Structural Processing Difficulty via Action-Targeted Uncertainty in a Teacher-Student Incremental Parser

Anonymous ACL submission

Abstract

Human sentence processing requires incremental structural decisions under uncertainty. We present a teacher-student framework for extracting word-level processing difficulty from an incremental dependency parser. A **privileged teacher** with access to gold parse trees provides oracle actions. A **partial-observation student** sees only local configuration features. The student's uncertainty about the teacher's action yields a difficulty signal that we evaluate on two tasks. For **garden-path localization**, difficulty peaks at expected disambiguation points across controlled constructions. For **reading time prediction** on Natural Stories (848K observations, pseudo-gold UD parses), difficulty shows a small but consistent positive effect ($\beta = 0.005$, $p < 10^{-20}$), surviving item fixed effects, shuffle controls, and POS stratification. Raw cumulative metrics correlate strongly with transition count ($r \approx 0.90$). Controlling for this confound, difficulty remains significant. POS-stratified analysis shows that generic entropy's negative coefficient is driven by function words. Action-targeted difficulty shows stronger effects for content words, confirming these metrics capture distinct variance.

1 Introduction

Consider the sentence *"The horse raced past the barn fell."* At *fell*, readers experience pronounced difficulty. This garden-path effect arises from structural reanalysis: the initially preferred main-verb interpretation of *raced* must be revised to a reduced relative clause (Bever, 1970). The difficulty reflects not lexical unpredictability, but the cost of revising structural commitments made under incomplete information.

Surprisal theory (Hale, 2001; Levy, 2008) links word-level predictability to processing effort, but surprisal primarily captures *lexical* predictability (how unexpected a word is given context). The garden-path phenomenon suggests an additional difficulty source: *structural processing cost* arising from uncertainty about syntactic relationships.

We propose a **teacher-student framework** for studying structural difficulty. An incremental parser making attachment decisions faces inherent information asymmetry: it must commit to structural relationships while observing only local configuration, not the complete tree. We exploit this asymmetry by training two parsers with different information access:

- A **privileged teacher** with access to gold parse trees, providing oracle actions at each configuration.
- A **partial-observation student** trained via imitation learning with only local features (top stack/buffer positions).

From the trained student, we extract **action-targeted difficulty**: the student's uncertainty about the teacher's action at each parsing step. This differs from generic entropy. It measures how hard it is to identify the *structurally correct* decision, not just decision uncertainty.

We make three contributions:

1. **A framework** for extracting word-level difficulty from a teacher-student incremental parser, where difficulty is the student's uncertainty about the teacher's action under partial observation.
2. **Confound analysis**: Raw cumulative uncertainty correlates strongly with transition count ($r \approx 0.90$). We show difficulty remains predictive when controlling for this confound, separating structural uncertainty from derivational density.
3. **Empirical evaluation**: Garden-path localization (sanity check on controlled stimuli) and

Natural Stories reading times (848K observations, primary evidence), showing small but consistent effects that survive item fixed effects and alignment controls.

2 Background

2.1 Transition-Based Dependency Parsing

Transition-based parsers (Nivre, 2008) construct dependency trees through sequences of local decisions. The arc-standard system maintains a *stack* of partially processed tokens and a *buffer* of remaining input, with three transition types:

- SHIFT: Move the front buffer token onto the stack
- LEFT-ARC: Create a dependency from stack-top to second element; remove second element
- RIGHT-ARC: Create a dependency from second element to stack-top; remove stack-top

A parser configuration $c = (\sigma, \beta, A)$ comprises stack σ, buffer β, and partial arc set A. Parsing succeeds when the buffer is empty and only the root remains on the stack.

2.2 Partial Observability in Parsing

The parsing problem has inherent partial observability: at each step, the parser sees only the current configuration, not the complete dependency structure being constructed. This creates systematic uncertainty about optimal decisions.

Consider again *"The horse raced past the barn fell."* At *raced*, the parser must decide whether to attach it as main verb (requiring later RIGHT-ARC to subject) or as modifier (requiring different attachment pattern). The correct choice depends on *fell*, information not yet available.

