THE COST OF INFORMATION:
LOOKING BEYOND PREDICTABILITY IN LANGUAGE PROCESSING

Précis

Jacob Hoover Vigly
McGill University, Linguistics

As you read the words of this text, one after another, you build an understanding of the meaning it conveys. How do humans accomplish the task of language comprehension? Important clues about the mechanisms underlying this core cognitive capacity can be found in the patterns of effort during processing. A prominent approach in psycholinguistics has been that of surprisal theory (Hale, 2001; Levy, 2008), the hypothesis that the effort a word incurs is proportional to its negative log probability—an information-theoretic quantity known as surprisal, which quantifies how unexpected or surprising it is, given context. This influential hypothesis provides a direct link between the statistical predictability of words and human behaviour during language comprehension, based on the intuition that the cognitive cost of a word is fundamentally driven by the amount of information it contributes.

However, this focus on next-word prediction may be too narrow: This dissertation argues that standard surprisal theory has some significant and under-discussed shortcomings. Foremost, no known processing algorithm has complexity that scales directly proportional to surprisal. Additionally, there are empirical phenomena in human language processing behaviour that cannot be explained by standard surprisal theory. Based on core motivations from prior work, I develop a reframing of the central hypothesis of surprisal theory: I propose that processing cost directly reflects the computational complexity of updating probabilistic beliefs, which can be measured by the divergence between belief distributions. I argue that, by proposing that cost is a function of belief-update size, we are afforded the possibility for a much-needed theoretical connection between computational theories of processing difficulty and known inference algorithms. Namely, this proposal provides an intrinsic link to a wide family of potential theories of comprehension at an algorithmic level, such as sampling-based algorithms for probabilistic inference. Additionally, this proposal provides explanation for empirical phenomena wherein words are processed easily, even though they are unpredictable—such phenomena are inherently problematic for standard approaches that use surprisal as the measure of cost.

Another area where the predictability of words in context is potentially relevant to explaining language comprehension is in the description of latent linguistic structure—in terms of the dependency relationships between words. Linguistic dependency structures are widely used to describe the grammatical relationships that govern how a sentence is interpreted. At the same time, words display robust statistical relationships with each other, in a way that is intrinsically related to grammatical structure—for instance, as the result of agreement or selectional requirements. This dissertation contributes an analysis of the relationship between these two kinds of word-to-word dependencies, extracting dependency parses using probability estimates from large language models (LMs), and finding that the relationship is more tenuous than previously supposed.

The contributions of this dissertation are interdisciplinary in nature, bridging cognitive psychology, artificial intelligence, and linguistics. It explores central questions about the cognitive science of language using formal tools from information theory and models from natural language processing, and offers connections between psycholinguistics and literature on the computational complexity of incremental inference algorithms. This work provides evidence that human language processing costs arise not just from the challenge of predicting upcoming words, but from the computational demands of inferring and updating beliefs about meaning. These results contribute to an understanding of the relationship between distributional patterns of language use and the structures and mechanisms by which language is processed.

Quantifying processing cost with belief-update

During comprehension, the amount of cognitive resources required to integrate each word is variable and context dependent. What mechanism can explain why a given word is harder or easier to process? Chapter 1 provides an overview of the approach I take to answering this question, proposing to measure processing cost with divergence between belief distributions, and situating this proposal with respect to prior literature. This chapter also gives derivations of novel predictions within this framework, which are further explored and tested in subsequent chapters.

This dissertation follows a growing body of previous literature in modelling comprehension as probabilistic (Bayesian) inference, within the larger framework of rational analysis of cognition (Anderson, 1990; Anderson & Schooler, 1991; Chater et al., 1999), taking the view that the processing cost can be measured by the size of the Bayesian update in beliefs about the latent interpretation, given the new information contained in the word. This intuition has been encapsulated as the key justification for the influential *surprisal theory* (Hale, 2001; Levy, 2008), which hypothesizes that the cognitive effort associated with processing a word w in context is proportional to its surprisal, an information theoretic quantity defined as the negative log probability of the word given context:

> *Surprisal theory*—quantifying cost as unpredictability.
>
> The processing cost of a word w in context increases proportional to its surprisal, defined as $s(w) := -\log p(w \mid \text{context})$. That is,
>
> $$\text{cost}(w) \propto s(w)$$

Intuitively, the more surprising a word is, the larger a change it causes in a comprehender's beliefs about the meaning of the utterance, in a rational inference setting. Empirically, it is well documented that words that are less expected are harder to process—for example, during reading, people spend more time looking at words which are less predictable given context, as has been known for decades (Ehrlich & Rayner, 1981; Balota et al., 1985; McDonald & Shillcock, 2003a, 2003b), and this relationship has offered empirical support to surprisal theory as a broad explanation of processing difficulty (Smith & Levy, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020; Wilcox et al., 2023; Hofmann et al., 2022; Shain et al., 2024).

