THE COST OF INFORMATION
LOOKING BEYOND PREDICTABILITY IN LANGUAGE PROCESSING

Jacob Louis Hoover

Department of Linguistics
McGill University, Montréal

August 2024

A thesis submitted to McGill University
in partial fulfillment of the requirements of
the degree of Doctor of Philosophy

Abstract

What are the structures by which words are related to one another within a sentence, and how are these structures processed? This dissertation approaches these central questions from the perspective of information theory, examining how the statistical patterns of language use can inform theories of linguistic structure and incremental language processing.

The main focus of this work is to explain incremental processing difficulty, the question of what makes a word harder or easier to process when it is encountered in context. During comprehension, as each new word is observed, it provides some amount of information to be incorporated into an understanding of the overall message. In recent decades, a prominent approach to explaining human processing cost has been that of surprisal theory, based on the hypothesis that the cost of a word is fundamentally a function of the amount of information it contains, quantified as its negative log probability, or surprisal. This hypothesis has broad empirical support and is widely accepted, but also has some significant and under-discussed weaknesses. Primarily, no known processing algorithm has complexity directly proportional to surprisal. Additionally, from an empirical perspective, there are phenomena of human processing cost that cannot be predicted by surprisal alone, such as constructions in which a word is highly unpredictable but does not incur high processing cost. In this dissertation, building on previous research, I develop a reframing of the central hypothesis of surprisal theory, to measure processing cost with divergence between belief distributions. This quantification of information gain is mathematically equivalent to surprisal only with certain simplifying but potentially problematic assumptions, which previous literature has implicitly or explicitly assumed. This proposed generalization of traditional surprisal theory builds on its established strengths while affording important advantages, both theoretical and empirical. Namely, it allows an intrinsic link to a wide family of potential algorithmic theories, such as sampling-based belief-update algorithms for approximate inference. It also offers the ability to explain phenomena in human processing cost that standard surprisal theory cannot, including the efficient processing of minor production mistakes, such as typographical errors. In this work, I develop these theoretical predictions and present empirical studies to predict human reading time using estimates of contextual word probability from a variety of large language models.

Another area where the distributional patterns of language use are relevant to explaining human language processing 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 intuitive connection between linguistic and statistical relationships between words raises the question: Do words that are strongly statistically dependent upon each other tend to be related by linguistic dependency, and vice versa? This work contributes an analysis of the relationship between these two kinds of word-to-word dependencies, using probability estimates from large language models, and finding that the relationship is more tenuous than previously supposed.

Together, the theoretical and empirical contributions in this dissertation contribute to an understanding of the relationship between the distributional patterns of language use and the structures and mechanisms by which it is processed.

Abrégé

Quelles sont les structures qui relient les mots dans une phrase, et comment sont-elles assimilées ? Cette thèse aborde ces questions fondamentales du point de vue de la théorie de l'information, examinant comment les motifs statistiques de notre usage de la langue peuvent informer les théories sur la structure et le traitement incrémental de la langue.

L'objectif principal de ce travail est d'expliquer ce qui rend un mot plus difficile ou plus facile à traiter lorsqu'il est observé dans son contexte. Au cours de la compréhension, chaque nouveau mot observé fournit une certaine quantité d'informations à intégrer dans la compréhension du message global. Au cours des dernières décennies, « surprisal theory » a constitué une approche principale pour expliquer le coût du traitement mental de l'information. Cette théorie repose sur l'hypothèse selon laquelle le coût d'un mot est fondamentalement une fonction de la quantité d'informations qu'il contient, quantifiée comme son logarithme négatif de probabilité, ou communément appelée « surprisal ». Cette hypothèse bénéficie d'un large soutien empirique et est largement acceptée, mais elle présente également des faiblesses importantes et peu discutées. Tout d'abord, aucun algorithme de traitement connu n'a une complexité directement proportionnelle au « surprisal ». En outre, il existe des phénomènes empiriques qui ne peuvent être prédits seule par « surprisal », tels que les constructions dans lesquelles un mot est hautement imprévisible mais n'entraîne pas de coût de traitement élevé. Dans cette thèse, je développe une reformulation de l'hypothèse centrale de « surprisal theory », afin de quantifier l'information avec une divergence entre distributions de probabilité. Cette quantification du gain d'information n'est mathématiquement équivalente au « surprisal » qu'avec certaines suppositions simplificatrices et potentiellement problématiques. La reformulation proposée s'appuie sur les points forts de la théorie traditionnelle tout en offrant d'importants avantages, tant théoriques qu'empiriques. Elle établit un lien intrinsèque avec une large famille de théories algorithmiques potentielles, comme les algorithmes de mise à jour des croyances basés sur l'échantillonnage pour l'inférence approximative. Elle permet également d'expliquer des phénomènes que la théorie standard ne peut expliquer, notamment le traitement efficace d'erreurs de production mineures. Dans ce travail, je développe ces prédictions théoriques et présente des études empiriques pour prédire le temps de lecture humain en utilisant de la probabilité contextuelle des mots estimée à partir de grands modèles de langage.

La description de la structure linguistique latente est un autre domaine dans lequel les tendances distributionnelles de l'utilisation de la langue sont pertinentes pour expliquer le traitement du langage humain. Les structures de dépendance linguistique sont couramment utilisées pour décrire les relations grammaticales qui régissent l'interprétation d'une phrase. Parallèlement, les mots présentent des relations statistiques entre eux, d'une manière qui est intrinsèquement liée à la structure grammaticale. Ce lien intuitif entre les relations linguistiques et statistiques entre les mots soulève la question suivante : Les mots qui dépendent fortement les uns des autres d'un point de vue statistique ont-ils tendance à être liés par une dépendance linguistique, et vice versa ? Ce travail propose une analyse de la relation entre ces deux types de dépendances entre les mots, en utilisant des estimations de probabilité provenant de grands modèles de langage, et en constatant que la relation est plus fragile que ce que l'on supposait auparavant.

Les contributions théoriques et empiriques de cette thèse contribuent à notre compréhension de la relation entre les tendances distributives de l'utilisation de la langue et les structures et mécanismes par lesquels elles sont traitées, puis assimilées.

Contents

List of figures

List of regression formulæ

List of tables

Acknowledgements

First and most of all, I want to express my gratitude to my supervisor, Tim O'Donnell. Tim has provided me with a model of how to be a scientist, offered consistent and conscientious mentorship for the past six years, and continually inspired me with his creativity, tenacity, and breadth of intellectual curiosity. I have thoroughly enjoyed the many times when his willingness to rigorously engage with a question would lead to an hours-long open-ended discussion while filling a whiteboard, with the conversation often continuing into a walk across Montreal. It has been a fulfilling journey getting to this landmark, and I look forward to where it may go from here. Thank you so much, Tim.

I would also like to thank the other people who have served as advisors and mentors during my time at McGill and Mila, in particular Morgan Sonderegger, whose advice, instruction, and encouragement were instrumental in the completion of this work. Conversations with Morgan have often pushed me toward a better and more thorough understanding of methodological questions, and I have always appreciated his attention to detail and honesty in approaching all aspects of the research endeavor. I am also fortunate to have had Alessandro Sordoni as a co-advisor on my first evaluation paper; I found discussions with Alessandro extremely valuable as I worked to develop my interests and motivations for research. Thank you also to Steve Piantadosi, Wenyu Du, and Peng Qian, collaborators on the projects presented here, all of whom I am privileged to have worked with, and to Vikash Mansinghka for being on my committee. I am also grateful to Michael Wagner and Vera Demberg, for serving as my dissertation examiners, and for their insightful and helpful comments and questions—and also to Vera specifically for reading the defence draft closely enough to notice an embarrsasingly large number of typos (and taking the time to highlight each one, a detail I found somewhat delightfully ironic, given the topic of chapter 3).

I have had the enormous good fortune to spend these past years in a truly supportive department and community. Many thanks to Giuliana Panetta and Andria De Luca for their kindness and fantastic work as administrators, and to my fellow students and labmates for their comradery. In particular, to Michaela Socolof, it has been a true pleasure to navigate the singular journey of graduate school with you on a parallel trajectory the whole way. Also thank you to Jonny Palucci, for boundless enthusiasm and generosity; to Ben LeBrun, for many helpful discussions and insights; to Eva Portelance, for refreshing perspective and encouragement when I really needed it (and for translation help). To the other students I've shared time with these past years, and whose conversation and companionship has enriched my experience intellectually and beyond: Emily Kellison-Linn, Emily Goodwin, Emily Baylor, Connie Ting, Gaurav Kamath, Amanda Doucette, Laurestine Bradford, Mathieu Paillé... Merci and thank you, all.

Thank you to Jessica Coon for advice, help, and opportunities to branch out (even as far outside of Linguistics as a workshop on handstands). Likewise thank you Prakash Panangaden and Brendan Gillon, for some especially stimulating conversations at the very beginning of my time at McGill, about research directions that I did not end up pursuing, but still hope to, partially based on how much I enjoyed those conversations. More recently, I am also grateful to Roger Levy welcoming me into his lab at MIT as a visiting student in my penultimate semester. For their conversation and help along the way to completing this dissertation, thank you also to Meghan Clayards, Vojkan Jakšić, Tiwalayo Eisape, Thomas Hikaru Clark, and Alex Lew.

From the time leading up to my coming to McGill, I am indebted to Sabine Iatridou, David Pesetsky, Mark Steedman, and Polly Jacobson for enthusiastically welcoming me to the field of

Linguistics, including inviting me to audit courses, despite my not having any academic affiliation, and taking the time to help, advise, and encourage me in pursuing the career switch from dancer to academic. Thanks also to Marguerite Mahler, Andrew Nevins, Michael Becker, and the infectious enthusiasm of Paul Bamberg, the lasting impression of whose variety of topics courses in undergrad Mathematics more than a dozen years ago were part of what eventually led me back to pursue mathematical interests.

A special thank you also to my parents Michele and Jeff, and to Zev and Zoe for giving their full attention to some under-prepared practice talks, and for enthusiastically reading my draft manuscripts, and of course to Zev for many late-night discussions in the kitchen. And to Katie, for constant support, patience, advice, summer squashes, daily porridge, for helping me keep moving throughout, and for always being willing to listen despite sometimes being impelled to roll on the floor in order to cope with the theoretical abstractness of it all. You are true!

0

Overview

This dissertation can be broadly situated as an examination of questions about how linguistic structures and the mechanisms of language processing are related to statistical patterns of language use, focusing on computational theories of incremental processing difficulty (chapters 1 to 3), and also examining the connection between linguistic structure and statistical dependence between words (chapter 4). I explore these questions using formal tools from information theory, and present empirical studies using probability estimates from a variety of pretrained large language models to explain human reading behaviour and linguistic structure.

Throughout this work I am interested in forming and evaluating hypotheses about how aspects of human language can be explained via the distributional patterns that are observed in linguistic data, viewing language comprehension as probabilistic inference. There has been increasingly robust evidence that comprehenders make rational inferences about the intended message within a noisy environment, assessing relative likelihoods at the level of phonemes, and also morphemes/-words, and structure (R. Levy, 2008a; Clayards et al., 2008; Piantadosi et al., 2011; Gibson et al., 2013; Ryskin et al., 2018). Fundamental to the approach to comprehension as inference is the perspective that the general linguistic environment can be viewed in a probabilistic light, where the predictability or unpredictability of a linguistic item in context is a central property of interest (for reviews of the ways in which linguistic knowledge and behaviour can be viewed in a probabilistic light, see, e.g., Chater & Manning, 2006; R. Levy, 2013; Lau et al., 2017). In particular, (un)predictability can be quantified as the log inverse probability—an information-theoretic quantity

known as *surprisal*, which I will denote with $s(\cdot)$.

$$s(x) \coloneqq \log \frac{1}{p(x)} = -\log p(x) \tag{1}$$

Surprisal quantifies the amount of information contained in a particular observation x of a discrete random variable. It is zero when the observation x is certain, and increases to infinity as probability approaches zero. With the observation of interest being a word (or other unit of linguistic input), this simple information-theoretic quantity has a long history of applications to explaining aspects of linguistic structure and processing.

Brief history of surprisal in psycholinguistics

In the aftermath of Shannon (1948)'s groundbreaking paper introducing the fundamental concepts of "A Mathematical Theory of Communication" (founding the field which has come to be known as information theory), there was considerable interest in developing a variety of applications of this new way of understanding communication, including many in the cognitive sciences (e.g., Hick, 1952; Hockett, 1953; Miller, 1957; Attneave, 1959; Garner, 1962). Use of the term *surprisal* to refer to the log inverse probability of an observation of a discrete random variable traces back to this time, being, as far as I am aware, first introduced by Samson (1953, The Surprisal Property of Present Events, p. 293). He proposed this name for the mathematical quantity best corresponding to the "natural mental concept" of surprise at an outcome, and noted that the expected value of this quantity is precisely Shannon's definition of the uncertainty or entropy of the random variable.

More recently (following a decades-long period during which information-theoretic approaches received relatively little attention; see Luce, 2003), surprisal was introduced to quantify cognitive effort in incremental sentence processing by Hale (2001), in the particular context of a parsing strings into the tree structures of a probabilistic context-free grammar, under the intuition that a more strongly held belief (a more likely structure) takes more effort to rule out. This idea was further developed and generalized by R. Levy (2005, 2008a) who formalized this intuitive justification by equating surprisal with relative entropy between prior and posterior distributions—a precise quantification of the shift in understanding incurred by observing the word.

Language models as surprisal estimators Computing the surprisal value for a word in context requires an estimator of conditional probability of words, a function which, given some context, defines a probability distribution over words that might come in that context. Such a probabilistic model is known as a *language model* (LM)—this term simply refers to any model of the conditional

probability of words. Early 'n-gram' language models approximated the desired model of conditional probability by limiting the context-dependency to a fixed number n of preceding words (see overview in e.g., Jurafsky & Martin, 2024). Estimates of this n-gram probability based on relative frequency in a corpus of text form a useful way of computing the distributions that define such a language model, either at the level of characters, or of words (with somewhat surprising accuracy, given their simplicity, as noted in work as early as Markov, 1913; Shannon, 1948, 1951). But, accurately estimating the probability of words in context is a nontrivial task, given the combinatorial explosion of possible contexts. To get an intuition for why this is the case, consider the fact that for any particular sentence—for example, the one you are presently reading—it is quite possible that the particular sequence of words within it have never before existed in that precise order. This becomes even more likely as sentences are strung together into longer utterances. Simple relative frequency estimates are useless when encountering a word sequence that did not occur in the corpus they were computed from, with no way to distinguish between a continuation that is novel but plausible versus one that is truly implausible/impossible. The problem of how to improve such probability estimates led to the development of increasingly sophisticated smoothing techniques for statistical n-gram models (e.g., Nádas, 1984; Gale & Church, 1994; Kneser & Ney, 1995), but such models still have fundamental challenges in generalization to unseen data.