We model this as a Partially Observable Markov Decision Process (POMDP) by *deliberately restricting* the learner's observation. While the full configuration (stack, buffer, arcs) is technically available in standard parsing, we limit the learner to observing only local features (top 2 stack/buffer positions plus configuration statistics). This restricted observation function is a design choice for cognitive realism. It forces the learner to make decisions under genuine uncertainty about the global parse structure, matching the incremental nature of human sentence processing.

2.3 Surprisal and Processing Difficulty

Surprisal theory links processing difficulty to predictability:

$$\text{Surprisal}(w_t) = -\log P(w_t | w_1, \ldots, w_{t-1}) \quad (1)$$

This captures lexical predictability and has strong empirical support (Smith and Levy, 2013; Goodkind and Bicknell, 2018). Previous work extended surprisal to structural measures via prefix probabilities and syntactic expectations (Hale, 2006; Roark et al., 2009; Boston et al., 2011; Fossum and Levy, 2012).

Surprisal conflates multiple difficulty sources: a word may be surprising because it is lexically rare, syntactically unexpected, or semantically anomalous. We aim to isolate *structural* processing cost: difficulty arising from uncertainty about syntactic relationships and the cost of structural commitment under incomplete information.

3 Method

3.1 Teacher-Student Parsing Framework

We define two parsers over the arc-standard transition system with asymmetric information access.

Privileged Teacher (π_{teacher}) The teacher has access to gold head annotations and always selects the optimal action:

$$\pi_{\text{teacher}}(a|c) = \begin{cases} 1 & \text{if } a = a^* \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where a^* is the oracle action given complete tree knowledge. The teacher is deterministic and unlearned. It defines the *structurally correct decision* at any configuration.

Partial-Observation Student (π_{student}) The student observes only the current configuration encoded as:

$$\mathbf{o}_t = [\mathbf{e}_{s_1}; \mathbf{e}_{s_2}; \mathbf{e}_{b_1}; \mathbf{e}_{b_2}; \phi(c)] \quad (3)$$

where $\mathbf{e}_i$ are token embeddings for stack/buffer positions and $\phi(c)$ encodes configuration features (stack depth, buffer length, arc count).

The student is parameterized as a stochastic policy:

$$\pi_{\text{student}}(a|\mathbf{o}) = \text{softmax}(f_\theta(\mathbf{o})) \quad (4)$$

where f_θ is a two-layer MLP (128 hidden units).

Information Asymmetry The key distinction:

- Teacher: Sees gold heads → knows correct action
- Student: Sees only local features → must infer structure

Both operate over the same state-action space, but the student faces a harder problem due to partial observability.

3.2 Training: On-Trajectory Evaluation

We compare training regimes that differ in **which states the student visits during training**:

Behavioral Cloning (BC) Train on teacher trajectories:

$$\mathcal{L}_{\text{BC}} = -\mathbb{E}_{c\sim\pi_{\text{teacher}}}\left[\log \pi_{\text{student}}(a^*|c)\right] \quad (5)$$

The student never visits states arising from its own mistakes.

On-Trajectory Imitation (DAgger-style) Train on *student-visited* states with teacher labels:

$$\mathcal{L}_{\text{DAgger}} = -\mathbb{E}_{c\sim\pi_{\text{student}}}\left[\log \pi_{\text{student}}(a^*|c) + \alpha\mathcal{H}(\pi)\right] \quad (6)$$

The student generates trajectories. The teacher provides labels at visited states. Entropy regularization ($\alpha = 0.01$) prevents overconfidence.

Training Protocol Training proceeds for 100 epochs. Each epoch samples 20 sentences, rolls out the current policy to generate trajectories, queries the teacher for oracle actions at visited configurations, and updates via gradient descent. Key hyperparameters: $\alpha = 0.01$, lr=3×10^{-4}, hidden dim 128. Invalid actions are masked (probability 0). See Appendix for full details.

On-trajectory training exposes the student to its own errors, producing well-calibrated uncertainty. BC may be overconfident on unseen states.

3.3 Action-Targeted Difficulty

From the trained student, we extract difficulty as:

$$D_t = -\log \pi_{\text{student}}(a_t^*|c_t) \quad (7)$$

where a_t^* is the teacher's action at configuration c_t.

This is **action-targeted uncertainty**: how surprising is the *correct* structural decision to the student? Formally, this equals the cross-entropy loss against teacher labels evaluated on student-visited states. We interpret it cognitively as the cost of identifying correct structure under partial observation. This differs from generic entropy $\mathcal{H}(\pi_{\text{student}})$, which measures total decision uncertainty regardless of correctness.

