Towards a Bayesian hierarchical model of lexical processing

Cassandra L. Jacobs
Department of Linguistics
State University of New York at Buffalo
Buffalo, NY, USA
cxjacobs@buffalo.edu

Morgan Grobol
MoDyCo
Université Paris Nanterre
Nanterre, France
lgrobol@parisnanterre.fr

Abstract

In cases of pervasive uncertainty, cognitive systems benefit from heuristics or committing to more general hypotheses. Here we present a hierarchical cognitive model of lexical processing that synthesizes advances in early rational cognitive models with modern-day neural architectures. Probabilities of higher-order categories derived from vector representations extracted from the middle layers of an encoder language model have predictive power in accounting for several reading measures for both predicted and unpredicted words and influence even early first fixation duration behavior. The results suggest that lexical processing can take place within a latent, but nevertheless discrete, space in cases of uncertainty.

1 Introduction

Skilled readers are able to quickly and accurately leverage real-world and linguistic knowledge to understand texts. Lexical and syntactic factors strongly influence the speed and accuracy of sentence processing (Levy, 2008; Brennan and Hale, 2019). In addition to factors such as lexical frequency, word length, and syntactic processes, there is also lexico-semantic structure in language as it unfolds in time. Such higher-order abstractions are posited to be advantageous for any cognitive system to track (Kwisthout et al., 2017), such as anticipating or quickly integrating the semantic category that a word belongs to into one's understanding of a sentence (Federmeier and Kutas, 1999; Roland et al., 2012).

Despite clear macro structure in the predictability of individual words (e.g., the mention of *couch* versus *sofa*), it has been less clear how semantic structure influences reading times. The present paper aims to account for such macro structure and better understand how the semantic predictability of words shapes reading behavior. We quantify this structure using Bayesian Gaussian mixture models trained over embeddings of cloze responses, which we apply to a standardized dataset of reading times with associated predictability norms. Then, we obtain "semantic" estimates using cluster probabilities derived from the above Bayesian Gaussian mixture modeling process, which we incorporate into models of "early" predictive processing measures of reading times and a later, a more "semantic" reading time measure.

2 A hierarchical model of reading times

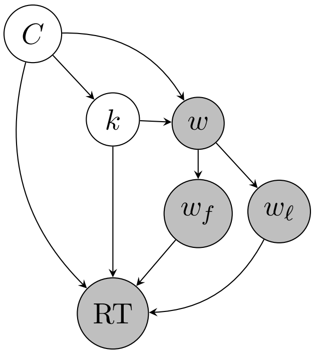

Figure 1: Our hierarchical model relating linguistic variables to reading time. C is the context (potentially including extralinguistic components), k a semantic cluster, w the observed word and w_f and w_ℓ its frequency and length respectively, RT is a measure of reading time such as fixation go-pass duration (FGPD) ; arrows denote random variable dependencies ; shaded variables are observable, unshaded ones are latent/unobserved.

Reading times partially reflect the contextual or conditional probability of a linguistic event (such as a word or syntactic structure), in that low-probability events are correlated with longer reading times (e.g., de Varda et al., 2024; Shain et al., 2024). Since the advent of neural language models, researchers have taken a strongly lexical approach to these analyses, though there is a growing appreciation that estimates of a word or syntactic structure's probability in context do not tell the full story

Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 165–171
May 3, 2025 ©2025 Association for Computational Linguistics

about what makes written language easier or harder to read (Gruteke Klein et al., 2024). Reliance on lexical estimates of predictability may over-estimate the uncertainty of the linguistic future (Kwisthout et al., 2017) and thus under-estimate the importance of higher-order prediction and over-estimate processing difficulty (Ozaki et al., 2024). For example, a comprehender may make more general linguistic predictions (e.g., the CAT concept) or more specific ones (e.g., "oriental longhair"; Degen et al., 2020; Roland et al., 2012; Federmeier and Kutas, 1999). We present a clustering method that estimates these higher-order, lexico-semantic hypothesis spaces K, which we demonstrate provides additional predictive power in explaining reading times beyond word-level information alone. Quantifying higher-order structure ties the modern lexical approach to hierarchical Bayesian processing models and early surprisal approaches (e.g., Levy, 2008). We outline such a model in Figure 1.

