the held-out set, thus eliminating the need for multiple trials correction. All observations were log-transformed prior to model fitting.

5.2 Predictors

The baseline predictors commonly included in all regression models are word length measured in characters, index of word position within each sentence, and 5-gram surprisal. The 5-gram surprisal predictor is calculated from a 5-gram language model estimated using the KenLM toolkit (Heafield et al., 2013) trained on the Gigaword 4 corpus (Parker et al., 2009).[7]

In addition to the baseline predictors, surprisal predictors were calculated from the full content-sensitive processing model, the content-ablated model, and the category-ablated model trained as part of Experiment 1 (*FullSurp*, *NoConSurp*, and *NoCatSurp*). To account for the time the brain takes to process and respond to linguistic input, it is standard practice in psycholinguistic modeling to include 'spillover' variants of predictors from preceding words (Rayner et al., 1983; Vasishth, 2006). However, as including multiple spillover variants of predictors leads to identifiability issues in mixed-effects modeling (Shain and Schuler, 2019), the *FullSurp*, *NoConSurp*, and *NoCatSurp* predictors were all spilled over by one position. Moreover, preliminary analysis showed that the surprisal predictors are highly collinear, which may result in identifiability issues for the regression model if included together as predictors. In order to mitigate this problem, the difference between the surprisal predictors from the ablated model and those from the full model (Δ*ConSurp*, Δ*CatSurp*) were also calculated as predictors that represent the contribution of the full model over an ablated model. All

[6] Source code is available at `https://github.com/modelblocks/modelblocks-release`.

[7] Although word frequency is also often included as a baseline predictor in the form of unigram surprisal, it was excluded in the current study in light of results showing no significant effect of unigram surprisal over and above 5-gram surprisal when predicting reading times from the Natural Stories Corpus (Shain, 2019).

predictors were centered and scaled prior to model fitting.

5.3 Likelihood Ratio Testing

Two sets of nested linear mixed-effects models were fitted to reading times in the held-out set using using `lme4` (Bates et al., 2015). The first set manipulated the contribution of propositional content by including Δ*ConSurp* in the full regression model over the base model that contains the baseline predictors and *NoConSurp*. Similarly, the second set manipulated the contribution of syntactic categories by including Δ*CatSurp* in the full regression model over a base model that contains the baseline predictors and *NoCatSurp*. All regression models included by-subject random slopes for all fixed effects and random intercepts for each word and subject-sentence interaction. Subsequently, a series of LRTs were conducted between nested regression models in order to assess the contribution of surprisal predictors from the full processing model to regression model fit. As there were three variants of each surprisal predictor, a total of nine (3×3) LRTs were performed for each ablated surprisal predictor.[8]

5.4 Results

The results show that the Δ*CatSurp* predictor made a statistically significant contribution to model fit over *NoCatSurp* in eight out of nine LRTs,[9] which is highly significant according to a binomial test ($p < 0.001$). In contrast, no significant contribution of Δ*ConSurp* over *NoConSurp* was observed, with none of the nine LRTs indicating significantly improved model fit.[10] This demonstrates that the full processing model captures the influence of propositional content and syntactic category information differentially, the latter of which contributed to predicting self-paced reading times.

6 Experiment 3: Eye-tracking Data

In order to examine whether the results observed in Experiment 2 generalize to other latency-based measures, linear-mixed effects models were fitted on the Dundee eye-tracking corpus (Kennedy et al., 2003). Following similar procedures to Experiment 2, a series of LRTs were conducted to test the contribution of propositional content and syntactic category information.

6.1 Procedures

The set of go-past durations from the Dundee Corpus (Kennedy et al., 2003) provided the response variable for the regression models. The Dundee Corpus contains gaze durations from 10 subjects that read 20 newspaper editorials consisting of 51,502 tokens. The data were filtered to exclude unfixated words, words following saccades longer than four words, and words at starts and ends of sentences, screens, documents, and lines. This resulted in the full set with a total of 195,296 observations, which were subsequently partitioned into an exploratory set of 97,391 observations and a held-out set of 97,905 observations. In the base regression models, word length in characters, index of word position in each sentence, and saccade length were included. Additionally, either *NoConSurp* or *NoCatSurp* spilled over by one position was included as a baseline predictor. Similarly to Experiment 2, the first set of LRTs examined the contribution of propositional content by including Δ*ConSurp*, and the second set of LRTs examined the contribution of syntactic category information by including Δ*CatSurp* in the full regression models.

6.2 Results

The results show that the Δ*ConSurp* predictor made a statistically significant contribution to model fit over *NoConSurp* in all nine LRTs.[11] A significant contribution of Δ*CatSurp* over *NoCatSurp* was observed as well, with three of the nine LRTs indicating significantly improved model fit ($p = .008$ according to a binomial test).[12] Interestingly, contrary to Experiment 2 that showed only a robust contribution of syntactic category information to

[8] Despite the risk of convergence issues, the LRTs were also replicated with full regression models that include raw *FullSurp* in addition to the baseline predictors and either *NoCatSurp* or *NoCatSurp*.

[9] Any LRT in which either the base or full regression model failed to converge was considered as a null result. Regression models in one LRT failed to converge. In the replication using raw *FullSurp*, regression models in five LRTs failed to converge. However, the remaining four LRTs were statistically significant, which is highly significant according to a binomial test ($p < 0.001$).