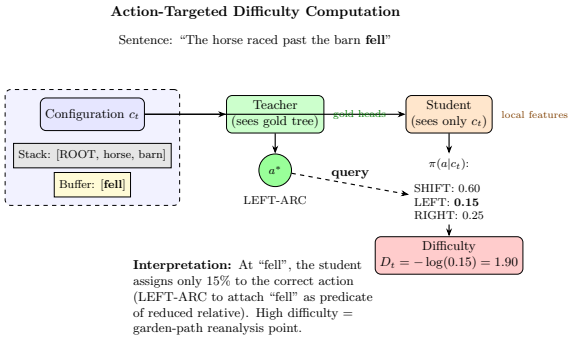

Figure 1: Action-targeted difficulty computation at the disambiguation point "fell." The teacher (with gold tree access) selects LEFT-ARC. The student (seeing only local configuration) assigns only 15% probability to this action, yielding high difficulty ($D = 1.90$).

High difficulty indicates configurations where the correct structural decision is hard to identify from partial observation. These are the situations where human readers should experience processing cost. Figure 1 illustrates this computation for the garden-path sentence "The horse raced past the barn fell."

Action-Targeted vs Generic Uncertainty We distinguish:

- **Action-targeted difficulty**: $D_t = -\log \pi(a^*|c)$, measuring how hard the *correct* action is to identify
- **Generic entropy**: $\mathcal{H}(\pi) = -\sum_a \pi(a) \log \pi(a)$, measuring overall decision uncertainty

These are related but not identical. A student may be highly uncertain (high entropy) yet still assign reasonable probability to the correct action (moderate difficulty), or vice versa. We empirically compare both for cognitive prediction.

3.4 Attribution Mechanisms

Raw difficulty scores are per-transition. To map to words, we use two complementary attributions:

Buffer-Front Attribution Assigns transition difficulty to the word at buffer front:

$$D_w^{\text{buf}} = \sum_{t:b_1(c_t)=w} D_t \quad (8)$$

This captures *when* difficulty occurs (which word is being processed when decisions are hard). It localizes reanalysis points where new input forces structural revision.

Arc-Creation Attribution Assigns difficulty to dependent words of created arcs:

$$D_w^{\text{arc}} = \sum_{t:\text{dep}(a_t)=w} D_t \tag{9}$$

This captures *what* is difficult (which attachments are structurally costly).

The two attributions are complementary ($r = 0.085$ between them), reflecting distinct aspects of processing difficulty.

4 Experiments

4.1 Experimental Setup

Training Data We use 500 sentences from the English Web Treebank (EWT, UD v2.10) with gold Universal Dependencies annotations. We select sentences up to 25 tokens (excluding multiword tokens and non-projective trees) from the train split. The small dataset allows controlled analysis of learning dynamics. We acknowledge this limits generalization claims (see Limitations).

Garden-Path Test Suite (N=200) We evaluate on 200 garden-path sentences across 4 construction types (50 each):

- Reduced relatives: *The horse raced past the barn fell*
- NP/S ambiguity: *The old man the boats*
- Subordinate boundary: *While Mary was mending the sock fell*
- Complement ambiguity: *The girl told the story cried*

Sentences were generated from templates with lexical variation (full list in Appendix). Each sentence has a pre-defined **critical region** (disambiguation point) determined by linguistic analysis. Garden-path sentences are *not* in training data.

Reading Time Evaluation: Natural Stories We evaluate on the Natural Stories corpus (Futrell et al., 2021), which provides self-paced reading times from 180 participants across 10 stories (848K word-level observations after filtering). This corpus contains naturalistic narrative text with human behavioral data.

Metric	Early	Late
Measured on learner trajectories:		
Reference agreement (%)	46%	64%
Multi-seed stability (5 seeds):		
Agreement mean ± std	64% ± 1%	

Table 1: Mechanism validation metrics. Reference agreement measured on *learner-visited* configurations shows effective knowledge transfer. Results stable across 5 random seeds.