3 Applying cloze data to the study of rational language comprehension

Accounting for hierarchical processing in reading requires suitable resources for estimating these hierarchical categories. While the contemporary approach to estimating lexical predictability overwhelmingly relies on language model surprisal, human beings and language models do not align make the same prediction (Smith and Levy, 2011). We aggregate cloze production data (Taylor, 1953) into quasi-semantic clusters using Bayesian Gaussian mixture modeling. We focus on the Provo Corpus, in which participants read an incomplete text and guessed the identity of the next word for each word in a sentence ("serial cloze"; Luke and Christianson, 2016, 2018; Lowder et al., 2018), which is one of the only parallel datasets with reading time and cloze production statistics.

In the Provo corpus, college-age American English speaking participants incrementally guessed the identity of each non-initial word w_i for every preamble $p = w_1 \dots w_{i-1}$ in order. The resulting cloze corpus consists of 41 236 unique continuations across 2398 unique preambles, from a collection of 55 short, multi-sentence web texts. Cloze probabilities are defined as follows:

$$P(\text{word} \mid p) = \frac{\text{count}(\text{word} \cap p)}{\text{count}(p)}$$

The Provo corpus also includes reading time data for each of these texts from 84 participants from the same population. We focus on two reading time measures for their relative cognitive transparency and to minimize the number of statistical comparisons (Von der Malsburg and Angele, 2017): first fixation duration (FFD) and first go-past duration (FGPD). FFD is often conceptualized as reflecting early-stage visual processes while FGPD is thought to reflect additional time for semantic integration. Both FFD and FGPD measures are sensitive to quantitative indicators of lexical and syntactic predictability (Staub, 2015).

4 Clustering model

We model semantic predictability using a Dirichlet process mixture (Antoniak, 1974) of Gaussians trained with variational inference (Blei and Jordan, 2006) on the set $E \subset \mathbb{R}^d$ of token embeddings of participants' best guesses in a serial cloze task. More precisely: we model E as a sample drawn from a weighted sum of d-dimensional multivariate Gaussian variables k_i (components) $\sum_i \pi_i k_i$. This can be reformulated as a two-step process of first sampling one component k from a set K, then sampling an embedding from k. If we identify each component with the set of the embeddings it generated, K can then be seen as a clustering of E, which can be approximated by estimating a probability distribution over components $P(e \in k)$ for each embedding $e \in E$ and assigning e to its maximum-likelihood component $\text{argmax}_{k \in K} P(e \in k)$.

The number of unique word forms in a cluster (of approximately 36 000 completions) ranged from 1 to 1534, showing substantial skew with a mean/median/mode of 128/5/2 words per cluster. Clustering results in a drop in uncertainty during naturalistic reading that reduces the complexity of the prediction process by lowering the size of the hypothesis space from that of the whole vocabulary $|V|$ to an average of $|K| + \mathbb{E}[|k|]$, making it a more tractable (and therefore more plausible) problem for readers to solve. We demonstrate an example case in Figure 2. More details of our implementation can be found at the clamp repository github page.[1]

Part-of-speech labels are strong predictors of clusters, with further subcategorizations being evident by assessing the component words for each cluster. The resulting clusters partially encode part-of-speech, with cluster agreement index (Rabbany and Zaïane, 2017) of 0.42 between part-of-speech