More recently, artificial neural network language models have become prevalent. Based on distributed representations of words (Hinton, 1986), the first such models were feed-forward networks (Bengio et al., 2000, 2003), and then using recurrent neural networks (Hochreiter & Schmidhuber, 1997; Sundermeyer et al., 2012; Mikolov, Yih, & Zweig, 2013), and, more recently, attention-based architectures (Bahdanau et al., 2016; Vaswani et al., 2017; Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023). These models make use of continuous-space representations that have the capability to assign similar contexts similar representations, effectively getting around some of the fundamental generalization challenges of statistical n-gram models. The success of modern pretrained large language models at the fundamental language modelling task (Zhao et al., 2023; Chang et al., 2024) has led to the modern era of large language models as the 'foundation models' forming the engine of a nascent artificial intelligence industry (Bommasani et al., 2022; Bubeck et al., 2023).

In this dissertation, in order to investigate theories of human language structure and processing, I will make critical use of pretrained large language models as statistical models which are trained to accurately predict words in context. Recent such models (e.g., Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023) provide the most accurate estimates of the statistical properties of language use currently available, and thus they provide valuable estimators of surprisal. At the same time, the theoretical and empirical results presented in this work also call into question the extent to which measures based purely on language model surprisal can be used

directly to explain phenomena of linguistic processing and linguistic structure.

The first and main focus of this dissertation is on information theoretic approaches to explaining incremental processing cost. When humans understand language—for instance, while reading—we incrementally observe words one after another, forming some idea of the meaning these words are intended to express, and updating this latent representation as we observe successive words of input. During comprehension, the amount of cognitive resources required to integrate each word is variable and context dependent: Sometimes a word will take more effort, sometimes less. How can we explain why are some words harder to process than others?

Chapter 1 provides an overview of the approach I take to answering this question, derives novel predictions within this framework, and presents a general introduction, situating this work with respect to prior research. I follow previous literature in framing the question from the perspective of Bayesian incremental inference, within the larger framework of rational analysis, 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. That is to say, the effort involved in processing a word is a function of the amount of information it contains. This idea has been encapsulated in the influential *surprisal theory* (Hale, 2001; R. Levy, 2008a), which posits that the difficulty of a word in context scales proportional to its surprisal. However, the standard arguments for surprisal theory rely on two assumptions: first, that surprisal is in fact a good measure of the amount by which expectations change upon observing the word, and second that the linking function between information gain and processing effort is linear. In this work, I investigate the ramifications of relaxing these two assumptions, leading to what I call *divergence theory*, a generalization of standard surprisal theory. I argue that this generalization has multiple benefits: It is more directly linked to some of the fundamental original motivations for surprisal theory, and offers a potential connection to a broad family of sampling-based inference algorithms which I argue are promising as models of incremental processing. It also provides the flexibility to capture empirical phenomena in human processing behaviour that the more restricted standard surprisal theory cannot.

I investigate two consequences of this generalization in the following two chapters, by relaxing each of the above assumptions one at a time. First, in **chapter 2**, I investigate 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 surprisal and processing cost. These predictions are tested and supported with results from nonlinear regressions fit to model the effect on human reading times of surprisal, as estimated by pretrained LMs. 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, something which cannot be explained under standard surprisal theory, but which can be accounted for by relaxing the assumption that justifies surprisal within the more general theory of processing cost as quantified by divergence between belief distributions. This intuition is evaluated with a self-paced reading study, using a variety of LMs as probability estimators.

In **chapter 4**, 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.

Linguistic dependencies are latent structures which are widely used to describe the grammatical relationships that govern how a sentence it interpreted, described by connecting words in a sentence to form a tree. These trees are constructed by connecting words that depend upon each other grammatically: for instance connecting a verb and its subject, or an adjective to the noun it modifies. According to such a system, each word in a sentence can be ascribed exactly one other word (called its head) upon which it is dependent, thus describing a directed tree structure over the entire sentence. At the same time, there are natural statistical relationships that exist between words in linguistic data, according to the distributional patterns of their use. The words in a sentence can be seen as observations of discrete-time stochastic process—a sequence of random variables taking on values from the set of words in the vocabulary. The statistical dependence between two particular observed words can be quantified using *pointwise mutual information*, which quantifies the amount of information about one word that is gained by knowing the other. For two observations x and y (of two discrete random variables), the pointwise mutual information (PMI; Fano, 1961) between them is defined as the log ratio of their joint probability to the product of their marginal probabilities. This can, equivalently, be expressed as a difference in surprisals: $\text{pmi}(x; y)$ is the amount by which the surprisal of x reduces when conditioning on y (or vice versa, swapping x and y; PMI is symmetric).

$$\text{pmi}(x; y) \coloneqq \log \frac{p(x, y)}{p(x)p(y)} = \underbrace{\log \frac{1}{p(x)}}_{s(x)} - \underbrace{\log \frac{1}{p(x \mid y)}}_{s(x \mid y)} \tag{2}$$

This quantity is zero when the two random variables are independent (since then the joint probability is by definition equal to the product of the marginals). It is positive when knowing y decreases

the surprisal of x, and negative when it increases it.

Thus we have two different kinds of structure between words in a sentence: those described by the head-dependent arcs of linguistic dependency trees, and the statistical dependencies that can be described in terms of pointwise mutual information. The statistical relationships between words are, intuitively, related to grammatical structure. Words which are grammatically dependent one upon the other would be expected to constrain each other's occurrence: One would not expect head and dependent to co-vary freely in terms of their occurrence in aggregate examples of language use. This intuitive connection between linguistic dependencies statistical dependence can be stated strongly as the following hypothesis: Connecting words into tree structures that maximize mutual information will result in recovering linguistic dependency structures. Such a hypothesis has been present implicitly in decades of work on unsupervised dependency parsing, and has even been stated explicitly in some work, including most notably Futrell et al. (2019). However, previous investigations of this hypothesis did not make use of language models which were capable of leveraging context in their estimates of pointwise mutual information. Motivated by this shortcoming of previous work, we obtain contextual pointwise mutual information values using probability estimates of pretrained masked language model, and extract tree structures which maximize this quantity. We then compare these trees to standard linguistic dependency trees on the same sentences, to understand the extent to which these two kinds of structure correspond. We find that while the correspondence between these two types of word-to-word dependence are consistently and substantially above a a random baseline, statistical dependence arcs only correspond to linguistic dependencies roughly as often as a simple baseline that connects adjacent words.

Chapter 5 concludes and presents discussion of limitations and further directions. Literature review is distributed among the content chapters. The definitions and notation for information-theoretic concepts such as surprisal and pointwise mutual information given above are repeated along with others which will be defined and discussed in the following chapters in a glossary at the end of the text.

Contributions of authors

The work represented in this dissertation was, to a lesser or greater extent, all the result of collaboration. The four chapters which form the body of this thesis constitute original scholarship and distinct contributions to knowledge. Chapter 1 presents a general introduction and theoretical background and also presents novel theoretical work forming a framework motivating the questions taken up in the second and third chapters. All writing in the first chapter is my own, conceptualized and revised based on conversations with and multiple rounds of feedback from Timothy O'Donnell, and one round of comments from Morgan Sonderegger. Chapters 2 and 4 consist of co-authored published papers and are reproduced here without meaningful modification. Chapter 3 consists of a co-authored manuscript in preparation for publication, not yet submitted. I was lead author of all three manuscripts. Individual intellectual contributions to each of these co-authored chapters are as follows.

Chapter 2 (along with accompanying supplemental material in appendix A) consists of an article titled "The Plausibility of Sampling as an Algorithmic Theory of Sentence Processing" which has been published in the journal *Open Mind: Discoveries in Cognitive Science* (Hoover et al., 2023, ©2023 MIT, CC BY 4.0). This article was co-authored with Morgan Sonderegger, Steven T. Piantadosi, and Timothy J. O'Donnell. The work was carried out by me under the supervision of MS and TJO at McGill, with STP of the University of California, Berkeley. All authors contributed to the conceptualization and research questions. I was responsible for all code, visualization, and initial writing, and I carried out all statistical analyses, with input from MS. TJO and MS contributed detailed comments and revision during multiple rounds of editing in the preparation of the final manuscript. I presented a preliminary version of this at Architectures and Mechanisms for Language Processing (Hoover et al., 2022). One additional appendix for this chapter, giving runtime derivations for probability-ordered search (appendix A.8), is included here that was not in the published version.

Chapter 3 (accompanied by appendix B) comprises new work, to be submitted for publication. This work was carried out under the supervision of Morgan Sonderegger and Timothy J. O'Donnell at McGill, with additional advising from Peng Qian at Harvard/MIT. I was responsible for the conceptualization and research questions and overall design of the study in close discussion with TJO and I developed the experimental design in conversation with PQ. I was responsible for statistical analyses, with input and advising from MS. I designed the stimuli, coded and administered the experiment, and contributed visualizations and all writing, with comments from MS and TJO.

Chapter 4 (with appendix C) consists of a paper titled "Linguistic Dependencies and Statistical Dependence" which has been published in the *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Hoover et al., 2021, ©2020 ACL, CC BY 4.0). This

paper was co-authored with Wenyu Du of the University of Hong Kong, Alessandro Sordoni of Microsoft Research, and Timothy J. O'Donnell. The work was carried out by me under the supervision of TJO and AS, and the three of us were jointly responsible for conceptualization and design of the project. WD contributed the contextualized pointwise mutual information (CPMI) estimates for the Ordered-Neuron LSTM models. I performed all other data-gathering, and contributed all code, conducted experiments, performed analysis, created visualizations, and contributed initial writing. TJO and AS also contributed detailed comments and revision to the draft in preparing the manuscript for publication.

I did not use generative AI to assist in any of the writing or revision of this dissertation.

1

Introduction: Processing cost as information gain

What makes a given word harder or easier to process, when it is encountered in context? Here we approach this question from the perspective that a word's processing cost is a consequence of the amount of information it communicates. This intuition has been a main underlying justification for expectation-based theories of processing cost (Hale, 2001, 2003b, 2016; R. Levy, 2008a, 2013; Futrell & Levy, 2017; Hahn et al., 2022). The hypothesis that cost can be quantified directly with the amount by which the comprehender's expectations change upon observing a word in context provides a connection to a broad family of algorithms for incremental inference as potential cognitive mechanisms for human processing, with the potential to explain how processing difficulty arises. Here we take this hypothesis seriously, and derive a computational framework for theories of processing cost as information gain, quantified with divergence between belief distributions. This approach can be seen as generalizing the standard hypothesis of surprisal theory—that the cost of an item scales proportional to its log inverse probability (Hale, 2001; R. Levy, 2008a). This generalization provides novel empirical predictions about processing cost in specific situations, depending on what the relevant meanings are, about which the observed word provides information. This chapter will lay out the conceptual and mathematical framing of processing cost as quantified by information gain, and describe the ways in which this account differs from standard surprisal theory, motivating theoretical and empirical questions to be explored in the following chapters.

1.1 Background

1.1.1 The rational analysis approach

The problem of explaining incremental processing, and how it is related to the distributional patterns of language use, can be situated within a larger scientific program aimed at understanding how human cognition is adapted to its environment. There is a rich history in cognitive science whereby aspects of cognition are modeled under the assumption of that they are in some way optimal (based on seminal work by, e.g., Newell, 1981; Marr, 1982; Pylyshyn, 1984; Shepard, 1987; Anderson, 1990). In particular, this optimality assumption can be defined in what Anderson termed the *general principle of rationality*—the assumption that "the cognitive system operates at all times to optimize the adaptation of the behaviour of the organism" (Anderson, 1990, p. 28). Note that, despite the name, this hypothesis needn't be interpreted to imply a claim about behaviour being the result of conscious application of rationally correct logical reasoning; it simply consists of the stipulation that a system behaves in an optimal, or approximately optimal, way. This hypothesis is sometimes justified via an evolutionary argument (evolution would be expected to exert pressure towards optimal system), but this also is not a necessary component of the framing.[1]

Starting from this hypothesis, Anderson outlined a metatheoretical strategy to understanding an aspect of human cognition, forming an approach referred to as *rational analysis*. In this approach, one first specifies the goals of the system, and formalizes a model of the environment, and of the information-processing problem that the system must solve. Then, one derives how an optimal system would behave, and examines these predictions empirically, refining the model if the predictions are not met (Anderson, 1990, 1991a, 1991b). Inspired by the arguments of Marr (1982, §1.2), this methodological approach is advocated from the perspective that if we want to understand the mechanisms for achieving the goals of a particular system, we should focus on understanding the nature of the problem being solved. Rational analysis gives a constrained framework within which to build computational level theories of the information-processing system. One benefit of this approach is that it de-emphasizes the primary importance of process-level theories, allowing the formulation and testing of explanations for why a particular relationship between behaviour and environment should exist. Then, after formulating a robust and empirically well-supported explanation of the optimal behaviour from this perspective, one may subsequently formulate algorithms as hypotheses for how this relationship arises at the level of cognitive processes and mechanisms.

Applying the rational analysis approach to the case of incremental processing means first

[1] While an evolutionary argument was an original motivation for proposing the principle of rationality, he also conceded that "evolutionary connections are as much of a hindrance as a help" (Anderson, 1991b), and suggested that perhaps a less contentions and perhaps clearer term would be the 'principle of adaptation.'

formalizing the nature of the problem of extracting meaning from linguistic input, and deriving the properties of an optimal system for achieving this goal. Then, we can investigate the degree to which human behaviour can be accurately predicted by this model, and ask questions about the mechanisms that can explain such behaviour.

1.1.2 Processing as incremental probabilistic inference

The task of language comprehension can be viewed fundamentally as an inference problem. Confronted with an utterance, received as sensory input, the comprehender must infer the intended meaning. From the perspective of probabilistic inference, the comprehender's inference can be modeled with a space of hypotheses about the intended meaning, within which different hypotheses are held to varying degrees of belief. Moreover, this inference task is naturally incremental: As a sequence of words is observed, the inference about the meaning can be sequentially updated. A comprehender need not wait until the entire utterance is complete to begin inferring the intended message.