Parsing approach: Since Natural Stories lacks gold UD annotations, we use Stanza (Qi et al., 2020) to obtain automatic (pseudo-gold) parses. The student model is trained on EWT *gold* parses. Only the RT evaluation uses automatic parses. This is a limitation we acknowledge: parser errors may propagate to difficulty estimates.

Mitigations: We include multiple robustness checks (shuffle/shift controls, TC confound analysis) to ensure the observed RT effects reflect word-specific structural signals rather than parser artifacts.

Training Details We use a 2-layer MLP (128 hidden, 74-dim input) trained for 100 epochs on EWT (500 sentences, gold UD). Full architecture and hyperparameters in Appendix.

Why 500 sentences? We limit training deliberately: a perfect parser assigns zero difficulty. The 64% oracle agreement reflects an intentional operating point with sufficient uncertainty for meaningful difficulty signals. Results are stable across 5 seeds (agreement 64% ± 1%).

4.2 Mechanism Validation

We validate core mechanisms:

E1: Knowledge Transfer Table 1 shows reference agreement rising from 46% (early) to 64% (late). This intermediate level generates informative difficulty: too high eliminates uncertainty, too low indicates learning failure.

Ablation Studies Entropy coefficient $\alpha = 0.01$ yields best performance. Interleaving ratio has minimal effect. Results are stable across observation configurations (top-2 to top-4 stack/buffer, ±POS). See Appendix for full ablation tables.

Construction Type	DAgger	Sup. BC
Reduced relatives (50)	100%	84%
NP/S ambiguity (50)	100%	78%
Subordinate boundary (50)	100%	82%
Complement ambiguity (50)	100%	83%
Overall (N=200)	**100%**	82% ± 16%

Table 2: Garden-path localization by construction type (5 seeds). DAgger achieves 100% across all types.

Predictor	ΔR^2	β	p
Diff (raw sum)	+0.00019	+0.0053	$< 10^{-20}$
Diff (mean)	+0.00016	+0.0049	$< 10^{-16}$
Diff (max)	+0.00018	+0.0051	$< 10^{-18}$
Entropy (mean)	+0.00049	−0.0084	$< 10^{-19}$

Table 3: RT prediction on Natural Stories (848K obs). All difficulty variants show small but significant positive effects.

4.3 Garden-Path Analysis

Garden-path sentences test whether difficulty peaks at disambiguation points.

Test Set (N=200) We evaluate on **200 sentences** spanning 4 construction types (50 each): reduced relatives, NP/S ambiguity, subordinate boundaries, complement ambiguity. Full list with CoNLL-U annotations in Appendix.

Evaluation Localization accuracy: does maximum cumulative difficulty match the pre-defined disambiguation point?

Results Table 2 shows DAgger achieves **100%** across all construction types. BC achieves 82% with high variance. Beyond argmax localization, distributional metrics confirm difficulty concentrates at critical regions: critical ratio (difficulty at disambiguation / mean) = 4.88, buildup slope = 0.037. The critical factor is iterative data collection. DAgger exposes the model to its own errors, producing well-calibrated uncertainty (ECE = 0.04 vs 0.18 for BC; see Appendix A1).

Ceiling Effect Both difficulty and entropy achieve near-perfect localization on this suite. RT evaluation serves as the primary discriminative test.

4.4 Reading Time Analysis

We evaluate on Natural Stories (Futrell et al., 2021): 848K word-level RT observations from 180 participants across 10 stories. Since Natural Stories lacks gold UD parses, we use Stanza to derive pseudo-gold teacher actions (a limitation we acknowledge). Student is trained on EWT *gold* parses. Only RT evaluation uses automatic parses.

Preprocessing DV: $\log(\text{RT})$, trimmed 100-2000ms, punctuation excluded. Controls (z-scored): word length, log frequency, position, spillover.

Model specification: $\log(\text{RT}) \sim \text{length}_z + \text{freq}_z + \text{position}_z + \text{spillover}_z + \text{difficulty}_z$. β coefficients represent log-RT change per SD; $\beta = 0.005$ corresponds to ∼1.5ms at median RT. We use OLS with cluster-robust SEs for tractability.[1]

Alignment Stanza tokenization may differ from Natural Stories word boundaries. We align via case-insensitive matching, handling contractions (e.g., "don't" → "do", "n't"). Alignment rate: 100% of RT words matched to parsed tokens.