[1] https://github.com/calicolab/clamp

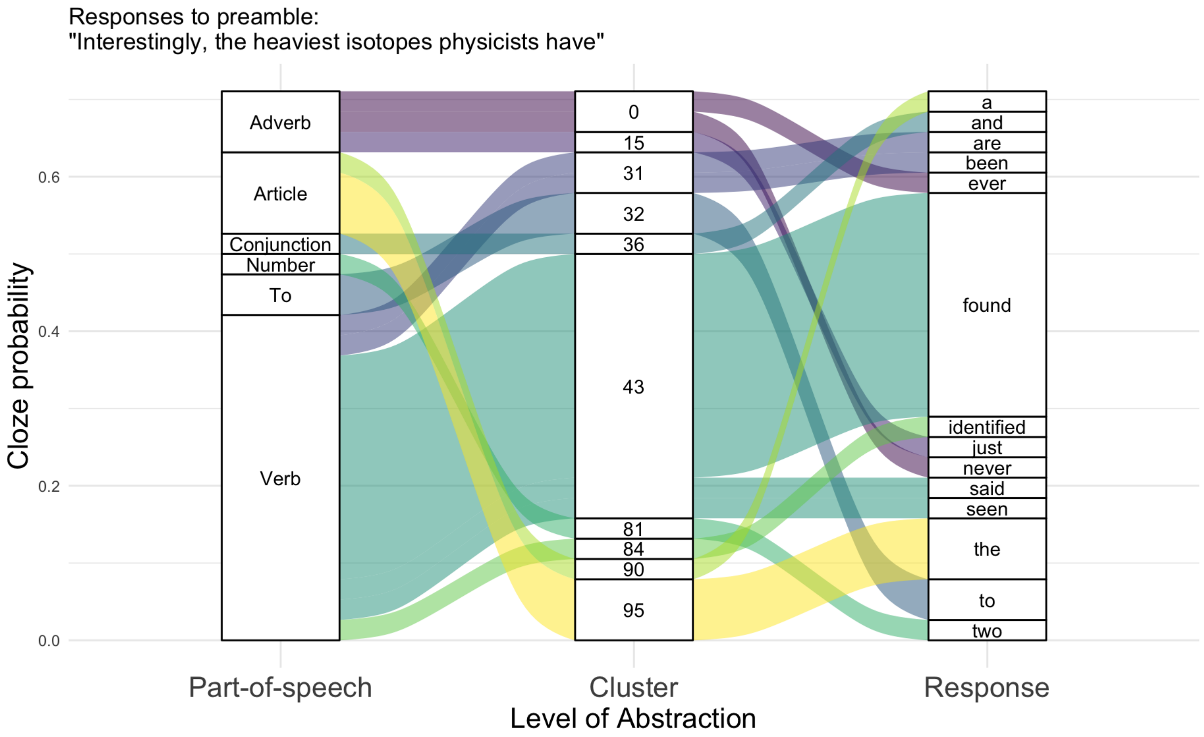

Figure 2: Word and POS repartition by cluster for responses to the preamble "Interestingly, the heaviest isotopes physicists have…". Plot made using ggalluvial (Brunson, 2020).

labels and our clustering. We present a visualization of cluster structure by word and by POS for a single preamble in Figure 2.

In contrast to lexical approaches and in keeping with the hierarchical nature of prediction (Kwisthout et al., 2017), out-of-sample words (i.e., words with a cloze probability of 0 that are the empirical next word in a sentence) may also attain a non-zero probability, which we explore in the next section.[2]

5 Predicting reading times

Hierarchical prediction mechanisms empower readers to make less precise predictions in cases of uncertainty, and result in greater ease of processing even at early stages (Kwisthout et al., 2017). We clustered human responses in the cloze (next-word prediction) portion of the Provo corpus by extracting their contextual representations from the hidden layers of RoBERTa (Liu et al., 2019). As described above, we apply Bayesian Gaussian mixture models and interpret the resulting clusters as approximations of higher-order lexico-semantic categories. In that setting, the probability of a cluster C for a given preamble p is the sum of the cloze probabilities (eq. 3) of its elements:

$$P(C \mid p) = \sum_{\text{word} \in C} P(\text{word} \mid p, C)$$

We constructed linear mixed effects models of FFD and FGPD measures for words in the Provo corpus that were either responses produced in the cloze task ($P(\text{word} > 0$; *guessed*; Table 1) or were not observed (*unguessed*; Table 2). Such cases are precisely where we would expect uncertainty to promote maintaining a general hypothesis rather than a very specific one about upcoming words (Bannon et al., 2024; Kwisthout et al., 2017; Giulianelli et al., 2024).