However, the standard arguments for deriving surprisal theory require two important assumptions: These are **(1)** that surprisal is equivalent to belief update size; and **(2)** that the linking function between belief update and processing cost is linear. These assumptions are explicitly described in the original work motivating surprisal theory (Levy, 2005, 2008), and while there has been a lively debate about the form of the linking function (Levy & Jaeger, 2006; Brothers & Kuperberg, 2021; Meister et al., 2021; van Schijndel & Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024; Shain et al., 2024), the theoretical justifications for these two assumptions have largely gone unquestioned in much subsequent work. I argue that there are important and timely reasons to question both these assumptions, from theoretical as well as empirical perspectives.

Relaxing these two assumptions gives a more general hypothesis, which I refer to as *divergence theory*—measuring cost as the divergence between belief distributions over intended meanings before versus after observing a word:

> *Divergence theory*—quantifying cost with belief-update size.
>
> The processing cost of a word w increases as a function of the amount of information it communicates, as quantified by the Kullback-Leibler (KL) divergence (a.k.a. relative entropy) between the posterior distribution $p_{Z|w}$ and the prior p_Z. That is,
>
> $$\text{cost}(w) = f\left(D_{\text{KL}}(p_{Z|w} \,\|\, p_Z)\right)$$
>
> where f is a monotonically increasing function, and Z is the latent variable that is the target of inference—the meaning of the utterance—about which beliefs are updated upon observing w.

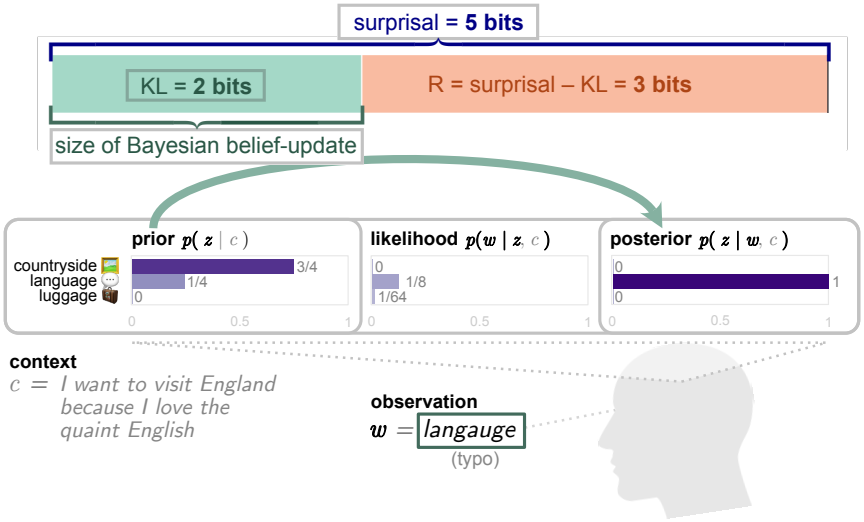

Figure 1: Diagram illustrating a toy example of surprisal and information gain of an observation w in context, with prior and likelihood chosen such that surprisal markedly larger than KL divergence. The remaining bits constitute the portion of the total information about the precise form of the observation that does not contribute to belief update. [Note that a minimally different alternative example with a likelihood function that assigned probability 1 rather than $1/8$ to the inferred meaning would have resulted in surprisal being equal to KL at 2 bits.]

This hypothesis captures several key insights, detailed in this chapter. First, it formalizes the motivation for surprisal theory, while generalizing it: The surprisal of a word provides an upper bound on the belief-update cost measured by divergence. Second, it provides the flexibility to capture empirical phenomena where processing difficulty is low on words that are unpredictable. Third, it offers a potential connection to a broad family of well-studied sampling-based inference algorithms, whose complexity scales in such a divergence, and which form a promising yet under-explored space of models for language processing.