This perspective formalizes the intuition that language comprehension on a basic level involves maintaining and updating expectations about what is being expressed, in real time during processing. For a rational comprehender, a given observation will cause a shift in the belief distribution, favoring those hypotheses that better explain the observation, and disfavoring those that do not. One need not be committed to the perspective that human behaviour in language processing behaves in a strictly rational manner. However, to the extent that it does, probabilistic inference gives a precise framework in which to define what it means for a system to be optimal with respect to the environment: An optimal solution to the inference problem is one which, on average, makes accurate predictions about the observations. The optimal way to update beliefs in the light of a new observation is to redistribute weight on hypotheses according to the laws of probability (Keynes, 1921; de Finetti, 1972), shifting a prior distribution to a posterior distribution in what has become known as the Bayesian approach to inference (Earman, 1992; Howson & Urbach, 2006). Such probabilistic models have been productively employed to characterize linguistic comprehension (see e.g., Chater et al., 1998; Jurafsky, 2003; Chater & Manning, 2006; Kuperberg & Jaeger, 2016; Degen, 2023), with the comprehender performing inference over the intended meaning as successive pieces of input are received.

In this work we approach comprehension from this perspective, treating it explicitly as an incremental inference problem, wherein the latent variable (meaning) is inferred from the stream of observations (words), within the larger framework of rational analysis.

1.1.3 Surprisal theory

In recent decades, a substantial body of research in the field of computational psycholinguistics has pursued expectation-based theories of human processing cost (e.g., Hale, 2001, 2003b, 2014; Narayanan & Jurafsky, 2001, 2004; R. Levy, 2008a, 2013; Smith & Levy, 2013; Rasmussen & Schuler, 2018; Futrell et al., 2020; Hahn et al., 2022), which can be situated within the framework of comprehension as incremental probabilistic inference. Perhaps the most prominent among such approaches has been that of *surprisal theory* (Hale, 2001; R. Levy, 2008a), which proposes that a word's incremental processing cost can be quantified as Shannon information, or surprisal—the inverse log probability of the word, conditioned on the previous words in the sentence (and possibly also conditioned on additional extra-sentential or even non-linguistic context). This proposal can be stated explicitly as the following hypothesis.

Hypothesis 1.1 (standard surprisal theory)**.** The processing cost of a word $\breve{w}$ in context increases proportional to its surprisal. That is,

$$\text{cost}(\breve{w}) = \beta \cdot s(\breve{w}) \tag{1.1}$$

where β is some positive constant, and surprisal is defined as $s(\cdot) \coloneqq -\log p(\cdot \mid \text{context})$.[a]

[a] Surprisal as defined here is by definition context-dependent, and could be more explicitly written $s(\cdot \mid \text{context})$. I choose to suppress the context variable here and for the rest of this chapter, for brevity.

A weaker version of this hypothesis relaxes the assumption that the relationship is linear.

Hypothesis 1.2 (general surprisal theory)**.** The processing cost of a word $\breve{w}$ in context increases monotonically with surprisal. That is,

$$\text{cost}(\breve{w}) = f(s(\breve{w})) \tag{1.2}$$

where f is a monotonically increasing function.

A special case of this weaker version of surprisal theory has also been proposed with the stipulation that f belongs to some particular family of monotonically increasing functions.[2] In this chapter I will refer to the stronger version of the hypothesis that explicitly assumes the relationship is linear as standard surprisal theory, and will refer to the weaker hypothesis with an arbitrary linking function as general or nonlinear surprisal theory. Examining arguments and evidence about the form of this linking function within general surprisal theory will be the focus of chapter 2.

[2] For example, a function of the form $s(\breve{w})^k$ for some k is discussed in work such as R. Levy (2005), Meister et al. (2021), and Xu et al. (2023), which has been motivated by the Uniform Information Density hypothesis (Aylett & Turk, 2004; R. Levy & Jaeger, 2006; T. H. Clark et al., 2023)—see discussion in §2.3.3.

1.1.4 Justifications for surprisal theory

A number of related theoretical justifications have been given as motivations for the quantification of processing cost with surprisal (as reviewed and discussed in, e.g., R. Levy, 2013). Each has its own assumptions and framing, but generally these justifications can be classified into two types: update-cost arguments, and time-optimality arguments, all within the perspective of rational analysis, outlined above.

Reallocation cost arguments

One primary type of justification for surprisal theory comes from the perspective that the cost of processing a word derives from cognitive effort involved in reallocation of resources as the comprehender changes their expectations about the interpretation of the utterance. This type of argument was first suggested explicitly by Hale (2001; 2003a, Ch. 6, inspired by Attneave, 1959), who framed surprisal as a quantification of the cognitive effort associated with 'disconfirming' structures that were inconsistent with the observed word, when parsing into a probabilistic grammar. This suggestion was based on the intuition that the processing effort is quantified by the amount of probability mass that is shifted off of parses that are ruled out upon observing it.

This intuition was formalized and developed by R. Levy (2005, 2008a), who proposed that this shift be quantified specifically as the relative entropy (also known as Kullback-Leibler divergence) between distributions over latent representations of meaning (parses) before and after observing the word, in a probabilistic generative model. R. Levy showed that this relative entropy is in fact precisely equivalent to surprisal, under the critical assumption that the parses consist at least in part of observable words. That is, he assumed the latent representations over which the belief distributions range consist of structures (parses) each of which contains an observable string (the yield of the parse). This string may either match the observed word at the appropriate point (in which case the parse is consistent with the observation), or not (in which case it is not consistent with the observation and has probability zero under the posterior). This assumption that there is a deterministic relationship between the latent structures and the observable words is necessary for R. Levy's derivation of the equivalence between relative entropy and surprisal. I will review this derivation in the following section, and propose that the determinism assumption may not always be warranted.

The reallocation-cost arguments also suggest a natural connection for surprisal theory with potential algorithmic theories of processing: An inference algorithm whose complexity scales with the size of the shift it must effect between states before and after encountering a word would be able to intrinsically predict processing effort.

Time-optimality arguments

Other justifications for surprisal theory include those based on the assumption that the cognitive mechanisms involved in comprehension are optimized for the task of processing input as quickly as possible—that is, that the comprehender is a rational agent possessed of a cognitive apparatus that implements this optimal efficiency. One such optimality justification (developed in Smith & Levy, 2008a, 2008b, 2013) starts from the assumption that there is some linking function which describes an item's processing cost as a function of its probability, and that this function is independent of granularity at which items are considered (this is what they refer to as the "scale-free assumption", which says that the conjectured linking function remains the same whether considering phrases, words, morphemes, or so on), and proves that such a function is linear in log-probability.

Another such justification comes from the literature on Bayesian modelling of word recognition (Norris, 2006, 2009). Norris's argument starts from the implicit assumption that lexical identification is the key contributor to difficulty, and additionally assumes that this process can be modeled with an algorithm of sequential sampling from the visual input, until a threshold of certainty about word identity is reached (via the sequential probability ratio test model for the desicion process; Wald, 1947; Barnard, 1946). This work claims the average time complexity for word recognition is linear in negative log probability. Norris (2006, 2009) points to Baum and Veeravalli (1994) for a derivation, and Adelman and Brown (2008) for a similar derivation in a simplified setting, though this connection between sequential perceptual sampling and processing cost is not developed in detail. One important aspect of these derivations is that they are essentially all about lexical decision (deciding whether the input is or isn't a word)—they leave open the question of how to describe the cost incurred in integrating information about meaning contained in a word or across multiple words, or assume it is negligible or constant.

1.1.5 Shortcomings of standard surprisal theory

The predictions of standard surprisal theory have been broadly supported by a large number of empirical studies (e.g., Demberg & Keller, 2008; R. Levy, 2008a; Smith & Levy, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2023; Shain et al., 2024).[3] Yet, despite this general empirical success, a number of recent studies have begun to cast doubt on the degree to which surprisal alone can explain the patterns of human processing difficulty. For instance, when considering constructions that are difficult for humans, such as syntactically ambiguous constructions, there is evidence that a linear relationship with surprisal cannot account for the degree of human processing difficulty (van Schijndel & Linzen, 2021; Arehalli et al., 2022; Huang et al., 2024). Other work has independently raised empirical and theoretical questions about surprisal theory's assumption of a linear linking function (Brothers & Kuperberg, 2021; Meister et al., 2021; Xu et al., 2023). The topic of the

[3] See §2.3 for further discussion of this literature.

linking function will be explored in depth in chapter 2, with the finding that algorithms based on sampling predict a superlinear relationship. Detailed discussion of these predictions is deferred to that chapter.

Surprisal theory also makes some counterintuitive predictions that have not yet received attention in the broader literature, but which I propose are potentially problematic. Namely, it is intuitively plausible that a particular item of linguistic input might be highly unpredictable (and therefore have high surprisal) without having high processing cost, directly contradicting standard surprisal theory. One way this can happen is when the unpredictability of the item is mainly the result of a production error which is easily correctable in comprehension. When this error does not cause any meaningful change in the expected meaning, it should not incur much additional effort to process (under resource reallocation cost justification discussed above, there simply is not much reallocation required to process such an item). This stands in contrast to the case where a piece of input is unpredictable because it introduces surprising information which incurs a large change in the inferred message, requiring cognitive resources to integrate. This intuition is in direct contradiction to standard surprisal theory, which would be forced to predict high cost for any item which is unpredictable, however I will propose below that this potential problem can be addressed within a more general version of the theory. I will further develop this proposal in the next section, and explore empirical predictions by looking at processing of text with typographical errors in chapter 3.

Before examining these points in detail (which will be the topics of subsequent chapters), it is worth first understanding how the underlying motivations behind surprisal theory can inspire refinements and generalizations of its central hypothesis. Formalizing the components of such a generalization will be useful in understanding what aspects of surprisal theory can be modified by the adoption of alternative assumptions, without abandoning the fundamental motivation.

1.2 Quantifying information gain

As outlined above, surprisal theory offers a simple formalization of the underlying intuition that the amount of information contained in a word is important predictor of its processing cost, but also makes additional assumptions which are worth examining and may lead to undesirable implications. In this section I will generalize surprisal theory to recenter the central justification and retain this crucial intuition, and will tease out these assumptions explicitly.

Let us take a step back and start from the intuition that a word's cost is intrinsically related to the information that it contributes. This can be stated as the following broad conjecture.

> **Hypothesis 1.3** (information gain)**.** The processing cost of a word $\breve{w}$ is an increasing function

of the amount of information it contributes to the comprehender. That is,

$$\text{cost}(\breve{w}) = f\big(\text{information-gain}(\breve{w})\big) \tag{1.3}$$

for a nonnegative monotonic increasing function f, and some appropriate measure of information gain.

This conjecture is meant to encapsulate the intuition that processing cost reflects the amount of information a word contributes, motivated by the idea that if the goal of comprehension is to incorporate new information, then it is plausible that the quantity of resources consumed to do this will scale as a function of the amount of information gained. A measure of the information gain associated with an observation should be nonnegative and should be zero when the observation is certain to occur. Operationalizing this conjecture requires both specifying a method for quantifying information gain, as well as proposing the form of the linking function f. Yet at this stage we needn't commit to a stance on precisely what the relevant representations are or how they are maintained, or how integration of new information is carried out.

1.2.1 Setting up the inference problem

In framing a computational theory of processing cost in terms of surprisal, we have thus far only referred to the probability of the observed word $\breve{w}$, which may be thought of as the outcome of a random variable, W, ranging over possible observations in the context.[4] In order to formalize a relevant notion of the information-gain associated with this observation, let us introduce another random variable, Z, ranging over some set $\mathcal{Z}$ of latent structures (representing meanings or interpretations) about which the words are informative. In this setting, the task of comprehension for a single word is the task of inferring the intended meaning Z given an observed outcome $\breve{w}$ of W, in context. For what follows, we will leave the nature of the meaning structures comprising the set $\mathcal{Z}$ intentionally unspecified (for example, the individual meanings might be modelled as consisting of linguistic structures such as parse trees, continuous vectors, discourse representation structures, or any other particular space of meaning representations). In this probabilistic inference setting, a comprehender's beliefs about the intended interpretation of an utterance are represented by a probability distribution over this space of latent structures.

In a general joint model $p_{Z,W}$, the Bayesian update about the latent Z upon observing some outcome $\breve{w}$ of W can be described as starting with the **prior** distribution, p_Z, representing the comprehender's beliefs before encountering the word $\breve{w}$, and a **likelihood** function, mapping any

[4] Remark on notation: In this chapter and what follows, I mark an outcome of a random variable with the breve diacritic (˘) as a visual reminder that it represents a particular fixed outcome ($\breve{w}$ is a specific observed word). This notation serves for instance to clarify that an expression such as $p(\breve{w} \mid z)$ is intended to represent the likelihood of z—that is, the conditional probability of thought of as a function of z, with $\breve{w}$ fixed. This notation is purely cosmetic, and the reader may ignore the diacritic without consequence.

point in the meaning space to the probability of outcome $\breve{w}$'s being observed given that meaning. Bayes' rule then gives the **posterior** distribution, $p_{Z|\breve{w}}$, as proportional to the prior reweighted by the likelihood.

Achieving the Bayesian update consists of forming a representation of this posterior belief distribution, given the observation. We can think of this update as the result of a procedure which takes as input an observation $\breve{w}$ and, starting with a prior belief distribution p_Z, outputs a new belief distribution which represents the posterior.

$$\breve{w}, p_Z \rightsquigarrow p_{Z|\breve{w}}$$

Or, somewhat more generally, we could consider a process that approximates the posterior using some more general **proposal** distribution q_Z over $\mathcal{Z}$ (which is not necessarily the prior).

$$\breve{w}, q_Z \rightsquigarrow p_{Z|\breve{w}}$$

In this setting, the goal of the inference procedure is to compute, or approximate, the posterior, making use the prior, or more generally, some proposal distribution q_Z. In formulating a computational theory of processing effort, we will be interested in the difficulty of this estimation process. From an information-theoretic perspective this can be quantified with a measure of how bad the proposal is as an estimate of the posterior, as a quantification of the information that is gained in making the update. A worse proposal means more work must be done in order to transform it into a good approximate of the the posterior. I will start by formalizing information as relative entropy using the prior, and then generalize to an arbitrary distribution in the subsequent section.

1.2.2 Introducing divergence theory

Consider the update from prior to posterior. To formalize the notion of information gain in this setting requires some function that quantifies how much a given observation $\breve{w}$ affects the distribution over Z. I will give the name *divergence theory* to the general hypothesis that processing cost increases with increasing belief divergence.