Following Luke and Christianson (2016), we include several basic predictors to model reading times for each word — log word frequency, word number, sentence number, word length, LSA Context Score, and cloze probability (where applicable) to the base model with maximal random intercepts and slopes. LSA Context Score was defined as the fit between the empirical next word and the surrounding context using a cosine similarity distance metric, which was reported to significantly influence processing in Luke and Christianson (2016).

[2] In principle, it is possible to leverage the uncertainty in a mixture model's assignment of a data point to clusters. In practice, most embeddings are assigned to a cluster with probability 1 due to properties of the embedding space that make lexical representations highly distinct from each other.

Coefficient	β	E	t	p	β	E	t	p
Intercept	–0.02	0.04	–0.64	n.s.	0.01	0.03	0.26	n.s.
Cluster Probability	–0.03	0.01	–3.35	***	–0.02	0.01	–2.36	*
Word frequency	–0.05	0.03	–1.73	.	—	—	—	—
Sentence Number	–0.01	0.01	–1.23	n.s.	0.02	0.01	1.85	.
Word in Sentence	–0.02	0.01	–1.78	.	0.01	0.01	1.04	n.s.
Word Length	0.18	0.03	7.20	***	0.04	0.02	2.60	**
LSA Context Score	–0.01	0.01	–1.07	n.s.	–0.02	0.01	–1.84	.

Table 1: Linear mixed effects model for FGPD for words with non-zero cloze probability. Singularity issues affecting model convergence led to the removal of the word frequency term from the FFD model. . represents $p < 0.1$; ** $p < 0.01$; *** $p < 0.001$.

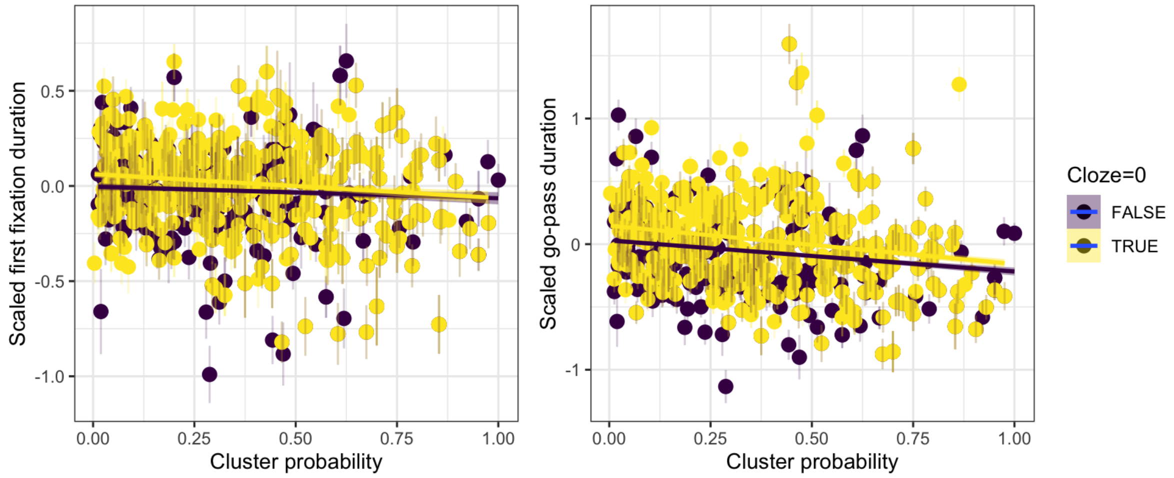

Figure 3: First fixation duration (FFD) and go-pass duration (FGPD) as a function of cluster probability for words that were guessed in the cloze norms. Cluster probability has a facilitative effect on both word types for both eyetracking measures.

Our analyses focus on FGPD and FFD specifically.