In general, divergence between posterior and prior can be decomposed, pulling out a term for surprisal (as was first shown in Levy, 2005). This decomposition can be arranged to describe a way of partitioning of the information (bits) of surprisal of into two nonnegative quantities:

$$\overbrace{\log \frac{1}{p(w)}}^{s(w)} = \overbrace{\mathop{\mathbb{E}}_{p_{Z|w}}\left[\log \frac{p(z \mid w)}{p(z)}\right]}^{D_{\mathrm{KL}}(p_{Z|w} \,\|\, p_Z)} + \overbrace{\mathop{\mathbb{E}}_{p_{Z|w}}\left[\log \frac{1}{p(w \mid z)}\right]}^{R(w)} \tag{1}$$

The first term, $D_{\mathrm{KL}}(p_{Z|w} \,\|\, p_Z)$, quantifies the size of the Bayesian belief update induced upon observing w. The second term, $R(w)$, defined as the expected value of the likelihood over the posterior, quantifies the remaining bits of information that do not contribute to belief update. This second quantity is zero if there is a deterministic relationship between latent representations (Z) and observable words, however may be nonzero if this is not the case. If surprising words will tend to cause commensurately large changes in beliefs about the meaning of the utterance, then the magnitude of surprisal will tend to be be similar to that of KL divergence. This is always the case in the original setting for surprisal theory, which explicitly assumed a probabilistic model in which there was a deterministic relationship between latent linguistic structures over which belief distributions range, and the observable words (Levy, 2008). Adopting the divergence theory hypothesis allows us to relax this assumption, offering the flexibility to also capture situations where a word is highly unpredictable, yet does *not* incur a large change in beliefs.

Figure 1 illustrates the decomposition of surprisal in eq. 1 for a toy example where the magnitude of surprisal and divergence differ markedly. This illustration is meant to demonstrate the way in which surprisal forms only a loose

upper bound on belief update, in any Bayesian inference setting where the likelihood function is not deterministic. In this example the value of the likelihood is small even for the meaning that is inferred, as a consequence of the observed word's containing a typographical error (such examples are modelled and explored empirically with a reading-time experiment in Chapter 3).

As noted above, the hypothesis that cost scales as a function of divergence between prior and posterior distributions reduces completely to standard linear surprisal theory if two assumptions are made. This framing naturally leads to the question of whether these assumptions are justified. The following chapters investigate the theoretical and empirical justifications for relaxing each of these assumptions of standard surprisal theory, one at a time. Chapter 2 investigates the form of the linking function, presenting novel theoretical arguments based on the computational complexity of sampling algorithms, which predict a superlinear (rather than linear) linking function between belief-update and processing cost. These predictions are tested and supported with results from nonlinear regressions fit to model the effect on human reading times on surprisal, as estimated by pre-trained language models. Then, in Chapter 3, I take up the question of whether the other assumption is justified, proposing that typographical errors intuitively present an example of input which may be high surprisal but not be difficult to process. Such behaviour cannot be explained under standard surprisal theory, but can be accounted for under the proposed divergence theory. This intuition is evaluated with a self-paced reading study, using a variety of LMs as probability estimators.

Arguments and evidence for a superlinear linking function

Chapter 2 investigates the linking function between belief update size and processing cost. In order to focus on the question of the linking function specifically, this chapter follows all previous literature in this area in explicitly assuming that divergence is equivalent to surprisal (leaving aside the question about whether this assumption is always merited, which is taken up as the focus of the subsequent chapter).

A processing algorithm can be conceived of as a mechanism for building a representation of the posterior distribution given an observation. In this probabilistic inference framework, the natural way in which computational cost might be related to belief-update is if the algorithm gives priority to high-probability regions of the space of meanings, when building its representation of the posterior. A broad class of algorithms which privilege likely meanings are those which sample hypotheses from a prior distribution. This chapter contributes an analysis of some fundamental examples of such algorithms, revealing that they predict runtime to increase in surprisal superlinearly, and with increasing variance. This is also the case for more sophisticated algorithms based on importance sampling, where the number of samples required scales exponentially in KL divergence (and therefore in surprisal, under the standard assumption of their equivalence).