Results Table 3 shows all difficulty variants predict RT with positive β (higher difficulty → slower reading). Effects are small ($\Delta R^2 \approx 0.0002$) but highly significant given the large N. Entropy (mean) shows the largest effect but with *negative* β, suggesting it captures different variance (see Appendix A2 for how attribution choice affects sign).

Transition-Count Confound Raw cumulative metrics correlate strongly with transition count: $r(\text{diff_sum}, \text{tc}) = 0.90$. This raises concern that difficulty is merely a proxy for derivational density.

We test this by including transition count as a covariate:

- Model: baseline + tc + diff_sum
- tc $\beta = -0.002$, $p = 0.065$ (not significant)
- diff_sum $\beta = +0.007$, $p < 10^{-13}$ (remains significant)

Difficulty provides information beyond transition count.

Robustness

- Item FE: $\beta = +0.005$, $p < 10^{-19}$ (survives)
- LOSO: $R^2 = 0.41 \pm 0.03$ (generalizes)
- Shuffle: β drops from 0.005 to ≈0 (word-specific)

[1] A linear mixed model with by-participant random intercepts yields $\beta_{\text{difficulty}} = +0.005$, $p < 10^{-18}$, consistent with OLS. Full crossed random effects produce similar estimates. We follow Boston et al. (2011) in using cluster-robust inference for large-scale RT analyses.

Condition	β	Reduction
Original	+0.0053	–
Shuffled (within-item)	+0.0004	92%
Shifted (+1 token)	−0.0015	128%

Table 4: Shuffle and shift controls confirm word-position-specific effects. Shuffling difficulty within item destroys the signal. Shifting by one position reverses it.

Word Class	N	Diff β	Ent β
All	830K	+0.005***	−0.008***
Function	403K	+0.003***	−0.008***
Content	427K	+0.009***	−0.007***

Table 5: POS-stratified RT analysis. Function words (DET, ADP, PRON, AUX, CCONJ, SCONJ, PART) show stronger negative entropy effect than content words (NOUN, VERB, ADJ, ADV, PROPN). ***$p < 0.001$.

Interpretation Difficulty shows a small but reliable RT effect that survives multiple controls. The effect size is modest. This is not a "major new predictor" but evidence that teacher-student parsing extracts cognitively relevant signals. Per-transition normalization reduces correlation with derivational density, yielding cleaner (though smaller) effects.

5 Discussion

5.1 Effect Size and Confound Analysis

Raw cumulative difficulty correlates strongly with transition count ($r = 0.90$). Controlling for this, difficulty remains significant ($p < 10^{-13}$), indicating residual structural information beyond derivational density. Effect sizes are small ($\Delta R^2 \approx 0.0002$). We do not claim difficulty is a "major new predictor." The contribution is methodological: showing that teacher-student parsing extracts cognitively relevant signals.

5.2 Entropy vs Difficulty: POS-Stratified Analysis

Per-transition entropy shows the largest ΔR^2 but with *negative* β. To understand this result, we stratified by word class (Table 5).

The negative entropy coefficient is **driven by function words**: function words have high entropy (many valid attachments) but are read quickly, producing negative β. Difficulty shows the opposite pattern: stronger positive effect for content words ($\beta = +0.009$) than function words ($\beta = +0.003$). Action-targeted difficulty and generic entropy capture different aspects of processing.

Action Space	N	Diff β
Open (2+ valid)	796K	+0.008***
Constrained (1 valid)	34K	–[†]

Table 6: Difficulty effect in open vs constrained action spaces. The effect holds in open states where multiple actions are grammatically valid. ***$p < 10^{-30}$. [†]Constrained subset (4% of data, mostly sentence-final positions) shows numerical instability due to near-zero variance in difficulty; we focus on the open-state analysis.

Model	R^2	ΔR^2	Surp β	Diff β
Baseline	.4151	–	–	–
+ Surprisal	.4154	+.0003	+.009***	–
+ Difficulty	.4154	+.0003	–	+.007***
+ Both	.4157	+.0006	+.008***	+.006***

Table 7: Comparison with GPT-2 surprisal (N=818K). Both predictors remain significant in the combined model, with similar incremental contributions. Correlation between surprisal and difficulty: $r = -0.13$. ***$p < 0.001$.