We tested for the importance of the Cluster Probability measure $P(C \mid p)$ through model comparison against a base model that did not include Cluster Probability as a predictor. Including Cluster Probability in the model resulted in significant improvements in ΔLL via a likelihood ratio test for all measures and datasets. All models showed the same pattern, such that next words belonging to higher-probability clusters were read more quickly, for both early visual stages (FFD) and higher order semantic stages (FGPD) and for guessed and unguessed words. For words that were guessed, cloze probability did not significantly predict either gaze measure and was thus excluded from the final model.

Furthermore, for such zero-cloze probability words, the effect of Cluster Probability on FPGD was larger (Satterwhaite $t(640) = -4.37$) than the effect of lexical frequency ($t(440) = 3.85$) for words that had zero-probability cloze but non-zero probability of that cluster. We visualize this relationship for FGPD in Figure 3 and present the results for zero-cloze FFDs in Table 2.

6 Related work

We are not the first to cluster language model representations. Others modeled semantic processing in analyses of reading times, typically comparing static word vectors for next words against prior context with cosine similarity (e.g., Luke and Christianson, 2016; Staub et al., 2015) or, more recently, used such similarities as a smoothing factor for surprisal distributions —slightly improving surprisal theory fits to reading time measures (Meister et al., 2024). Gaussian processes are particularly common in cognitive modeling of linguistic category learning (Kleinschmidt and Jaeger, 2015; Toscano

Coefficient	β	E	t	p	β	E	t	p
Intercept	−0.03	0.04	−0.72	n.s.	−0.01	0.03	−0.27	n.s.
Cluster Probability	−0.04	−0.01	−4.37	***	−0.03	0.01	−3.29	**
Word frequency	−0.08	−0.02	−3.85	***	−0.06	0.02	−3.45	***
Sentence Number	−0.05	0.01	−4.09	***	—	—	—	—
Word in Sentence	−0.01	0.01	−0.93	n.s.	—	—	—	—
Word Length	0.19	0.02	9.67	< .001	0.02	0.02	1.12	n.s.
LSA Context Score	−0.02	0.01	−1.72	.	−0.01	0.01	−1.38	n.s.

Table 2: Linear mixed effects model for FGPD for words with 0 cloze probability but non-zero cluster probability. Backwards elimination from the FFD model recommended removal of other control variables. . represents $p < 0.1$; ** $p < 0.01$; *** $p < 0.001$.

and McMurray, 2010). Modeling semantics using Dirichlet distributions (as in topic modeling Blei and Jordan, 2006) has also proven successful in modeling human semantic memory (Steyvers et al., 2006).

Other work in computational psycholinguistics has tested whether language processing involves a semantic comparison between alternatives in contextual language space (Giulianelli et al., 2023). We believe the current proposal that readers represent semantics as scalar, but nevertheless quasi-discrete, categories is a novel synthesis of these areas. The present results support the proposal that efficient, rational language processing can be achieved by combining levels of granularity of linguistic predictions.

7 Conclusion

Here we presented a hierarchical cognitive model of lexical processing that synthesizes early rational cognitive models with modern-day neural architectures. We argue that language model representations can be combined with human cloze data to infer higher-order structure. Cluster probabilities had predictive power in accounting for several reading measures for both predicted and unpredicted words and influence even early first fixation duration behavior. The results suggest that lexical processing can take place at a featural level in cases of uncertainty (Federmeier and Kutas, 1999; Roland et al., 2012; Kwisthout et al., 2017).

8 Limitations

This work is meant as a proof of concept for a hierarchical model of lexical processing and the use of transformer language models as predictors of reading times, not only through next-word probabilities, but also through their internal contextual representations of words. This study is by no mean exhaustive, and further replications and refinements using other datasets should be undertaken in the future using a wider variety of datasets.

Our work did not explore the potential semantic capacity of next word prediction-based language models. We did not consider larger models, nor simpler ones. We did not vary the number of clusters or manipulate the hyperparameters we used for the mixture model; future work should determine the optimal number of clusters.

The cloze norms and the eye tracking data here are relatively limited compared to real-world reading. The data are limited to American English which makes asking questions about other phenomena (e.g., morphosyntactic processing) more challenging. The data were gathered from highly literate populations at a prestigious university in the United States and are not representative of all people. Many individuals vary substantially in their language experience and this variability, which shapes processing fluency (e.g., Breen et al., 2024).