These predictions are notably in tension with the standard conception of surprisal theory, which assumes a linear linking function, and constant variance. Indeed, a majority of previous studies have simply assumed a linear, constant-variance linking function—either explicitly, or more often implicitly in their choice of statistical models for analysis (Mitchell, 1984; Reichle et al., 2003; Demberg & Keller, 2008; Frank, 2009; Fernandez Monsalve et al., 2012; Frank et al., 2013; Lowder et al., 2018; Hao et al., 2020; van Schijndel & Linzen, 2021; Kuribayashi et al., 2022). A smaller number of empirical papers have investigated the shape of the linking function directly (Smith & Levy, 2008, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020; Hofmann et al., 2022). For the most part, these studies have found support for the assumption of linearity. However, we argue there are a number of methodological reasons to revisit these results, in addition to the theoretical motivation provided by our analysis of sampling algorithms.

The second part of Chapter 2 provides an empirical analysis of the linking function, using generalized additive models (GAMs) to predict reading times on the Natural Stories corpus (Futrell et al., 2021), using surprisal estimates from a variety of pre-trained language models.

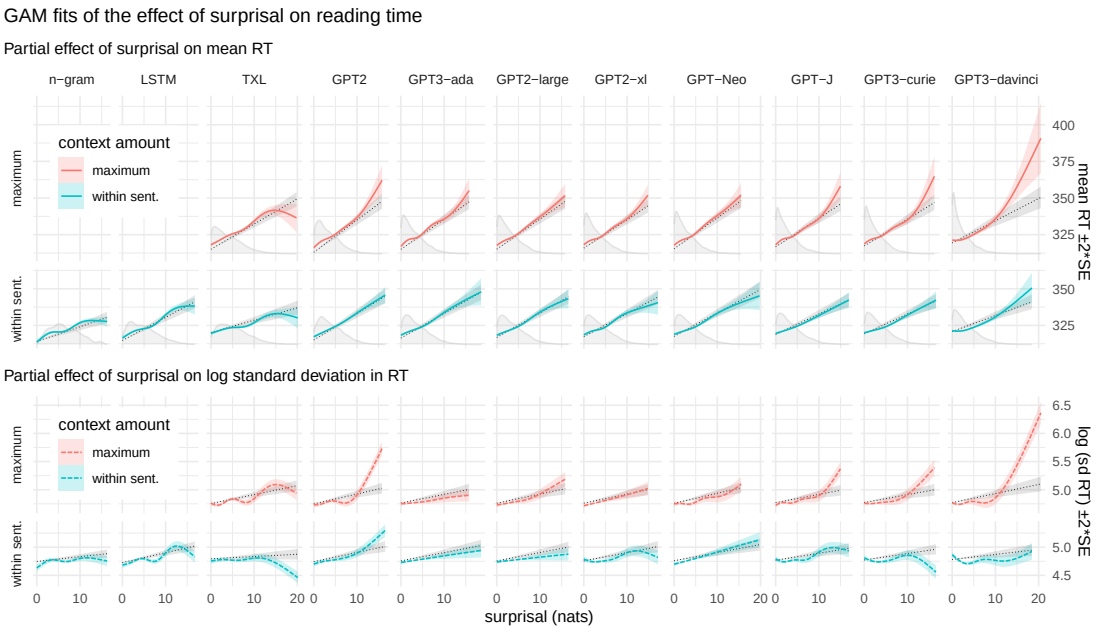

Figure 2: The effect of surprisal on self-paced reading time. Coloured lines are the fitted effects from the nonlinear GAMs, dotted black lines beneath are from the corresponding linear control GAMs. **Top two rows**: effect of surprisal on mean RT, with density plots of surprisal underlaid at the bottom. The top row (red) uses surprisals from LMs with full access previous context, the second row (blue) uses LMs with access only to within-sentence context. **Bottom two rows**: as the first two, but for the effect of surprisal on variance in RT (as log standard deviation).

As shown in fig. 2, our results give evidence that processing cost increases more steeply at higher levels of surprisal, particularly for higher-quality language models with access to full context, suggesting a superlinear linking function. We also find that variance in cost increases as a function of surprisal, consistent with the predictions of sampling-based inference algorithms. Further quantitative assessment confirms this general qualitative interpretation of our results: The more accurate the LM used to calculate surprisal is, the more superlinear the effect of surprisal on reading time. We interpret these results as evidence supporting the plausibility of sampling-based algorithms for sentence processing.