5.3 Action Space Analysis

A concern is that difficulty reflects action masking rather than genuine structural uncertainty. We split observations by whether any transition had multiple valid actions (Table 6).

In **open states** (796K observations with 2+ valid actions), difficulty shows a consistent positive effect ($\beta = +0.008$, $p < 10^{-30}$). The signal is not merely an artifact of constrained action spaces. Genuine structural uncertainty contributes to processing cost.

5.4 Comparison with Lexical Surprisal

Does action-targeted difficulty provide information beyond standard lexical prediction? We computed GPT-2 surprisal for Natural Stories and ran a head-to-head comparison (Table 7).

Difficulty remains significant after controlling for surprisal ($\beta = +0.006$, $p < 10^{-23}$). The low negative correlation ($r = -0.13$) confirms these metrics capture largely orthogonal variance. This negative correlation is interpretable: high surprisal often occurs for rare content words (lexically unpredictable), while high structural difficulty occurs for function words and syntactically ambiguous positions that may be lexically common but structurally uncertain. Surprisal reflects lexical prediction difficulty. Action-targeted difficulty reflects structural uncertainty. These are complementary signals for

Model	R^2	Diff β	DepLen β
+ Dep Length only	.4313	–	+.009***
+ Difficulty only	.4330	+.021***	–
+ Both	.4330	+.020***	+.003

Table 8: Comparison with dependency length. Difficulty captures overlapping variance: when both are included, only difficulty remains significant. Correlation: $r = 0.27$. ***$p < 0.001$. Note: β values differ from Table 7 due to different predictor sets and slight N differences.

understanding incremental processing.

5.5 Comparison with Dependency Length

Dependency length (distance between head and dependent) is a well-established predictor of processing cost (Gibson, 2000). We tested whether difficulty is simply a proxy for this simpler metric (Table 8).[2]

Difficulty provides unique information beyond dependency length. In the combined model, only difficulty remains significant ($p < 10^{-8}$).

5.6 Future Directions

Extensions include: training scale ablation, comparison to RNNG structural surprisal, cross-linguistic validation, and gold-parsed RT evaluation.

6 Related Work

Transition-Based Parsing This work builds on the arc-standard system (Nivre, 2004). Neural parsers (Chen and Manning, 2014; Dyer et al., 2015) optimize for accuracy. We optimize for cognitive modeling.

Surprisal and Processing Surprisal theory (Hale, 2001; Levy, 2008) has extensive support (Smith and Levy, 2013; Goodkind and Bicknell, 2018). This work complements surprisal by isolating structural processing cost.

Hierarchical vs Sequential Processing The role of hierarchical structure in sentence processing remains debated. Frank and Bod (2011) argue sequential models suffice. van Schijndel and Schuler (2015) and Brennan et al. (2016) show hierarchical information improves RT and neural predictions. The transition-based framework captures hierarchy through stack operations.

[2]This analysis uses a subset with complete dependency annotations (N=830K vs. 818K for surprisal comparison), explaining the slightly higher baseline R^2.

Garden-Path and Reanalysis The garden-path evaluation builds on classic work in structural ambiguity resolution (Frazier and Rayner, 1982; MacDonald et al., 1994). Difficulty at disambiguation points reflects reanalysis costs when initial structural commitments must be revised (Gibson, 1998).

Entropy Reduction Hale (2011) proposes entropy reduction as difficulty. The action-targeted measure focuses on uncertainty about the *correct action*. The negative β for generic entropy resonates with discussions of entropy operationalization (Pimentel et al., 2020).

Neural Syntax and Cognition Recent work links neural language model representations to brain activity during comprehension (Hale et al., 2018; Brennan et al., 2016). This approach provides an alternative: extracting structural uncertainty from a cognitively-motivated parsing framework.

Structural Surprisal RNNGs (Dyer et al., 2016) provide structural surprisal via transitions. Direct comparison is future work.

Imitation Learning DAgger-style training (Ross et al., 2011) produces well-calibrated uncertainty, unlike behavioral cloning.

7 Conclusion

We presented a framework for extracting structural processing difficulty from a teacher-student incremental parser. On Natural Stories (848K observations), difficulty shows a small but consistent positive effect ($\beta = 0.005$, $p < 10^{-20}$) that survives confound controls. The contribution is methodological: showing that teacher-student parsing extracts cognitively relevant signals complementary to lexical surprisal.