[10] Regression models in one LRT failed to converge. In the replication using raw *FullSurp*, regression models in five LRTs failed to converge, with the remaining four LRTs indicating non-significance. Additionally, removing 5-gram surprisal from the baseline did not change the pattern of significance.

[11] In the replication using raw *FullSurp*, regression models in five LRTs failed to converge. However, the remaining four LRTs were statistically significant, which is highly significant according to a binomial test ($p < 0.001$).

[12] Regression models in all LRTs converged. In the replication using raw *FullSurp*, regression models in five LRTs failed to converge, with two out of four remaining LRTs indicating statistical significance ($p = .071$ according to a binomial test).

predicting self-paced reading times, a strong influence of propositional content in predicting eye-gaze durations is observed. This corroborates the finding that the full processing model captures the distinct influence of propositional content and syntactic category information, the ablation of which results in qualitatively different predictions. In addition, this differential contribution of Δ*ConSurp* across self-paced reading and eye-tracking data suggests that these self-paced reading times and eye-gaze durations may capture different aspects of online processing difficulty.

7 Experiment 4: Filler-gap Constructions

Observing that surprisal from the full processing model did not contribute significantly to predicting broad-coverage self-paced reading times on top of its content-ablated counterpart in Experiment 2, we focus on filler-gap constructions,[13] in which information about the extracted object is thought to strongly influence the processing of the verb. In order to explore the extent to which integration costs associated with filler-gap constructions could be explained by the influence of propositional content, a series of LRTs were conducted to assess the contribution of surprisal from the full processing model to predicting reading times of object-extracted verbs.

7.1 Procedures

The subset of self-paced reading times from the Natural Stories Corpus corresponding to object-extracted verbs provided the response variable for the regression models. The object-extracted verbs were identified using a version of the Natural Stories Corpus that had been reannotated using a deep syntactic annotation scheme (Shain et al., 2018). Applying the same data exclusion criteria as Experiment 2 resulted in an exploratory set of 1,537 observations and a held-out set of 1,523 observations. As the number of data points for regression model fitting was substantially smaller in comparison to the full set used in Experiment 2, the regression models had to be simplified for reliable convergence. First, the 5-gram surprisal predictor was excluded as its effect estimate was not stable on the exploratory set. In addition, the random effects structure was simplified to include only the by-subject random intercept.

In the base regression models, word length in characters, index of word position within each sentence, and *NoConSurp* were fitted to the log-transformed reading times in the held-out set. The contribution of propositional content was incorporated by including *FullSurp* in the full regression models. *NoConSurp* and *FullSurp* were spilled over by one position, and all predictors were centered and scaled. The same three variants of each surprisal predictor were used, which resulted in a total of nine LRTs testing the contribution of *FullSurp*.

7.2 Results

The results showed that the *FullSurp* predictor made a statistically significant contribution to model fit over *NoConSurp* in all nine LRTs. The inclusion of *FullSurp* consistently improved model fit, indicating that integration costs associated with object-extracted filler-gap constructions can be partially explained by the influence of propositional content.

8 Conclusion

This paper presents a generative and incremental content-sensitive processing model which factors the contribution of propositional content and syntactic category information. This model can be cleanly ablated to calculate surprisal predictors that differentially isolate the influence of the two components. Subsequent experiments demonstrate the utility of both components in predicting human behavioral responses; the inclusion of propositional content resulted in significantly better fits to broad-coverage eye-gaze durations and self-paced reading times of object-extracted verbs. Additionally, the inclusion of syntactic category information significantly improved fits to both broad-coverage self-paced reading times and eye-gaze durations. Taken together, these results suggest a role for propositional content and syntactic category information in incremental sentence processing.

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

[13] For example, in the sentence *It was a match that the girl rubbed _ on the wall*, the extracted object *a match* has to be retrieved from memory and integrated to the transitive verb *rubbed*.

the authors and do not necessarily reflect the views of the National Science Foundation.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

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.

J. D. Bransford and J. J. Franks. 1971. The abstraction of linguistic ideas. *Cognitive Psychology*, 2:331–350.

Sarah Brown-Schmidt, Ellen Campana, and Michael K. Tanenhaus. 2002. Reference resolution in the wild: Online circumscription of referential domains in a natural interactive problem-solving task. In *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*, pages 148–153.

Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a distributional model of semantic complexity. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 12–22.

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.

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*, pages 10–18.

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 on Language Technologies*, pages 1–8.

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.

Robert J. Jarvella. 1971. Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior*, 10:409–416.

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

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

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.

Jeff Mitchell, Mirella Lapata, Vera Demberg, and Frank Keller. 2010. Syntactic and semantic factors in processing difficulty: An integrated measure. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 196–206.

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.

Ulrike Padó, Matthew W. Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science*, 33(5):794–838.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. English Gigaword LDC2009T13.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.

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.

Asad Sayeed, Stefan Fischer, and Vera Demberg. 2015. Vector-space calculation of semantic surprisal for predicting word pronunciation duration. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 763–773.

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.

Cory Shain. 2019. A large-scale study of the effects of word frequency and predictability in naturalistic reading. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Cory Shain, Marten van Schijndel, and William Schuler. 2018. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources (LREC 2018)*.

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

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.

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

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.

Shravan Vasishth. 2006. On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. In *Proceedings of the International Conference on Linguistic Evidence*, pages 96–100.