In the information theoretic setting, a natural choice for divergence function is the Kullback-Leibler (KL) divergence, also known as the relative entropy or discrimination information (Kullback & Leibler, 1951). The KL divergence is a standard quantification of information gain, in that for two distributions p, q the quantity $D_{\mathrm{KL}}(p \parallel q)$ can be interpreted as the amount of information that would be 'lost' by using distribution q when the true distribution is p, for example in the setting of statistical model comparison (e.g., Burnham & Anderson, 2004) or in terms of code length in information theory (Cover & Thomas, 2006).[5]

[5]The divergence $D_{\mathrm{KL}}(p \parallel q) := \mathbb{E}_p[\log \frac{\mathrm{d}p}{\mathrm{d}q}]$ is defined only if p is *absolutely continuous* with respect to q. A probability

The divergence $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) := \mathbb{E}_{p_{Z|\breve{w}}}\left[\log \frac{\mathrm{d}p_{Z|\breve{w}}}{\mathrm{d}p_Z}\right]$ quantifies the information gained by revising beliefs about Z from a prior distribution p_Z to a posterior distribution $p_{Z|\breve{w}}$ in a Bayesian inference setting. Another way to understand this relative entropy is that it consists of the expected reduction in surprisal of $\breve{w}$ that results under the posterior, as will become clear in derivation given in the next section.

Note, while here we focus on measuring cost with divergence in belief distributions, there is at least one salient alternative information-theoretic function measuring information gain of an observed word $\breve{w}$ about a latent meaning variable Z, that has been proposed in this literature: The reduction in the entropy of Z under $p_{Z|\breve{w}}$ compared to under p_Z (proposed as a predictor of processing cost in Hale, 2003b, ch. 3). For the moment I will just point out that entropy reduction has a similar interpretation to KL divergence, with the important difference being that it can be negative. In this work I will focus on KL divergence, but see §1.4 for discussion of entropy reduction as an alternative measure of information gain.

With the choice of KL divergence as our measure of information gain, we have the following hypothesis about incremental processing cost, which I will refer to as *KL divergence theory* from the prior, or simply *KL theory*.[6]

Hypothesis 1.4 (Divergence theory, using KL from the prior)**.** The processing cost of a word $\breve{w}$ is a monotonic increasing function of the amount of information it communicates, as quantified by the KL divergence between the posterior distribution $p_{Z|\breve{w}}$ and the prior distribution p_Z.

$$\text{cost}(\breve{w}) = f\left(D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)\right) \tag{1.4}$$

where f is a monotonically increasing function.

To keep this hypothesis general, I have not proposed a specific linking function f, beyond requiring that it be monotonically increasing, but a full theory would require specifying the form of this linking function.

Note that KL divergence has the desirable properties of (i) being nonnegative, and (ii) vanishing when the distributions are identical. It is also important to note that while it may be convenient to

measure p is said to be absolutely continuous with respect to another probability measure q on the same space, written $p \ll q$, iff $q(\zeta) = 0$ implies that $p(\zeta) = 0$, for any measurable set ζ. In the discrete case, absolute continuity of p with respect to q means simply any outcome given zero probability under q must be also zero probability under p; thus this guarantees against dividing by zero in the computation of KL divergence. Note that in the case of the KL divergence between prior and posterior, Bayes' rule guarantees that the posterior is absolutely continuous with respect to the prior, so the divergence $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$ is well defined.

[6]Note, there are other divergence functions that can be defined, which provide alternative ways of measuring the difference between two distributions. Choosing some other divergence function between distributions (and substituting in place of KL divergence in the cost hypothesis equation) would give an alternative realization of divergence theory. I do not explore such alternatives here, but see the box below, in §1.3.1.1, for brief motivation and discussion of alternative probability divergence functions.

think of KL divergence as quantifying the "distance" between two distributions, it does not satisfy the requirements of a distance metric: Importantly, it is not a symmetric function. Given this fact, it is natural to ask whether the choice of order of the arguments in $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$ is the right one. In fact, the version of this divergence with the arguments transposed so the prior is the first argument—that is, $D_{\mathrm{KL}}(p_Z \,\|\, p_{Z|\breve{w}})$—has been proposed as a quantification of '(Bayesian) surprise' (proposed in Baldi, 2002; developed in Baldi & Itti, 2010, with applications to vision), based on similar intuitions as those presented here. However, their motivation to choose the transposed version seems to be mostly one of convenience, without a particular theoretical motivation (and indeed in some subsequent work, the same authors use the version with posterior as the first argument: e.g., Itti & Baldi, 2009). Here, in hypothesis 1.4, I explicitly propose to use the KL with the posterior as the first argument, following previous arguments which were used to motivate surprisal theory (since R. Levy, 2005), for two reasons. In this direction, it is a proper quantification of the amount of information gained when moving from from prior to posterior, and is defined even when the posterior has a smaller support, as is often plausible (when some amount of the prior support is ruled out by the observation). Additionally, this KL divergence has direct relevance to the complexity of sampling algorithms, as will be explored below (§1.3).

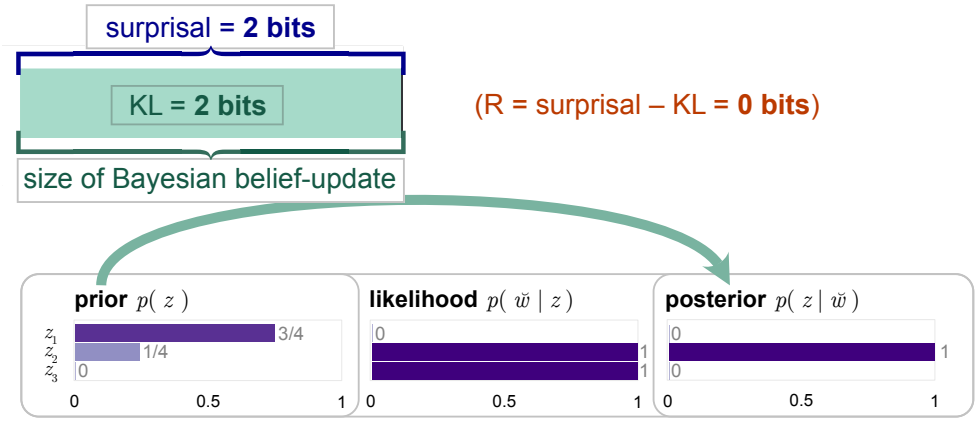
Decomposing KL divergence and relating to surprisal

Without making any additional assumptions, the KL divergence between posterior and prior can be decomposed in the following way.

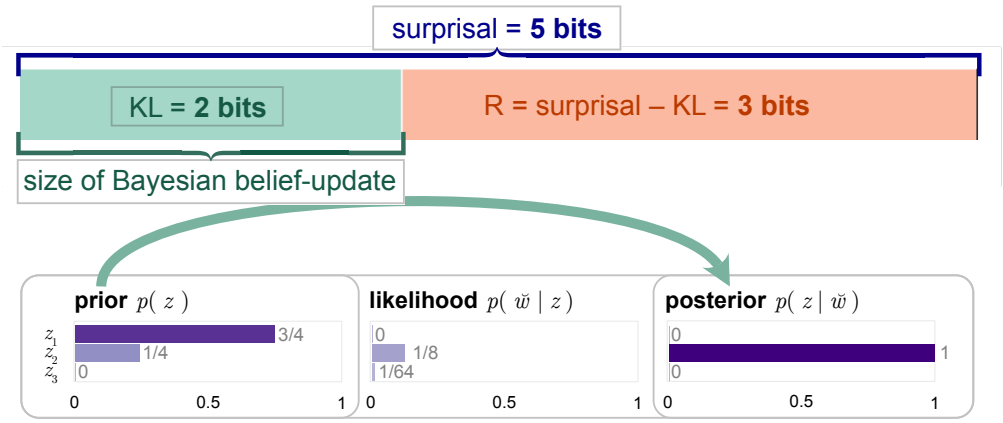
$$\begin{aligned} D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) &:= \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(z \mid \breve{w})}{p(z)}\right] = \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(z, \breve{w})}{p(z)p(\breve{w})}\right] \\ &= \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(\breve{w} \mid z)}{p(\breve{w})}\right] \\ &= \underbrace{\log \frac{1}{p(\breve{w})}}_{s(\breve{w})} - \underbrace{\mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{1}{p(\breve{w} \mid z)}\right]}_{:= R(\breve{w})} \end{aligned} \tag{1.5}$$

So, in general, this divergence consists of the surprisal, $s(\breve{w}) \geq 0$, minus a second nonnegative term which I will denote as $R(\breve{w})$ and refer to as the *reconstruction information*[7]—the expectation under the posterior of the negative log of the likelihood function, that is, the expected surprisal of $\breve{w}$, conditioning on $z \sim p_{Z|\breve{w}}$. I call this quantity the reconstruction information because it is

[7] As far as I am aware, there is no existing standard name for this information theoretic quantity in the context of Bayesian inference, hence the introduction of this novel terminology. This choice of name for the expected negative log likelihood under the posterior reflects the mathematical similarity of this term to what is sometimes called the '(negative) reconstruction error' in variational inference (in which context the expectation is taken under a variational approximation of the posterior—see, e.g., Blei et al., 2017; Liang et al., 2018).



(a) In this example, the likelihood is a binary function over the support of the prior, thus surprisal is equal to KL at 2 bits.



(b) Here, as above, the KL is 2 bits, but the likelihood is *not* a binary function over the support of the prior, and surprisal exceeds KL at $-\log(\frac{1}{4} \cdot \frac{1}{8}) = 5$ bits.

Figure 1.1: Diagram illustrating surprisal $s(\breve{w})$ partitioned into the sum of two nonnegative components: $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$, which quantifies the amount by which the observation causes belief to change (information gain); and the remainder, $R(\breve{w})$—which quantifies the information irrelevant to belief update.

the amount of information that would be needed under the posterior distribution to describe the precise value of $\breve{w}$ that was observed (that is, to 'reconstruct' the observation). Note that if the likelihood of $\breve{w}$ is uniformly 1 under the posterior, then this quantity is minimized to zero, and KL is equal to surprisal (as is assumed in justifications for surprisal theory; R. Levy, 2005, 2008a). The non-negativity of these terms gives that $0 \leq R(\breve{w}) \leq s(\breve{w})$.

Rearranging the above equation to solve for surprisal makes it clear that this derivation essentially describes a way of partitioning the bits of Shannon information (surprisal) of $\breve{w}$ into two nonnegative terms:

$$s(\breve{w}) = D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) + R(\breve{w}) \tag{1.6}$$

The first term, $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$, quantifies the size of the Bayesian belief update induced upon observing $\breve{w}$, and therefore the second term, $R(\breve{w})$, quantifies the remaining bits of Shannon information that do not contribute to belief update.

This decomposition of surprisal is illustrated with two examples in fig. 1.1, with the magnitude of surprisal indicated by a horizontal bar partitioned into parts representing $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$ and $R(\breve{w})$. These examples both use the same example prior distribution over a discrete space $\mathcal{Z} = \{z_1, z_2, z_3\}$.The difference between the examples is in the use of two different likelihood functions, where in both cases the resulting posterior is the same—in each case, after the Bayesian update all probability mass is concentrated on one outcome, z_2, resulting in a divergence of 2 bits from the prior distribution. The values in these illustrations were chosen to make the information theoretic quantities simple to compute, while illustrating the non-equivalence of KL and surprisal. In the first example (fig. 1.1a), the likelihood is binary—either consistent or inconsistent with the observation—and thus the surprisal is equal to KL. By contrast, in the second example (fig. 1.1b), the likelihood less than 1 for some (in fact all) values of Z supported in the posterior, and this results in a situation where the surprisal of the observation is substantially larger than the first example, while the KL is the same.

As illustrated in these examples, the decomposition of surprisal in eq. 1.6 expresses the fact that in this general setting surprisal $s(\breve{w})$ forms an upper bound on the amount by which the posterior distribution after the Bayesian update diverges from the prior. In terms of quantifying information gain, we can interpret $R(\breve{w})$ as a quantification of the how many bits of surprisal are, in a sense, wasted. When $R(\breve{w})$ is negligible, nearly all bits of information measured by surprisal contribute to belief update, and surprisal is thus a good measure of information gain. But when $R(\breve{w})$ is large, despite an observation containing a large amount of Shannon information, it does not result in a correspondingly large shift in Bayesian belief distributions about Z.

1.2.3 Divergence theory using a proposal distribution

Now, in the final step of generalization, let us consider the more general case when the distribution we compare to the posterior is not necessarily the prior. For the present, I will simply state this generalization as a mathematical fact, but it becomes relevant in the situation where the distribution we are interested in comparing to the posterior is not the naïve prior, but some alternative proposal distribution which might be strategically chosen, as will be discussed in the next section.

To state this generalization, for a fixed outcome $\breve{w}$ of W, let $q_{Z;\breve{w}}$ be a **proposal** distribution (which may depend on $\breve{w}$). This could be any distribution over Z, with one requirement: For the divergence $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}})$ to be well-defined, assume the posterior is absolutely continuous with respect to the proposal.

Hypothesis 1.5 (Divergence theory, using KL from a proposal)**.** The processing cost of a word $\breve{w}$ is a monotonic increasing function of the KL divergence between the posterior distribution $p_{Z|\breve{w}}$ and a proposal distribution $q_{Z;\breve{w}}$.

$$\text{cost}(\breve{w})_q = f\left(D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}})\right) \tag{1.7}$$

where f is a monotonically increasing function.

As before, it is worth noting that while I have chosen to measure the difference between distributions with KL divergence in particular, an alternative probability divergence function could be substituted for KL in this definition, giving an alternative realization of divergence theory—see the box at the end of §1.3.1.1 for a brief discussion of such alternatives.