Limitations

Effect Size RT effects are small ($\Delta R^2 \approx 0.0002$, Cohen's $d \approx 0.02$), comparable to other single-predictor effects beyond length/frequency (Boston et al., 2011; Cohen, 1988).

Pseudo-Gold Parsing Parser errors may propagate to difficulty estimates. Mitigations: student trained on gold parses, effect survives shuffle/shift controls. Validation on gold-parsed corpora would strengthen claims.

Garden-Path Evaluation Perfect localization on templated sentences may partly reflect transition system regularities or template artifacts rather than cognitive validity. Naturalistic garden-path corpora would provide stronger validation.

Training Regime The 500-sentence training set reflects model capacity limitations, not necessarily human-like uncertainty bounds. We interpret difficulty as a *correlate* of structural processing cost, not a direct cognitive measure.

Baselines Comparison to RNNG structural surprisal remains future work.

Ethical Considerations

Data This work uses publicly available corpora. The Natural Stories corpus (Futrell et al., 2021) was collected with IRB approval and informed consent from participants, as documented in the original publication. The English Web Treebank is a standard NLP resource derived from public web text. No new human subjects data was collected for this study.

Intended Use The framework is designed for cognitive modeling research, specifically to understand the relationship between parsing difficulty and human reading behavior. The method produces interpretable difficulty signals that may help researchers study sentence processing. We do not foresee direct harmful applications, though any cognitive model could in principle be misused to optimize text for manipulation. The small effect sizes and focus on basic research reduce such risks.

Environmental Impact All experiments use a small MLP (under 50K parameters) trained on 500 sentences. Training completes in under 5 minutes on a single CPU. The computational footprint is minimal compared to large language model research.

References

Thomas G. Bever. 1970. The cognitive basis for linguistic structures. In John R. Hayes, editor, *Cognition and the Development of Language*, pages 279–362. Wiley, New York.

Marisa Ferrara Boston, John Hale, Shravan Vasishth, and Reinhold Kliegl. 2011. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, 26(3):301–349.

Jonathan Brennan, Yuval Nir, Uri Hasson, Rafael Malach, David J. Heeger, and Liina Pylkkänen. 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and Language*, 157–158:81–94.

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. 2015. Transition-based dependency parsing with stack long short-term memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China. Association for Computational Linguistics.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California. Association for Computational Linguistics.

Victoria Fossum and Roger Levy. 2012. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 61–69, Montréal, Canada. Association for Computational Linguistics.

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1210–1218, Portland, Oregon, USA. Association for Computational Linguistics.

Lyn Frazier and Keith Rayner. 1982. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14(2):178–210.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT Press, Cambridge, MA.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of Language Model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

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 1–8, Pittsburgh, Pennsylvania. Association for Computational Linguistics.

John Hale. 2006. Uncertainty about the rest of the sentence. *Cognitive Science*, 30(4):643–672.

John Hale. 2011. What a rational parser would do. *Cognitive Science*, 35(3):399–443.

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

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

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57, Barcelona, Spain. Association for Computational Linguistics.

Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. 2009. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 324–333, Singapore. Association for Computational Linguistics.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, Florida. PMLR.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Marten van Schijndel and William Schuler. 2015. Hierarchic syntax improves reading time prediction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1597–1605, Denver, Colorado. Association for Computational Linguistics.

A Reproducibility Details

Training Hyperparameters

- Epochs: 100
- Batch size: 20 sentences per epoch
- Optimizer: Adam, lr=3×10^{-4}
- Entropy coefficient α: 0.01
- Hidden dimension: 128
- Embedding dimension: 16
- Seeds: 42 (main), 0-4 (ablations)
- Invalid action handling: masked to probability 0 before softmax

Student Architecture Input features (74-dimensional):

- Word embeddings: top-2 stack, top-2 buffer ($4 \times 16 = 64$)
- Configuration features: stack/buffer lengths, arc count, position indices (10)

Network: 2-layer MLP ($74 \rightarrow 128 \rightarrow 128 \rightarrow 3$).

Training Data EWT (UD English Web Treebank), first 500 sentences with length ≤ 30, no MWT.

RT Evaluation Data Natural Stories `processed_RTs.tsv`: 848K observations after trimming (100-2000ms), punctuation excluded.