When unpredictable does not mean difficult to process

Chapter 3 shifts focus away from the form of the linking function, to question the assumption that surprisal is equivalent to KL divergence, which is implied if the relationship between latent structure and observed words is deterministic. I argue that there are real-world situations in which we can expect surprisal to be only a loose upper bound on the divergence from prior to posterior. Such cases would provide examples where the predictions of standard surprisal theory about processing cost should differ most drastically from the proposed divergence theory.

An ideal situation in which to distinguish whether effort is driven by KL divergence or surprisal would be one where the divergence is identical across conditions, but surprisal is manipulated. As illustrated in the toy example in fig. 1, the divergence incurred between prior and posterior upon observing a word with a typographical error may reasonably be expected to depend primarily on the meaning it contributes: Namely, the processing effort may be small even if the observation is highly unpredictable. Comparing processing effort on a highly predictable word, with and without a typographical error, provides precisely the kind of situation required to evaluate divergence theory versus surprisal.

This chapter presents a self-paced reading time experiment on a data set of hand designed example sentences with

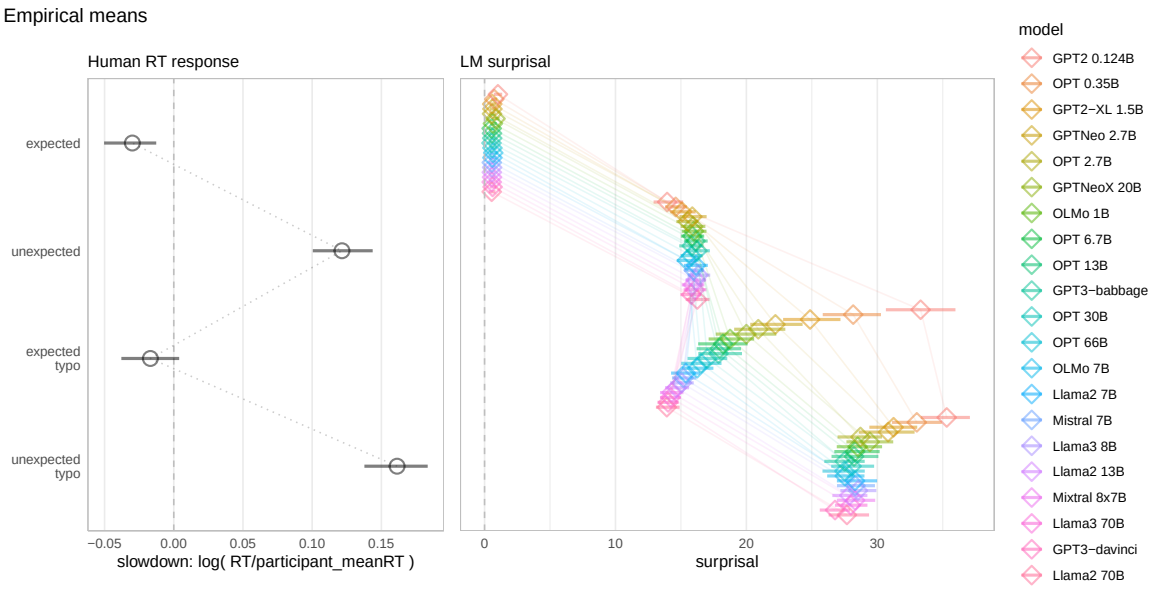

Figure 3: Empirical means of human reading time response and mean LM surprisal, across the four experimental conditions. Diamonds mark mean values, with horizontal lines indicating 99% CIs. **Left**: Reading time response represented on the horizontal axis as "slowdown", the log RT on the target region time relative to the participant's overall mean log RT. **Right**: Horizontal axis is surprisal; each LM is plotted in a separate colour.

target words in identical contexts for each of four conditions—either an expected or unexpected meaning, and with or without a typographical error—as a case study to compare the predictions of surprisal versus KL divergence. It presents a self-paced reading time experiment to assess human processing cost on these materials and compare against surprisal estimates from a collection of LMs.