Some believe that lexicalized language model probabilities are the best probabilistic predictor of reading time and neural data (Shain et al., 2024), though this claim may not hold for all types of stimuli (de Varda et al., 2024; Szewczyk and Federmeier, 2022). Perhaps more pertinently, the use of surprisal as a measure of linguistic predictability is not central to our question; we demonstrated that cloze responses are highly structured and that modeling this structure accounts even for early language processing dynamics. Future work should examine how to automate the discovery of semantic probabilities using language models directly.

The appropriateness of modeling LLM embeddings as samples drawn from a mixture of multivariate Gaussians has not – to our knowledge – been extensively studied. The semantic structure of cloze responses can be approached from several angles, ranging from ontologies such as WordNet (Miller, 1995), to feature sets (Turton et al., 2020), to representing words by their distributional semantics, the approach we take here. However, the clustering results presented here are suggestive of meaningful distributional sub-structure, and could in principle be replicated by many other clustering algorithms, such as k-means or agglomerative clustering.

References

Charles E. Antoniak. 1974. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174.

Julie Bannon, Tamar H Gollan, and Victor S Ferreira. 2024. Is predicting during language processing worth it? effects of cloze probability and semantic similarity on failed predictions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

David M. Blei and Michael I. Jordan. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Mara Breen, Julie Van Dyke, Jelena Krivokapić, and Nicole Landi. 2024. Prosodic features in production reflect reading comprehension skill in high school students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Jonathan R Brennan and John T Hale. 2019. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741.

Jason Cory Brunson. 2020. ggalluvial: Layered grammar for alluvial plots. *Journal of Open Source Software*, 5(49):2017.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A bayesian approach to “overinformative” referring expressions. *Psychological Review*, 127(4):591.

Kara D Federmeier and Marta Kutas. 1999. A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4):469–495.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024. Generalized measures of anticipation and responsivity in online language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11648–11669.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

Keren Gruteke Klein, Yoav Meiri, Omer Shubi, and Yevgeni Berzak. 2024. The effect of surprisal on reading times in information seeking and repeated reading. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 219–230, Miami, FL, USA. Association for Computational Linguistics.

Dave F Kleinschmidt and T Florian Jaeger. 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, 122(2):148.

Johan Kwisthout, Harold Bekkering, and Iris Van Rooij. 2017. To be precise, the details don’t matter: On predictive processing, precision, and level of detail of predictions. *Brain and Cognition*, 112:84–91.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, 42:1166–1183.

Steven G. Luke and Kiel Christianson. 2016. Limits on lexical prediction during reading. *Cognitive Psychology*, 88:22–60.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16485–16498, Miami, Florida, USA. Association for Computational Linguistics.

George A. Miller. 1995. WordNet: a lexical database for English. *Commun. ACM*, 38(11):39–41.

Satoru Ozaki, Aniello De Santo, Tal Linzen, and Brian Dillon. 2024. CCG parsing effort and surprisal jointly predict RT but underpredict garden-path effects. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 362–364.

Reihaneh Rabbany and Osmar Zaïane. 2017. A General Clustering Agreement Index: For Comparing Disjoint and Overlapping Clusters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Douglas Roland, Hongoak Yun, Jean-Pierre Koenig, and Gail Mauner. 2012. Semantic similarity, predictability, and models of sentence processing. *Cognition*, 122(3):267–279.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Nathaniel Smith and Roger Levy. 2011. Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Adrian Staub, Margaret Grant, Lori Astheimer, and Andrew Cohen. 2015. The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, 82:1–17.

Mark Steyvers, Thomas L Griffiths, and Simon Dennis. 2006. Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, 10(7):327–334.

Jakub M. Szewczyk and Kara D. Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123:104311.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Joseph C Toscano and Bob McMurray. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, 34(3):434–464.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 1–8, Marseille, France. European Language Resources Association.

Titus Von der Malsburg and Bernhard Angele. 2017. False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, 94:119–133.