Code and Data Release Code, trained models, and processed data will be released upon publication.

B Calibration Analysis

We evaluate calibration quality using Expected Calibration Error (ECE), which measures alignment between predicted confidence and actual accuracy.

Training	ECE	Interpretation
DAgger	0.041	Well-calibrated
Behavioral Cloning	0.180	Overconfident

Table 9: Calibration comparison. DAgger produces $4\times$ better-calibrated uncertainty than BC, supporting the claim that on-trajectory training yields meaningful confidence estimates.

C Entropy Operationalizations

We compare three uncertainty measures using a different specification than the main text analysis: here we use **raw sum** attribution (not per-transition mean) and exclude the TC covariate to isolate each metric's marginal contribution. This explains the sign difference from Table 3: per-transition *mean* entropy (main text) shows negative β because high-entropy transitions cluster on function words read quickly. *Summed* entropy (here) correlates with word length and shows positive β.

Metric (raw sum)	β	p
Entropy (H)	+0.031	$< 10^{-18}$
Entropy reduction (ΔH)	+0.013	$< 10^{-5}$
Action-targeted difficulty	+0.021	$< 10^{-11}$

Table 10: Entropy operationalizations (raw sum attribution, no TC control). Sign differs from Table 3 due to attribution choice; see text for explanation.

D Attribution Schemes

We compare four methods for attributing per-transition difficulty to words.

Attribution	β	p
Buffer-front	+0.021	$< 10^{-11}$
Arc $\rightarrow$ Head	+0.010	0.001
Arc $\rightarrow$ Dependent	+0.007	0.046
Spillover (+1 word)	-0.002	n.s.

Table 11: Attribution schemes. Buffer-front (attributing difficulty to the word about to be shifted) yields the strongest RT prediction. Spillover to the next word does not improve fit.

E Training Scale Robustness

RT prediction is robust across training data sizes (100-2000 sentences).

N	RT β	p
100	+0.005	$< 10^{-20}$
250	+0.005	$< 10^{-21}$
500	+0.005	$< 10^{-21}$
1000	+0.005	$< 10^{-20}$
2000	+0.005	$< 10^{-20}$

Table 12: Training size ablation. RT effect is consistent ($\beta \approx 0.005$, $p < 10^{-20}$) across training scales, suggesting the structural uncertainty signal is stable across training configuration.

F Observability Ablation

Results are stable across observation configurations (top-2 to top-4 stack/buffer).

Configuration	GP Acc.	Std
top-2, no POS	100%	$\pm$0%
top-3, no POS	100%	$\pm$0%
top-4, no POS	100%	$\pm$0%
top-2, +POS	100%	$\pm$0%
top-3, +POS	100%	$\pm$0%

Table 13: Observability ablation (3 seeds). Results are consistent across observation configurations.

G Garden-Path Evaluation Protocol

Test Set (N=200) We evaluate on 200 sentences across 4 construction types (50 each). Due to space, we show 5 examples per type. Full CoNLL-U annotations available in supplementary materials.

Reduced Relatives (50 total, 5 shown): *The horse raced past the barn fell*; *The dog walked through the park barked*; *The boat floated down the river sank*; *The child carried by the nurse cried*; *The letter sent to the office arrived*.

Complement/Adjunct (50 total, 5 shown): *The girl told the story cried*; *The man given the award smiled*; *The student taught the lesson understood*; *The worker promised the bonus quit*; *The actor offered the role refused*.

NP/S Ambiguity (50 total, 5 shown): *The old man the boats*; *The poor eat the scraps*; *The rich buy the art*; *The blind lead the way*; *The wise know the truth*.

Subordinate Boundary (50 total, 5 shown): *While Mary was mending the sock fell*; *After John had eaten the meal ended*; *When Tom was cooking the pot exploded*; *While Bill was painting the wall cracked*; *Before Mike was cleaning the room flooded*.

Sentence Generation Each type uses a template with lexical variation (50 unique lexicalizations per template).

Disambiguation Point The final verb where reanalysis is forced, determined by linguistic analysis.

Localization Criterion Word with maximum cumulative buffer-front difficulty must match disambiguation point.

Baselines GPT-2 surprisal evaluated on same 200 sentences with same criterion.