As shown in fig. 3, we found that while surprisal estimates from language models (right subplot) consistently predicted high processing cost for words with typos (regardless of whether the underlying word was expected or unexpected), human reading times (left subplot) showed a different pattern. In particular, typos on expected words showed relatively little processing cost compared to typos on unexpected words, despite having high surprisal values. This pattern in human processing cannot be explained by standard surprisal theory, but is precisely what is predicted under divergence theory. Surprisal intrinsically cannot distinguish between words which are unpredictable due to their conveying an unexpected meaning versus words that are unpredictable for some other reason, unrelated to the meaning they convey (such as their containing a minor typographical error). Our results confirm that there exist situations in which surprisal is not adequate as an explanation of processing cost for humans, which instead may be explained by the size of the belief update, measured by KL divergence.

Comparing statistical and linguistic dependencies between words

In Chapter 4, the final content chapter, I put aside questions about incremental processing and cognitive effort, to focus instead on the relationships between words that describe the structure of language. This chapter presents an examination of the connection between linguistic structure and the distributional patterns of words, by comparing the word-to-word relationships represented in linguistic dependency structures to those encoded by statistical dependencies in context.

This investigation was motivated by the question of whether words that stand in linguistic dependency relationships with each other tend to also be dependent on each other in terms of their co-occurrence frequency. We extracted tree structures which maximize pointwise mutual information between words, in context, and compared these resulting

tree structures to linguistic dependency structures.

In this study we found that the word-to-word arcs in the statistical dependency trees corresponded with linguistic dependencies at a rate that was substantially above chance, and more so for the trees extracted using the language models which take surrounding context into account. This finding confirmed a tendency also noted in earlier work (Futrell et al., 2019) that words that are related to one another syntactically are likely to depend upon each other statistically. However, our analysis revealed that as a method of dependency parsing, extracting statistical dependency trees is in general only roughly as good as the simple baseline heuristic of connecting adjacent words. This finding was robust across multiple languages and was not improved by using language models designed with an explicit bias for hierarchical structure, nor by adopting a delexicalized variant to our method. We interpret these results as evidence that while there are some superficial ways in which statistical dependence can be related broadly to linguistic dependencies, we do not see evidence of a deep and systematic relationship.

Discussion

This dissertation makes several key contributions to our understanding of human language processing, bridging cognitive psychology, computer science, and linguistics. The work develops and provides evidence for a fundamental reframing of how we think about processing difficulty in language comprehension—advocating moving beyond simple predictability in terms of surprisal to instead use models which consider the computational demands of probabilistic inference about meaning.

This re-framing generalizes existing surprisal-based accounts while maintaining their core motivations, and providing a connection to existing results on the computational complexity inference algorithms. Theoretical predictions within this framework were supported through complementary empirical studies: one showing evidence for super-linear scaling of processing cost with surprisal (as predicted by sampling-based inference algorithms), and another demonstrating that even highly unpredictable (high surprisal) strings can be easy to process when they don't require large updates to beliefs about meaning.

This research has several broader implications for cognitive science: It provides a bridge between computational-level theories of processing difficulty and algorithmic implementations, suggesting sampling-based probabilistic inference as a promising framework for modelling human sentence processing. It is situated along with other work in probabilistic modelling applied to language processing, in demonstrating how ideas from information theory and machine learning can inform our understanding of core cognitive processes, while maintaining theoretical connections to rational approaches to cognition. The finding that statistical dependencies between words correspond only loosely with linguistic dependencies challenges simplistic assumptions about the relationship between distributional and structural aspects of language.

The timing of this theoretical contribution is particularly relevant given the recent advent of large language models capable of increasingly accurate probability estimates. As these models become more sophisticated at predicting words in context, understanding the limitations of surprisal-based theories becomes crucial. This work proposes that while predictability is important, it does not tell the whole story about processing cost.

Several directions for future work emerge from these findings. First, developing explicit computational models that estimate divergence could provide more detailed predictions about processing patterns. Second, the framework could be applied to explain other phenomena from the theoretical linguistics and psycholinguistics literature where processing difficulty is smaller than would be expected under a surprisal-based model, such as grammatical illusions (Wason & Reich, 1979; Wellwood et al., 2018; Muller & Phillips, 2020). Third, the connection with sampling algorithms suggests implementations using such methods (as in Levy et al., 2008) are promising for building process-level models, toward a more complete picture of language comprehension.