Decomposing KL divergence from a proposal distribution

Similar to above (eq. 1.5), the KL divergence from a proposal[8] can be decomposed in the following way, introducing one additional term into the decomposition.

$$\begin{aligned} D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}}) &\coloneqq \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(z \mid \breve{w})}{q(z;\breve{w})}\right] = \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(z \mid \breve{w})}{p(z)} \frac{p(z)}{q(z;\breve{w})}\right] \\ &= \underbrace{D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)}_{\text{(eq. 1.5)}} + \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(z)}{q(z;\breve{w})}\right] \\ &= \underbrace{\log \frac{1}{p(\breve{w})}}_{s(\breve{w})} - \underbrace{\mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{1}{p(\breve{w} \mid z)}\right]}_{\coloneqq R(\breve{w})} - \underbrace{\mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{q(z;\breve{w})}{p(z)}\right]}_{\coloneqq D_{q_{Z;\breve{w}}}} \end{aligned} \tag{1.8}$$

This additional term, labelled $D_{q_{Z;\breve{w}}}$, which I will call the *proposal advantage*, has the form of a difference between two KL divergences (note this difference can be positive or negative).[9]

$$D_{q_{Z;\breve{w}}} \coloneqq \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{q(z;\breve{w})}{p(z)}\right] = D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) - D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}}) \tag{1.9}$$

[8] We could also consider the proposal not as a single distribution but as a *family* of distributions, as used in variational inference techniques (Blei et al., 2017), given the identity of $\breve{w}$ and defined by some additional parameters. It is worth noting in this context that the KL divergence being discussed here is the same quantity which is minimized in the expectation propagation (EP) approach for variational inference (introduced in Minka, 2001; see Bishop, 2006, §10.7; Opper, 2015), not the divergence generally used in the expectation maximization algorithm or other "variational Bayes" techniques (Attias, 1999; Wainwright & Jordan, 2007; Blei et al., 2017), which has the arguments transposed.

[9] This can be seen as an example of the general identity $D_{\text{KL}}(p \,\|\, r) - D_{\text{KL}}(p \,\|\, q) = \mathbb{E}_p\left[\log \frac{\mathrm{d}q}{\mathrm{d}r}\right]$, for any probability measures p, q, r, on the same space with $p \ll r$ and $p \ll q$.

Recall that one way to interpret the relative entropy $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$ is as quantifying the expected surprise that would result from drawing samples from p_Z if the actual distribution is $p_{Z|\breve{w}}$. From this same perspective, the quantity $D_{q_{Z;\breve{w}}}$ represents the expected reduction in surprise resulting from using $q_{Z;\breve{w}}$ instead of p_Z, when the actual distribution is $p_{Z|\breve{w}}$. This is to say, the term $D_{q_{Z;\breve{w}}}$ quantifies how much better the proposal $q_{Z;\breve{w}}$ is than the prior, as an estimate of the posterior.[10] This quantity vanishes when the proposal is equal to the prior, and is positive when the proposal is better than the prior, with an upper bound at $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) = s(\breve{w}) - R(\breve{w})$, which corresponds to when the proposal equals the posterior. It is negative when the proposal is worse than the prior, with no lower bound.

$$-\infty \leq D_{q_{Z;\breve{w}}} \leq D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) \tag{1.11}$$

This captures the fact that, at best, the proposal is equal to the posterior, resulting in zero cost, and at worst it can be arbitrarily far from the posterior, resulting in an arbitrarily large cost.

In the situation where the the proposal is at least as good as the prior, and thus $D_{q_{Z;\breve{w}}}$ is positive, KL theory from the proposal has an interpretation identical to that of KL theory from the prior: Surprisal is an upper bound on the cost, with the tightness of this bound specified by the amount of 'wasted' bits, which are quantified precisely as $R(\breve{w}) + D_{q_{Z;\breve{w}}}$.

1.2.4 Summary of hypotheses about processing cost

Table 1.1 summarizes all the hypotheses about how to quantify information gain discussed in this section. These hypotheses form a hierarchy, each contained within one another in terms of generality. Starting with only the broad claim that cost is an increasing function of some notion of information gain associated with an observation hypothesis 1.3, divergence theory consists of the hypothesis that information gain be quantified as a probability divergence between the posterior distribution over meanings given the observation, and a (potentially strategically chosen) proposal distribution. In this work we specifically choose the information-theoretically interpretable KL divergence to operationalize this hypothesis (hypothesis 1.5). Additionally assuming that the proposal is simply the Bayesian prior gives hypothesis 1.4—that cost is a function of the magnitude of the Bayesian belief update from prior to posterior. General (that is, nonlinear) surprisal theory (hypothesis 1.2) results from including the additional assumption that $R(\breve{w})$ is zero, and standard linear surprisal theory (hypothesis 1.1) results from the final assumption that the linking function is linear.

[10] Equivalently, it can be viewed as measuring the reduction in *cross-entropy* resulting in using the proposal instead of the prior, with a similar interpretation.

$$D_{q_{Z;\breve{w}}} = \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{1}{p(z)}\right] - \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{1}{q(z;\breve{w})}\right] = \mathrm{H}(p_{Z|\breve{w}}, p_Z) - \mathrm{H}(p_{Z|\breve{w}}, q_{Z;\breve{w}}) \tag{1.10}$$

cost hypothesis	$\text{cost}(\breve{w}) =$		assumptions
information gain (hypothesis 1.3)	$f($ information-gain$(\breve{w})$ $)$		cost scales with information gain
divergence theory	$f($ divergence$(p_{Z\|\breve{w}} \,\|\, q_{Z;\breve{w}}))$	$\wedge$	info. gain quantified as a divergence from proposal $q_{Z;\breve{w}}$ to posterior $p_{Z\|\breve{w}}$
KL from proposal (hypothesis 1.5)	$f($ $\underbrace{D_{\text{KL}}(p_{Z\|\breve{w}} \,\|\, q_{Z;\breve{w}})}_{s(\breve{w})-[R(\breve{w})+D_{q_{Z;\breve{w}}}]\ (1.8)}$ $)$	$\wedge$	info. gain quantified specifically with KL divergence
KL from prior (hypothesis 1.4)	$f($ $\underbrace{D_{\text{KL}}(p_{Z\|\breve{w}} \,\|\, p_Z)}_{s(\breve{w})-R(\breve{w})\ (1.5)}$ $)$	$\wedge$	proposal $q_{Z;\breve{w}}$ is the prior p_Z
general surprisal (hypothesis 1.2)	$f($ $s(\breve{w})$ $)$	$\wedge$	$R(\breve{w})$ is zero (likelihood is binary everywhere in support of prior)
standard surprisal (hypothesis 1.1)	$\beta\, s(\breve{w})$	$\wedge$	linking function f is linear

Table 1.1: Hierarchy of processing cost hypotheses. From top to bottom the hypotheses are ordered from most to least general: Starting with the broad information gain conjecture, and iteratively adding assumptions, we arrive first at KL theory, and ultimately to standard surprisal theory.

1.3 Justifications for divergence theory

1.3.1 Algorithmic complexity cost

As outlined in the background section above, a model developed within the conceptual framework of rational analysis has important interpretability advantage over an empirical model that simply fits free parameters without the explanatory power of such a justification. However, while a successful computational-level model may be seen as explaining why a pattern of behaviour should exist—because it is optimally adapted for its environment—it does not necessarily offer any clues about how this optimal behaviour is in fact achieved. For that, an algorithmic or process-level theory must be provided.

Such a process-level explanation for surprisal theory is conspicuously absent. One promising family of algorithms that may naturally require more work when an observation is less expected are those that involve guessing or sampling from the prior, in order to approximate the posterior. Investigating the complexity of such algorithms reveals that in fact they provide a more direct justification for the computational-level hypothesis that cost scales with KL divergence.

1.3.1.1 Runtime of importance sampling

An important potential process-level justification for divergence theory can be found in the runtime complexity of sampling-based algorithms for approximate inference. Say we want to approximate

some target distribution p using samples from some other distribution q. If we use samples from q to approximate p, using importance sampling (Doucet et al., 2001, §1.3.2; Chopin & Papaspiliopoulos, 2020, ch. 8), the number of samples required for an accurate approximation can be shown to scale exponentially in the KL divergence from q to p:

$$\#\text{samples}_{\text{IS}(p \leftarrow q)} \approx e^{D_{\text{KL}}(p \,\|\, q)} \tag{1.12}$$

This complexity result is due to Chatterjee and Diaconis (2018), who prove in a rather general setting that a sample size that is exponential in this relative entropy is sufficient (and also necessary, under some further assumptions) for importance sampling's approximation error to be close to zero with high probability.

Conditions for the importance sampling result

Technically, Chatterjee and Diaconis's result (eq. 1.12) requires that the log density of p with respect to q is likely concentrated around its expected value, $\mathbb{E}_p[\log \frac{\mathrm{d}p}{\mathrm{d}q}] = D_{\text{KL}}(p \,\|\, q)$, where $\frac{\mathrm{d}p}{\mathrm{d}q}$ denotes the density (Radon-Nikodým derivative) of p with respect to q—in the discrete case, this is simply the ratio of probability mass functions. Roughly, this is the requirement that the expected variance in importance weights is small.[a]

[a] More precisely, their result says that in order to bound the L^1-error of the estimate close to zero with high probability, a sample size of $\exp(D_{\text{KL}}(p \,\|\, q) + \mathcal{O}(s))$ is sufficient, where s is the typical amount by which $\log \frac{\mathrm{d}p}{\mathrm{d}q}(Z)$ fluctuates around its expected value, $D_{\text{KL}}(p \,\|\, q)$. Moreover they show that a sample size of $\exp(D_{\text{KL}}(p \,\|\, q) - \mathcal{O}(s))$ is necessary, for the case when the test function ϕ whose expectation is being estimated is the constant function $\phi(z) = 1$. This pair of necessary and sufficient conditions are given for normalized importance sampling Chatterjee and Diaconis (2018, Theorem 1.1—they also include very similar results for autonormalized importance sampling, in Theorem 1.2).

In our setting, the relationship in eq. 1.12 means that for an importance-sampling-based algorithm that draws samples of meanings from some proposal distribution $Z \sim q_{Z;\breve{w}}$ and uses importance sampling in order to re-weight these samples to form an approximate representation of the posterior, $p_{Z|\breve{w}}$, the cost in terms of the number of samples required is

$$\text{cost}(\breve{w}) = e^{D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}})} = e^{s(\breve{w}) - [R(\breve{w}) + D_{q_{Z;\breve{w}}}]} \tag{1.13}$$

where the distribution $q_{Z;\breve{w}}$ can potentially depend on $\breve{w}$—the general situation when using a proposal that may take into account the current observation when sampling meanings. This corresponds to divergence theory using KL from a proposal (hypothesis 1.5), with an exponential linking function.

If, in place of the proposal $q_{Z;\breve{w}}$, we assume that the samples are in fact simply drawn from

the prior, p_Z (not making use of the observation at all), then the proposal advantage term, $D_{q_{Z;\breve{w}}}$, vanishes, and we have divergence theory using KL from the prior (hypothesis 1.4). If we then additionally assume a binary likelihood, the reconstruction information term, $R(\breve{w})$, also vanishes, and we have general surprisal theory (hypothesis 1.2), still with an exponential linking function.

So, an algorithm whose complexity behaves like importance sampling's required number of samples gives a direct justification for an exponential version of KL divergence theory, whether this be the more general version, with a proposal, or the version which assumes samples are drawn from the prior. With the additional assumption that the relationship between meanings and observations is deterministic, this argument can likewise provide an algorithmic explanation for surprisal theory, though, importantly, with an exponential linking function, not a linear one.

Relating to probability divergences other than KL

The KL divergence is not the only way to quantify the amount one probability distribution differs from another. And indeed, the number of samples necessary for importance sampling has also been described in terms of other probability divergences, including the Hellinger distance, total-variation distance, or χ^2 divergence (see Y. Chen, 2005; Agapiou et al., 2017; Sanz-Alonso, 2018). It is worth noting that the choice of a different divergence may lead to a very different looking relationship between divergence and cost, but that this difference may be superficial.

For instance, importance sampling cost (in terms of the sample size necessary for approximating p with q) can be shown to scale *linearly* in $D_{\chi^2}(p \,\|\, q) \coloneqq \mathbb{E}_q\left[\left(\frac{\mathrm{d}p}{\mathrm{d}q} - 1\right)^2\right] = \mathbb{E}_p\left[\frac{\mathrm{d}p}{\mathrm{d}q}\right] - 1$ (see Sanz-Alonso, 2018, Thm. 4.2, 2). Compare this with the result, mentioned above, that this cost scales *exponentially* in $D_{\mathrm{KL}}(p \,\|\, q)$. For us, the upshot of this result is that if instead of using an operationalization of divergence theory using KL, we had instead chosen to use χ^2 divergence, the cost of importance sampling would suggest a linear relationship, cost $\approx D_{\chi^2}(p \,\|\, q)$, rather than an exponential one, cost $\approx e^{D_{\mathrm{KL}}(p \,\|\, q)}$. It is important to note that a hypothesis that processing cost increases linearly with χ^2 divergence instead of exponentially in KL does not necessarily constitute a radically different claim; such an apparent difference in linking function is a superficial result of the ways these two different divergences are defined (both of which, along with the others mentioned above, can be seen examples of the more general notion of *f-divergence*, $D_f(p \,\|\, q) \coloneqq \mathbb{E}_q[f(\frac{\mathrm{d}p}{\mathrm{d}q})]$; Rényi, 1961).[a]

In the current work I choose to focus on KL divergence as the quantification of belief-update size, due to its common use and information theoretic interpretability. However, I conjecture that it may be useful in the future to translate the approach to use quantify update cost differently if another function, such as χ^2 divergence, is most natural for understanding

a particularly promising inference algorithms' computational cost.

[a] In fact, one can show that in general $D_{\chi^2}(p \,\|\, q) + 1 \geq e^{D_{\mathrm{KL}}(p\|q)}$ (see, e.g., Gibbs & Su, 2002, Thm. 5)—essentially, the exponent of the KL divergence is the weighted geometric mean of the density $\frac{\mathrm{d}p}{\mathrm{d}q}$, while the χ^2 divergence is the weighted arithmetic mean (minus one). Loosely, these two divergences measure the same thing, but on a logarithmic versus linear scale. Yet, this loose interpretation should not be taken to mean that the two divergence are simply parametrically recoverable one from the other: No general lower bound on KL divergence can be expressed in terms of a function of χ^2 divergence (e.g., Polyanskiy & Wu, 2024, §7.6).

1.3.1.2 Simple guessing algorithms

A simpler class of sampling algorithms are those which simply make guesses according to some distribution, checking whether the guess can 'explain' the observation and continuing to make more guesses if not. If there is a deterministic relationship between latent representations and observable words, this reduces to simply checking whether the observed word corresponds to the sampled representation.

When the distribution from which guesses are drawn is the prior, this type of simple algorithm can also be shown to predict a runtime that is grows intrinsically with surprisal, since the expected number of guesses scales as an exponential function of the surprisal (which is equal to KL, given these assumptions). Chapter 2 gives proof of this relationship for this simple algorithm, and also a variant that samples without replacement, for which the number of samples can also be shown to grow superlinearly in surprisal, under some further assumptions about the prior distribution.