Bibliography

Anderson, J. R. (1990, January). *The adaptive character of thought*. Psychology Press. https://doi.org/10.4324/9780203771730

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. https://doi.org/10.1111/j.1467-9280.1991.tb00174.x

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 301–313. https://aclanthology.org/2022.conll-1.20

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, *17*(3), 364–390. https://doi.org/10.1016/0010-0285(85)90013-1

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Chater, N., Oaksford, M., Chater, N., Oaksford, M., Chater, N., Oaksford, M., Chater, N., Oaksford, M., Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65. https://doi.org/10.1016/S1364-6613(98)01273-X

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. https://doi.org/10.1016/j.cognition.2008.07.008

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Memory and Language*, *20*(6), 641. https://doi.org/10.1016/S0022-5371(81)90220-6

Fernandez Monsalve, I., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 398–408. https://aclanthology.org/E12-1041

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*. Retrieved October 12, 2022, from https://escholarship.org/uc/item/02v5m1hf

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2013). Word surprisal predicts N400 amplitude during reading. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 878–883. https://www.aclweb.org/anthology/P13-2152

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S. T., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*(1), 63–77. https://doi.org/10.1007/s10579-020-09503-7

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, 3–13. https://doi.org/10.18653/v1/W19-7703

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*. https://doi.org/10.18653/v1/w18-0102

Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://www.aclweb.org/anthology/N01-1021

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 75–86. https://doi.org/10.18653/v1/2020.cmcl-1.10

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, *4*, 730570. https://doi.org/10.3389/frai.2021.730570

Hoover, J. L. (2024, August). *The cost of information: Looking beyond predictability in language processing* [Doctoral dissertation, McGill University]. https://escholarship.mcgill.ca/concern/theses/r494vr42w

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510. https://doi.org/10.1016/j.jml.2024.104510

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10421–10436. Retrieved April 30, 2023, from https://aclanthology.org/2022.emnlp-main.712

Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity* [Doctoral dissertation, Stanford University]. https://www.proquest.com/docview/305432573

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Levy, R., & Jaeger, T. F. (2006). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. C. Platt, & T. Hofmann (Eds.), *Proceedings of the twentieth annual conference on neural information processing systems* (pp. 849–856). MIT Press. https://proceedings.neurips.cc/paper/2006/hash/c6a01432c8138d46ba39957a8250e027-Abstract.html

Levy, R., Reali, F., & Griffiths, T. L. (2008, December). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Proceedings of the twenty-second annual Conference on Neural Information Processing Systems* (pp. 937–944). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2008/hash/a02ffd91ece5e7efeb46db8f10a74059-Abstract.html

Lowder, M. W., Choi, W., Ferreira, F., & Henderson, J. M. (2018). Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive Science*, *42*(S4), 1166–1183. https://doi.org/10.1111/cogs.12597

McDonald, S. A., & Shillcock, R. C. (2003a). Low-level predictive inference in reading: The influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751. https://doi.org/10.1016/s0042-6989(03)00237-2

McDonald, S. A., & Shillcock, R. C. (2003b). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*(6), 648–652. https://doi.org/10.1046/j.0956-7976.2003.psci_1480.x

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.74

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading 1. In *New Methods in Reading Comprehension Research*. Routledge.

Muller, H., & Phillips, C. (2020, March 25). Negative polarity illusions. In V. Déprez & M. T. Espinal (Eds.), *The Oxford Handbook of Negation* (pp. 656–676). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198830528.013.42

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*(4), 445–476. https://doi.org/10.1017/s0140525x03000104

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*, 570–576. https://escholarship.org/uc/item/3mr8m3rf

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988. https://doi.org/10.1111/cogs.12988

Wason, P. C., & Reich, S. S. (1979). A Verbal Illusion. *Quarterly Journal of Experimental Psychology*, *31*(4), 591–597. https://doi.org/10.1080/14640747908400750

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The Anatomy of a Comparative Illusion. *Journal of Semantics*, *35*(3), 543–583. https://doi.org/10.1093/jos/ffy014

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of Surprisal Theory in 11 Languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470. https://doi.org/10.1162/tacl_a_00612

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713. https://www.cognitivesciencesociety.org/cogsci20/papers/0375/