1.3.2 An intuitive argument for generalizing surprisal theory

The framing of divergence theory as a generalization of surprisal theory brings into focus one main intuition: In a setting where $R(\breve{w})$ is not necessarily zero (i.e., where the likelihood function is not necessarily binary), surprisal will not always be a good measurement of the information an observed word $\breve{w}$ carries about a latent variable Z. Instead, surprisal may overestimate this quantity.

Divergence theory, as proposed in the current work using KL divergence, provides a formalization of the notion that surprisal may sometimes, or even often, overestimate the intuitively relevant concept of surprise at an observation. As far as I am aware, this idea has not been explicitly investigated in the body of research on surprisal theory carried out in the past quarter century. However, it can be connected to intuitions that trace all the way back to some of the earliest uses of the concept of surprisal in cognitive science, as illustrated in the following passage. Just after stating the definition of surprisal as the log inverse probability of an event, Attneave (1959) gave the following disclaimer.

> The reader should be warned, however, that events with equal surprisal values may not be equally surprising to an observer. If a number between 1 and 10 is randomly chosen, whatever number is actually drawn—say, 8—has the same surprisal ($\log \frac{1}{.10} =$

> $\log 10 = 3.32$) as a throw of "tails" with the biased coin [that lands tails one-tenth of the time]. It is less surprising, however, because the number drawn is no more improbable than any other number would have been.
>
> (Attneave, 1959, ch. 1, p. 6)

Attneave was pointing out a situation in which surprisal can be higher than intuitively appropriate as a measure of rational surprise, when the observed event is not distinguished as a separate alternative of interest—or, I would say, more generally, when not all alternatives are equally important to distinguish.[11] Continuing from this intuition, we can imagine there are a set of individual outcomes where each is individually unpredictable but most do not result in any large change from a priori expectation, as in Attneave's example. This intuition motivates a more psychologically appropriate measure being one that quantifies change in expectation directly, as information gain quantified with relative entropy does.

Using divergence theory to capture Attneave's intuition

A slightly modified version of his example will capture the spirit of Attneave's point and allow us to see how the inadequacy of surprisal can be remedied by using KL divergence to quantify information gain.

Imagine that a number is randomly chosen $X \sim \text{uniform}\{1, \ldots, 10\}$, as in his example, say by rolling a fair ten-sided die. Additionally imagine there is one specific outcome that is of particular interest to you—for instance, say that just as the die was thrown you stated aloud a guess that it would land on 5. Then the relevant set of events you care about correspond to just whether or not your guess is correct, $\mathcal{Z} = \{\mathsf{right}, \mathsf{wrong}\}$, and the prior distribution on this set puts the probability of wrong at $9/10$. Now suppose the actual observed outcome is 8. Given this observation, the posterior belief distribution collapses to certainty on $z = \mathsf{wrong}$. So, while surprisal is $\log 10 \approx 3.32$ bits, the information gained about Z is only $D_{\mathrm{KL}}(\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right] \parallel \left[\begin{smallmatrix}1/10\\9/10\end{smallmatrix}\right]) = \log \frac{10}{9} \approx 0.15$ bits. The reconstruction information in this case, $R(\breve{x}) = \log 9 \approx 3.17$ bits (the expected amount of information required to specify the exact observation, under the posterior), is nearly as large as surprisal—nearly all of the surprisal was comprised of 'wasted' bits, from the point of view of information gain.

For comparison, suppose instead that the actual outcome had in fact been 5. Then $R(\breve{x}) = 0$ and the KL from posterior to prior is equal to surprisal at $\log 10$ bits, just as it would be for the throw of tails with the biased coin.

Note, staying true to Attneave's exact example, where all observations are equally unin-

[11] This kind of abstraction whereby "a smaller world is derived from a larger by neglecting some distinctions between states" (Savage, 1972, §2.3, p. 9) is a ubiquitous component of the modelling of rational behaviour.

teresting, may be seen as a degenerate case of the above, where there is only one element in the set $\mathcal{Z}$, rather than two as in my modified example. In such a degenerate situation, there is no information gain, no matter how surprising the individual outcome of X is, since there is only one unique degenerate belief distribution over a singleton set, so there can be no shift in beliefs.[a]

[a] This is similar to the intuition in the canonical 'white snow' example of an unpredictable but uninformative observation, mentioned in e.g. Baldi (2002): For a television viewer interested in discerning what channel is being displayed, a screen that continually shows random pixel noise comprises maximally unpredictable input, but it is, in a sense, the least informative possible television program. Continued attention to this input induces no change in beliefs whatsoever for the viewer—once they have understood that they are looking at something which outputs white noise, the details of that noise are completely irrelevant and uninteresting, despite continuing to be extremely unpredictable.

1.3.3 Testing the assumptions of surprisal versus divergence theory

From the perspective of divergence theory, we can consider each of standard surprisal theory's implied assumptions in turn (starting in the last row of table 1.1 with standard surprisal theory, and moving upward as we relax assumptions). In the following chapters we will consider these questions one at a time keeping in mind that they can interact. First is the question of whether the linking function is linear; this is a question that has seen significant attention in previous literature, generally supporting a linear linking function, however there are also reasons to question this assumption, which will be reviewed below and discussed at length in the next chapter. Moving on from the question of the linking function, we can consider the assumption of a binary likelihood—the assumption that is necessary for surprisal to be equivalent to KL. Relaxing this assumption gives the prediction that surprisal may overestimate the cost of processing cost—in situations where the proportion of bits that are not relevant to belief update, quantified by $R(\breve{w})$ (or $R(\breve{w}) + D_{q_{Z;\breve{w}}}$), is non-negligible.

1.3.3.1 Nonlinear linking function: motivating chapter 2

While much of the literature on surprisal theory has either explicitly or implicitly assumed a linear linking function between surprisal and processing cost, there are a number of independent theoretical as well as empirical reasons to question this assumption. First, from the theoretical perspective, the potential algorithmic explanations of divergence theory in terms of sampling algorithms imply an exponential linking function, as described above. This argument supplements independent theoretical arguments supporting a superlinear linking function in earlier work in general surprisal theory (which will be discussed in detail in the next chapter, see §2.3.3). If the true linking function between divergence and cost is superlinear, then, assuming a binary likelihood, as all previous work has done, we should also expect the linking function with surprisal to be superlinear.

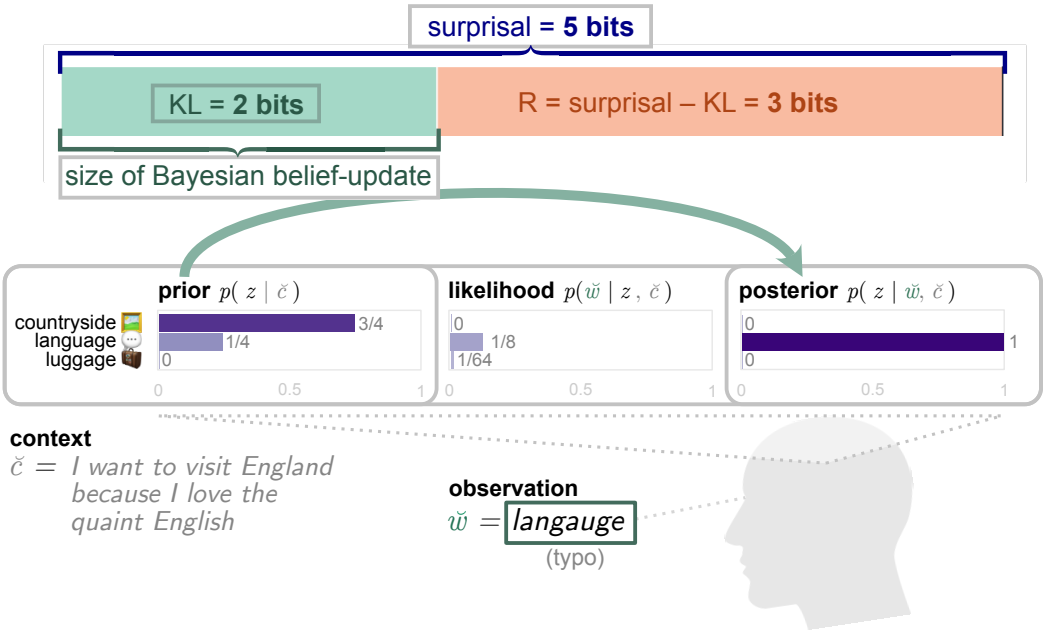


Figure 1.2: Diagram illustrating a toy example of surprisal and information gain of the observation $\check{w} =$ *langauge* in context, with prior and likelihood chosen to instantiate the example in fig. 1.1b, where surprisal is 5 bits, but KL is only 2 bits. The remaining 3 bits constitute the information about the precise form of the observation (in this example, the fact that it contains a spelling error), which are not relevant to the inference about intended meaning.

Additionally, many empirical studies of human reading time as a function of surprisal use a log transform on reading time (Boston et al., 2008; Roark et al., 2009; Aurnhammer & Frank, 2019; Merkx & Frank, 2021; J. Mitchell et al., 2010; Oh & Schuler, 2023a, 2023b; Oh et al., 2022, 2024). This transformation implies an exponential relationship between surprisal and processing time (as acknowledged in, for example Oh et al., 2024). We contribute to the ongoing debate about the linking function within general surprisal theory from both theoretical and empirical angles in chapter 2, providing evidence in favor of a superlinear relationship.

1.3.3.2 Explaining unexpected-but-easy phenomena: motivating chapter 3

Allowing for a non-deterministic relationship between meanings and utterances, with the immediate theoretical consequence of breaking the equivalence between surprisal and KL, is a novel direction of inquiry in this area. This opens the door to potential explanations for phenomena that surprisal intrinsically cannot capture, in particular those where surprisal as estimated by an accurate language model has an unexpectedly small effect on human cognitive effort.

Orthographic errors One such example is typographical errors in written text, which are not easily predicted, but this unpredictability is not due to their containing meaningful information. If processing cost is a reflection of information gain, the effort required to incorporate a word

with an unexpected form because of a minor typo should be much smaller than that required to incorporate a word that is unexpected because of the meaning it expresses, even if surprisal is the same in the two situations.

To illustrate this intuition, fig. 1.2 repeats the toy example given in fig. 1.1b of a prior and likelihood and the resulting surprisal and KL. Here, the example is assigned concrete values for the three meaning candidates, with the observation being a word containing a typographical error, as an example of a situation where surprisal may plausibly be substantially larger than KL. Given the context, $\breve{c}$, depicted in the figure, suppose the comprehender's prior over possible meanings of the utterance are as illustrated (representing the expectation that the meaning is likely be about the English countryside, or, less likely but still plausibly, the English language). Then, upon observing the typo $\breve{w} = \textit{langauge}$, the uncertainty about the intended meaning is entirely resolved, resulting in the posterior diverging from the prior by 2 bits. However, the expected likelihood for this observation is low—illustrated with the value $1/8$ (by contrast, for a correctly spelled version of the word, likelihood would be much closer to 1), and for this reason surprisal exceeds KL by three bits. This toy example illustrates the way examples of malformed input such as minor typographical errors may provide plausible test cases for comparing the predictions of KL versus surprisal. If such input can be processed roughly as if it did not contain the malformation which makes it unpredictable, then the belief update it engenders should be smaller than surprisal would predict. Investigating human reading times on items inspired by this intuition will be the focus of the empirical study presented in chapter 3.

Syntactic or semantic illusion effects Instances in which surprisal may be expected to substantially overestimate processing difficulty are not limited to orthographically anomalous tokens. Other types of situations which may be expected to behave in a similar way with respect to surprisal and KL include any observation where the particular form is extremely unlikely, under the prior, and remains relatively unlikely even under the posterior. This may potentially occur for phonological, syntactic, semantic or pragmatic reasons, as well as orthographic, any time when, according to the grammar of the comprehender, the observation is highly unlikely to be produced as a way of expressing the meaning.

This means that documented phenomena where processing of a word is easy despite being ungrammatical or unlikely to be produced, provide potential examples that may benefit from such a theoretical explanation. The broad family of phenomena known as *grammatical illusions* (Phillips et al., 2011; Muller & Phillips, 2020; Leivada, 2020; Paape et al., 2020) provide promising sources of such examples. These are constructions which are syntactically or semantically malformed, or pragmatically infelicitous when interpreted literally, yet are perceived as acceptable to humans when encountered and provide less difficulty in comprehension than would be otherwise predicted.

Examples of illusion effects include so-called *depth-charge illusions*, the canonical example of

which is the sentence *No head injury is too trivial to ignore* (Wason & Reich, 1979; Paape et al., 2020). The literal meaning of this sentence is the pragmatically extremely odd (that all head injuries should be ignored no matter how trivial), it is nearly universally interpreted as having a much more pragmatically plausible meaning (that no head injuries should be ignored). Another family of examples include *NPI illusions*, situations in which a negative polarity item (NPI) such as *any* or *ever*, are unexpectedly deemed acceptable, despite not appearing in the scope of negation (or other downward-entailing contexts)—normally a requirement for such words to be grammatical. A canonical example of an NPI illusion is the sentence *The bills that no senators voted for will ever become law* (Xiang et al., 2009; Muller & Phillips, 2020), where the NPI *ever* is not in the scope of negation, and yet this sentence provides less impediment to processing than other examples with NPIs in unlicensed positions. For the purpose of the current discussion, the important point about such illusions is that they contain words that are highly unpredictable and yet do not pose a commensurate amount of difficulty in processing. The structural and processing-related explanations of the attested unexpected acceptability of these various types of constructions has been the subject of research for decades, and this literature provides a rich source of possible example constructions to consider as additional test cases for KL versus surprisal theory.

Explaining task sensitivity A corollary of moving from a theory of surprisal to one of belief-divergence as the driver of processing cost is that the prediction becomes not a function of a single value in the probability distribution over possible observations (as surprisal is), but instead something that depends intrinsically on what the space of relevant meanings is, and their relationship to the observation via a likelihood function. The true surprisal of particular string of characters in context is a fixed quantity (which we may estimate with a language model), that does not depend on what about that string is relevant to the comprehender. However the same observation in the same linguistic context may be more or less informative depending on the cognitive task. For instance, a simple typo may cause almost no slowdown when reading quickly for general comprehension, when it is not relevant to the understanding of the message, but may cause a larger response when proofreading, when it is highly relevant to the reader (consistent with work finding task-sensitivity in the effect of surprisal on effort, e.g., Schotter et al., 2014). The empirical probability of that string of characters (and therefore the surprisal, as can be estimated by a language model) is not different between the two situations. But the KL will be different, if in the two situations the relevant space of hypotheses that are being distinguished differ.

1.4 Other measures of information gain

Relative entropy is not the only way to implement the intuition that processing cost reflects the change it induces in the state of a generative probabilistic processor. The most notable alternative proposal is the *entropy reduction* hypothesis (Hale, 2003a, 2003b, 2006; inspired by original

formulations in Wilson & Carroll, 1954; Lounsbury, 1954).

Hypothesis 1.6 (general entropy reduction)**.** The processing cost of a word $\breve{w}$ is a monotonic increasing function of the amount of information it communicates, as quantified by any reduction in entropy over meanings Z induced upon observing $\breve{w}$.

$$\text{cost}(\breve{w}) = f(\max\{0, \text{ER}(\breve{w})\}) \tag{1.14}$$

where f is a monotonically increasing function,[a] and $\text{ER}(\breve{w})$ is entropy reduction:

$$\text{ER}(\breve{w}) \coloneqq \text{H}(Z) - \text{H}(Z \mid \breve{w}) \tag{1.15}$$

$$= \mathop{\mathbb{E}}_{p_Z}\left[\log \frac{1}{p(z)}\right] - \mathop{\mathbb{E}}_{p_{Z\mid\breve{w}}}\left[\log \frac{1}{p(z \mid \breve{w})}\right] \tag{1.16}$$

[a] This is a slight generalization of the version given in Hale (2003a, ch. 3), which assumes f is linear:

> A person's reading time at a word in a sentence is linearly related to any downward change in the entropy of the set of derivations generating the observed words as a prefix.
>
> (Hale, 2003a, §3.2.3, *Hypothesis 1: Entropy Reduction Hypothesis — precise*)

Entropy reduction as a quantification of the information about a latent random variable Z given by an observation $\breve{w}$, has often been suggested in mathematical psychology and statistics (Lindley, 1956; MacKay, 1992; Chater et al., 1998), often in fact under the name *information gain*.

Both entropy reduction and the KL divergence between posterior and prior provide ways of quantifying the information that the particular observed outcome $\breve{w}$ of random variable W gives about the random variable Z. And (as discussed in e.g. Blachman, 1968) both have the property that taking their expectation over all possible observations $w \sim p_W$ gives the mutual information:

$$\mathop{\mathbb{E}}_{p_W}\left[D_{\text{KL}}(p_{Z|w} \,\|\, p_Z)\right] = \text{I}(Z : W) \tag{1.17}$$

$$\mathop{\mathbb{E}}_{p_W}\left[\text{ER}(w)\right] = \text{I}(Z : W) \tag{1.18}$$

however neither can be reduced to the other (so the ER hypothesis cannot be slotted into the hierarchy of hypotheses in table 1.1).[12]

[12] In Blachman's notation, entropy reduction is $I(Z; \breve{w})$, and the KL between posterior and prior is $J(Z; \breve{w})$, these being the two measures that he considers of information about Z contained in $\breve{w}$. The exploration of these measures of information gain can be traced back further to Lindley (1956) and Cronbach (1953), mostly used under expectation so that they were equivalent to mutual information $\text{I}(Z : W)$. Note however that while mutual information has also been explored as a measure of information gain in this context occasionally (see, e.g., Hale, 2003a) it has the undesirable property that it is determined by the random variable W globally, and does not depend on the actual identity of the particular observation $\breve{w}$, and thus may be appropriate only for measuring an anticipatory rather than responsive cost (in the sense of Pimentel et al., 2023).

Entropy reduction has theoretical and empirical weaknesses when compared to divergence theory and surprisal. For one, while it does provide plausible seeming alternative quantification of the concept of information gain, it does not benefit from the main justification that motivates surprisal theory and divergence theory in terms of a resource allocation cost, since the change in entropy is fundamentally not a quantification of how much the distributions differ. In particular, entropy reduction may be negative (when an observation leads to increased entropy), leading to unclear predictions about processing cost (and this simple fact may be seen as a general disadvantage of entropy reduction as a measure of information gain, when compared to KL divergence, as noted in e.g., Oaksford & Chater, 1996; Chater et al., 1998). Hale stipulates that the processing cost is zero in any case where entropy increases, but this has the undesirable consequence of predicting identical (zero) cost for an observation which causes no change in the distribution over Z as for an observation which increases entropy. Additionally, from an empirical perspective, while this has not seen attention in recent literature using large language models, there have been some studies comparing the entropy reduction hypothesis with surprisal theory: For instance, S. Wu et al. (2010) compared entropy reduction and surprisal (among other metrics) as predictors of self-paced reading times, found stronger evidence for an effect of surprisal. Similarly, Linzen and Jaeger (2014) compared surprisal to entropy (over parses in a PCFG), and found evidence for the effect of both on reading time.

Conclusion and roadmap for following chapters

In this chapter, I have developed a reframing of the central hypothesis of surprisal theory, in what I have termed divergence theory, proposing a quantification of processing cost that is mathematically equivalent to surprisal theory only with certain simplifying assumptions, which previous literature has implicitly or explicitly assumed. This modified framework builds on the established strengths of traditional surprisal theory while affording potential advantages, both theoretical and empirical. Namely, it allows a more direct and intrinsic link to a wide family of potential algorithmic theories, such as belief-update algorithms for approximate inference, and has the flexibility to explain phenomena in human processing cost that standard surprisal alone cannot.

Within this framework, the following two chapters will present investigations into relaxing two of the main assumptions which are implied by standard surprisal theory, one at a time. Chapter 2 will address the question of finding a class of algorithms which could explain general surprisal theory, looking broadly at the complexity of algorithms which sample from the prior in order to approximate the posterior. This work will follow all previous literature in assuming KL divergence and surprisal are equivalent, in order to focus just on the question of the form of the linking function. Then chapter 3 will explicitly question the binary likelihood, looking at typographical errors as an example of a situation where surprisal may reasonably be expected to substantially

overpredict processing cost that is due to information gain as quantified by KL divergence.

2

The plausibility of sampling as an algorithmic theory of sentence processing

Published as Hoover et al. (2023)

Words that are more surprising given context take longer to process. However, no incremental parsing algorithm has been shown to directly predict this phenomenon. In this work, we focus on a class of algorithms whose runtime does naturally scale in surprisal—those that involve repeatedly sampling from the prior. Our first contribution is to show that simple examples of such algorithms predict runtime to increase superlinearly with surprisal, and also predict variance in runtime to increase. These two predictions stand in contrast with literature on surprisal theory (Hale, 2001; R. Levy, 2008a), which assumes that the expected processing cost increases linearly with surprisal, and makes no prediction about variance. In the second part of this paper, we conduct an empirical study of the relationship between surprisal and reading time, using a collection of modern language models to estimate surprisal. We find that with better language models, reading time increases superlinearly in surprisal, and also that variance increases. These results are consistent with the predictions of sampling-based algorithms.

2.1 Introduction

One of the fundamental problems of computational psycholinguistics, going back to the earliest days of the field, is to provide an algorithmic theory of human sentence processing (see e.g., Yngve, 1960; Rasmussen & Schuler, 2018; Miller & Chomsky, 1963; Marcus, 1978; Frazier & Fodor,

1978; Roark, 2001; Stolcke, 1995; Collins & Roark, 2004; R. L. Lewis & Vasishth, 2005; Vasishth & Engelmann, 2021; Dotlačil, 2021). Such an algorithmic theory must satisfy a number of important empirical constraints. Amongst these are that the human processor is *incremental* and *predictive*—people process sentences eagerly, assigning as much meaning as possible as early as possible, and predicting likely continuations based on the current context (Marslen-Wilson, 1973, 1975; Frazier, 1987; Eberhard et al., 1995; Tanenhaus et al., 1995). Moreover, the effort needed to integrate each subsequent word (or smaller unit) depends on how predictable it is, in context, often quantified as *surprisal* (negative log probability given context; Hale, 2001; R. Levy, 2008a). The more surprising a word is, the more time it takes to integrate (e.g., Ehrlich & Rayner, 1981; Balota et al., 1985; McDonald & Shillcock, 2003b, 2003a; Wilcox et al., 2020; Brothers & Kuperberg, 2021; Meister et al., 2021).

However, despite the widespread recognition of these empirical facts, and the large number of studies looking at surprisal as an empirical predictor of incremental processing time (e.g., Demberg & Keller, 2008; Smith & Levy, 2008a, 2013; Goodkind & Bicknell, 2018, 2021; Wilcox et al., 2020; Meister et al., 2021; Hofmann et al., 2022), to our knowledge no sentence processing algorithm has been proposed for which incremental runtime intrinsically increases as a function of surprisal.

In §2.2, we review the kinds of algorithms that could possibly possess the desired properties, identifying and focusing on a class of approaches for which the desired relationship with surprisal is very natural—sampling based algorithms. The first contribution of this paper is to show that under some reasonable assumptions, sampling-based algorithms predict processing time to be a monotonic increasing function of surprisal. In particular, these algorithms predict runtime to increase as a superlinear function of surprisal. We also show that these algorithms make a novel prediction about processing times—under sampling based algorithms, we also expect variance to increase with surprisal.

However, as we discuss in §2.3, these two predictions are inconsistent with the assumptions made by the majority of published work in surprisal theory. In particular, empirical studies in this area have often assumed that the relationship between surprisal and processing time is linear (Demberg & Keller, 2008; Fernandez Monsalve et al., 2012; Frank et al., 2013), or at least that variance is constant (Smith & Levy, 2008a, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020; Meister et al., 2021). We review the status of the widespread assumptions of linearity and constant variance, identifying both theoretical and empirical reasons to question these properties.

We then present a new targeted study of the empirical relationship between surprisal and reading time (in §2.4). We obtain surprisal estimates from a variety of pre-trained language models (LMs), including GPT-3 (Brown et al., 2020) and then use generalized additive models (Wood et al., 2016) to examine the shape of the linking function between surprisal and reading time. We control for possibly nonlinear by-subject random effects, and also fit the relationship between surprisal

and variance in reading time. We find evidence that the overall shape of the linking function is in fact superlinear, especially for surprisals estimated by the most accurate LMs. We additionally find that variance in reading time increases with surprisal. Both these results are at odds with the assumptions typically made in surprisal theory, but they are consistent with the predictions of sampling-based algorithms for processing.

We situate our results in the context of earlier literature, speculating that our ability to detect this superlinear relationship rests on several ways our empirical study improves upon previous work. Namely, we use higher quality LMs to estimate surprisal, and fit statistical models designed to assess the possibly nonlinear relationship, controlling for individual differences. In the discussion, we also revisit previous proposals which are related to the analyses we give of sampling algorithms. Based on our theoretical and empirical results, we propose that sampling-based mechanisms form a promising yet under-explored family of algorithms for the modelling of human sentence processing.

2.2 Sampling algorithms for sentence processing

It is well documented that for humans, 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 (e.g., Ehrlich & Rayner, 1981; Balota et al., 1985; McDonald & Shillcock, 2003a, 2003b; Smith & Levy, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020; Brothers & Kuperberg, 2021; Meister et al., 2021; Hofmann et al., 2022). We may write this general relationship as:

$$\text{Time}(w_n) \approx f(s(w_n)) \tag{2.1}$$

where the linking function f is some monotonically increasing function, and

$$s(w_n) \coloneqq -\log p(w_n \mid w_{1:n-1}) \tag{2.2}$$

is the *surprisal* of word w_n. Thus, we seek an algorithmic model of sentence processing where the computational cost to perform each incremental update depends on the surprisal of the input at that step.

To clarify what is at stake, it is useful to consider the incremental sentence processing problem in more detail. Sentence processing can be viewed as a sequence of posterior inference problems: The comprehender updates their beliefs about the intended meaning, parse, or other latent structure as they successively observe linguistic input items (e.g., words, morphemes, or smaller units). Formally, we can define a probabilistic incremental parser as a map which, at each step, takes the sequence of linguistic inputs seen so far to a posterior distribution: $w_{1:n} \mapsto p(z \mid w_{1:n})$, where z ranges over meanings (or parses, etc.). Consider one step of this process, assuming that the comprehender has a representation of the exact posterior distribution given $w_{1:n-1}$, then encounters the next

word w_n. The job of this comprehender is to update their beliefs about meanings in light of the evidence, to obtain a new posterior:

$$p(z \mid w_{1:n}) = \frac{p(w_n \mid z)p(z \mid w_{1:n-1})}{\sum_z p(w_n \mid z)p(z \mid w_{1:n-1})} \tag{2.3}$$

Note that the denominator here is $\sum_z p(w_n \mid z)p(z \mid w_{1:n-1}) = p(w_n \mid w_{1:n-1})$, the marginal probability of the word given the preceding context—the negative logarithm of this quantity is the surprisal. This denominator represents the proportion of the prior meaning space that remains after posterior update. When it is small (and thus surprisal is high), this means that very little of the prior meaning space $p(z \mid w_{1:n-1})$ was consistent with the new word, when it is large (and thus surprisal is low), this means that much of the prior meaning space was consistent with the new word.

2.2.1 Algorithms that do not scale in surprisal

In the literature studying surprisal and processing cost, it has been common to use enumerative algorithms, such as Stolcke's probabilistic variant of Earley's chart-based algorithm (Earley, 1970; Stolcke, 1995) to estimate surprisal values (e.g., Boston et al., 2008; R. Levy, 2008a). Without further assumptions such as probability-based pruning (see below), such enumerative algorithms do not use the probability of chart items in deciding how much work to do, and thus do not scale in surprisal. The number of steps such an algorithm takes to integrate the next word into the chart can depend on the size and specification of a probabilistic grammar, but cannot depend on the probability of the word. This is also true of the many probabilistic or non-probabilistic bottom-up, top-down, or left corner parsing algorithms which have been studied over the years as models of sentence processing (Earley, 1970; Rosenkrantz & Lewis, 1970; Marcus, 1978; Abney & Johnson, 1991; Berwick & Weinberg, 1982; Roark, 2001; Nivre, 2008; Stabler, 2013; Graf et al., 2017), and likewise for RNN- or Transformer-based parsing models (e.g., Costa, 2003; Jin & Schuler, 2020; K. Yang & Deng, 2020; X. Hu et al., 2021, 2022).

Other parsing algorithms have properties which result in some correlation between surprisal and processing cost, without predicting the relationship directly. For instance, amortized parsing techniques that make use of *chunked* (Newell & Paul, 1981) parser moves or grammar fragments (as examined in, e.g., Hale, 2014; Luong et al., 2015), can predict broadly that common sequences of actions lead to lower surprisal. However, these accounts do not predict any direct link between individual word probability and the amount of computational work done by the processor. A similar argument can be made for theories which describe processing difficulty primarily in terms of distance-based measures such as dependency locality theory (DLT; Gibson, 1998, 2000), where certain common words may tend to have shorter dependencies, but the surprisal of a word is not

intrinsically related to its integration cost.

A final class of models to consider includes causal language models, which do not produce any observable representations of the meaning of their input, but rather simply predict the next word given some prefix (Hochreiter & Schmidhuber, 1997; Radford et al., 2018, 2019; Dai et al., 2019; Brown et al., 2020). The amount of work required by these algorithms may scale in quantities such as the length of the input or the size of the vocabulary, or other functions of the architecture of the model, but never directly as a function of the probability of the next word.

2.2.2 Algorithms that do scale in surprisal

As outlined above, highly probable words will necessarily tend to be associated with more likely meanings (parses) given the preceding words, while the least likely words will tend to be less compatible with these meanings. This suggests a natural way to relate processing algorithms' computational cost to the surprisal of the next word: When doing the posterior update, give priority to those meanings which are highly likely in the prior $p(z \mid w_{1:n-1})$. Since a word w_n with low surprisal will tend to be associated with highly probable prior meanings, privileging meanings in such a way will lead to algorithms with the desired dependence on surprisal.

In this work we focus on a broad class of algorithms that privilege high prior probability meanings: those that *sample* candidate meanings from the prior distribution $p(z \mid w_{1:n-1})$.[1] Another closely related class of algorithms with this property are those which perform a deterministic search over the space of meanings, in order of decreasing prior probability. Such an algorithm will naturally tend to take longer when confronted with an input word that has higher surprisal (see discussion in §2.6.3).

In what follows, we will consider two simple procedures for sampling from the prior and discuss their consequences for theories of incremental sentence processing.

2.2.3 Two simple sampling algorithms

In the analyses that follow, we consider the problem of integrating a single word w_n assuming that the comprehender has an exact representation of the true prior: $p(z \mid w_{1:n-1})$. Note that the probability that a random sample from the true prior will be consistent with observed word w_n is given by $\sum_z p(w_n \mid z)p(z \mid w_{1:n-1}) = p(w_n \mid w_{1:n-1})$. Thus, without loss of generality, we simplify the problem to analyzing the expected number of samples needed to exactly match w_n. Note, assuming an exact prior representation is highly conservative, since, in general, sampling-based algorithms for incremental processing will have to be approximate (e.g. using Markov chain or sequential Monte Carlo techniques) and so will accumulate errors. A similar observation can be made about modified versions of these algorithms which sample until some constant number of

[1] The particle filter model proposed in R. Levy et al. (2008) is a specific example of such an algorithm applied to parsing, but due to modelling choices, its runtime doesn't scale in surprisal. We will discuss this model in §2.6.2.

successes are achieved (rather than stopping at the first success). The runtime analyses we do here will thus provide a lower bound on runtime for the more general class of algorithms.

2.2.3.1 Simple guessing algorithm

Define the simple guessing algorithm[2] as follows: To get an exact sample from posterior $p(\cdot \mid w_{1:n})$, given prior $p(\cdot \mid w_{1:n-1})$, and observed next word w_n, repeatedly sample hypotheses (meanings) from the prior until getting one which explains the observed next word.[3]

The number of samples needed in this scheme, M, is geometrically distributed $M \sim \text{Geom}(p)$, where parameter $p = p(w_n \mid w_{1:n-1})$ is the probability of success. This random variable has expected value $1/p$ and variance $(1-p)/p^2$. Expressed as a function of surprisal, the expected value and variance are

$$\mathbb{E}[M] = \frac{1}{p} = e^{s(w_n)} \tag{2.4}$$

$$\text{Var}[M] = \frac{1-p}{p^2} = e^{2s(w_n)} - e^{s(w_n)} \tag{2.5}$$

So, the expected runtime of this sampling scheme (eq. 2.4) increases monotonically—in fact, exponentially—in surprisal. Likewise, the variance in runtime (eq. 2.5) also increases monotonically and superlinearly as a function of surprisal (to see this, note that all its derivatives are everywhere positive).

2.2.3.2 Guessing without replacement algorithm

In the simple guessing algorithm above, a meaning may be repeatedly sampled from the prior, despite not explaining the observation. So, we will also consider a more efficient version of the above scheme where sampling is carried out *without replacement* to avoid re-sampling meanings that have already been eliminated.

Define the simple guessing algorithm without replacement as follows: Let the meanings which do not explain the observation be indexed $1, \ldots, K$, with weights $\{u_i\}_{i=1}^K$. Consider one additional item, the target, assigned index 0, with weight, u_0, proportional to the total probability mass of the meanings which do explain the observation. At each step of the algorithm an item is sampled from the set $\{0, \ldots, K\}$ with probabilities proportional to the weights of the items not yet drawn.

[2]This simple sequential sampling algorithm, also mentioned in Freer et al. (2010), is sometimes informally referred to as 'rejection sampling.' We use the term 'guessing' to avoid confusion with the more general rejection sampling algorithm (as defined in, e.g., Chopin & Papaspiliopoulos, 2020, alg. 8.1), of which it is a special case.

[3]This is intentionally the simplest possible version of such an algorithm. Among the many possible refinements (which might be sensible in practice) would be to continue guessing until some reasonable number of successes, rather than stopping at the first success. Note that such a modification does not change the asymptotic complexity, simply adding a constant multiplier. As noted above, we do not analyze such particular modifications since we are not proposing a specific algorithm. Our goal with these analyses is to understand the general asymptotic complexity characteristics of the class of algorithms which involve iterative guessing from the prior.

The algorithm halts when the target item (0) is drawn.

Define binary random variables $\{X_i\}_1^K$ where $X_i = 1$ if item k is drawn before the target, else $X_i = 0$. Let random variable N be the number of guesses without replacement up to and including when the target is drawn. Then the runtime $N = 1 + \sum_{i=1}^K X_i$.

To derive runtime mean and variance for this algorithm, the following proposition will be useful.

Proposition 2.1. *In a guessing algorithm (with or without replacement) the probability of drawing item i before item j is* $\Pr(i \prec j) = \frac{u_i}{u_i+u_j}$.[a]

Proof. Consider a modification of the guessing-without-replacement scheme in which items i and j have been removed from the set and a new item $i \vee j$ is inserted instead, with weight $u_i + u_j$. If this item is drawn, then we say i is drawn with probability $\Pr(i \mid i \vee j) = \frac{u_i}{u_i+u_j}$, else j is drawn. The runtime of this scheme is identical to that of guessing without replacement. Let S_{K-1} be the set of permutations of $\{0, \ldots, K\} \setminus \{i, j\} \cup \{\{i \vee j\}\}$. First note that for any permutation $\sigma \in S_{K-1}$, the conditional probability $\Pr(i \prec j \mid \sigma) = \Pr(i \mid i \vee j)$. So $\Pr(i \prec j) = \sum_\sigma \Pr(i \prec j \mid \sigma) \Pr(\sigma) = \Pr(i \mid i \vee j) = \frac{u_i}{u_i+u_j}$. □

[a] Note the probability $\Pr(i \prec j)$ depends on the weights of items i and j, and no others. This means it is independent of the order the other items are drawn in, what their probabilities are, and even whether drawing is done with or without replacement.

So, with $\mathbb{E}[X_i] = \Pr(i \prec 0) = \frac{u_i}{u_i+u_0}$, we have that the expected runtime (number of draws), is

$$\begin{aligned} \mathbb{E}[N] &= \mathbb{E}[1 + \sum_i X_i] = 1 + \sum_i \mathbb{E}[X_i] \\ &= 1 + \sum_i \frac{u_i}{u_i + u_0} \end{aligned} \tag{2.6}$$

and the variance in number of draws is

$$\begin{aligned} \mathrm{Var}[N] &= \sum_i \left[\mathbb{E}[X_i] - (\mathbb{E}[X_i])^2\right] + \sum_{i \neq j} \Big[\mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]\Big] \\ &= \sum_i \left[\frac{u_i}{u_{i0}} - \left(\frac{u_i}{u_{i0}}\right)^2\right] + \sum_{i \neq j} \left[\frac{u_i}{u_{ij0}}\frac{u_j}{u_{j0}} + \frac{u_j}{u_{ij0}}\frac{u_i}{u_{i0}} - \frac{u_i}{u_{i0}}\frac{u_j}{u_{j0}}\right] \end{aligned} \tag{2.7}$$

using notation $u_{ab} \coloneqq u_a + u_b$ and $u_{abc} \coloneqq u_a + u_b + u_c$. See appendix A.1 for a derivation.

An important property to note here is that the individual weights of all items $\{u_i\}_{i=0}^K$ appear in the general expressions for mean runtime (eq. 2.6) and variance in runtime (eq. 2.7). This means that both mean and variance in runtime depend on how the weights are distributed across all the

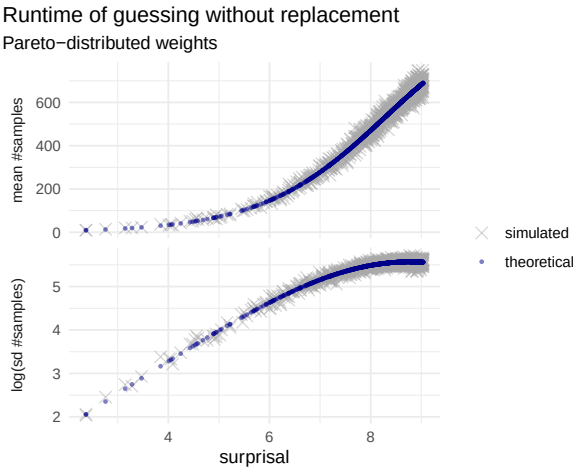


Figure 2.1: Relationship between surprisal (negative log probability) and guessing-without-replacement runtime for a set of 1000 weights sampled from a Pareto$(1, 1)$ distribution. Blue points show theoretical values for mean (top) and variance (bottom, transformed as log standard deviation). Grey crosses give average values in simulating 500 runs of the algorithm for each surprisal value.

items—not just the probability of success, as was the case in the simple guessing (with replacement) algorithm. Obtaining a concrete prediction for how the runtime scales as a function of surprisal requires making some assumption about the distribution from which we are sampling.

We will assume the item probabilities are heavy-tailed—specifically, that they are power-law distributed (a property ubiquitous in language, and word frequency distributions in particular; see Piantadosi, 2014). Figure 2.1 shows the empirical mean and variance of guessing-without-replacement runtime (number of samples until success) plotted against the surprisal of the target, for $K = 1000$ weights sampled from the power-law distribution Pareto$(1, 1)$, and normalized. Each of the discrete values on the horizontal axis corresponds to the negative log probability of one item in the set. The mean runtime to sample that item as the target is plotted in the top panel, and variance in the bottom panel. Blue points mark the theoretical values according to mean and variance derived in eqs. 2.6 and 2.7, and grey crosses indicate simulated values (estimated by simulating 500 runs of the algorithm for each item as the target).

We observe that the runtime of guessing-without-replacement increases as a superlinear function of surprisal, as is the case for the simple guessing algorithm with replacement. We also see that variance increases over most of the range of surprisal values, plateauing at the highest values of surprisal. Broadly, with respect to variance, we can say simply that it increases with surprisal, for both the with- and without-replacement algorithms.

2.3 Surprisal theory

The relationship between surprisal and human processing time has received attention in a large number of studies (Bicknell & Levy, 2010, 2012; Boston et al., 2008; Brothers & Kuperberg, 2021; Demberg & Keller, 2008; Fernandez Monsalve et al., 2012; Frank, 2009; Frank et al., 2013; Futrell, 2017; Futrell et al., 2020; Goodkind & Bicknell, 2018, 2021; Hale, 2001; Hofmann et al., 2017, 2022; Jin & Schuler, 2020; Jurafsky, 1996; R. Levy, 2005, 2008a, 2008b, 2011, 2013, 2018; Lowder et al., 2018; McDonald & Shillcock, 2003b, 2003a; J. Mitchell et al., 2010; Narayanan & Jurafsky, 2001, 2004; Rasmussen & Schuler, 2018; Reichle et al., 2003; Roark et al., 2009; van Schijndel & Linzen, 2021; Smith & Levy, 2008a, 2008b, 2013; Wilcox et al., 2020). We will refer to literature focusing on this relationship as work on surprisal theory, broadly. The question of the shape of the function linking surprisal and processing time goes back to early work in the area (Hale, 2001; Narayanan & Jurafsky, 2004; R. Levy, 2005). The majority of work, however, has either assumed or explicitly argued for a linear linking function, that is,

$$\text{Time}(w_n) = \alpha + \beta\, s(w_n) \tag{2.8}$$

for some constants α and β. This stands in contrast with the superlinear linking function predicted by sampling-based mechanisms, described above. A linear relationship has been motivated both empirically and on the basis of theoretical arguments. Nevertheless, as we review below, there are reasons to question the assumption of linearity, including relatively recent studies that provide evidence of a superlinear linking function as well as earlier theoretical models that have assumed or argued for superlinearity (see §2.3.3). Furthermore, as we note below, nearly all previous work has assumed the relationship between surprisal and variance in processing time to be constant.

2.3.1 Empirical studies in surprisal theory

Determining the correct functional relationship between surprisal and processing time is a long-standing problem in the field. A large number of studies have simply assumed a linear relationship, explicitly—or implicitly, by the use of linear statistical models for their analysis (e.g., D. C. 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).[4] A smaller number of papers, beginning with Smith and Levy (2008a, 2013), have investigated the shape of the linking function directly, using generalized additive models (GAMs; Wood, 2004, 2017), a family of statistical models which allows the fitting of arbitrary nonlinear

[4]Others have used linear models with a log-link, or log-transformed dependent variable (e.g., Boston et al., 2008; Roark et al., 2009; J. Mitchell et al., 2010; Aurnhammer & Frank, 2019; Merkx & Frank, 2021; Oh et al., 2022; Oh & Schuler, 2023a, 2023b), implying an exponential relationship between surprisal and reading time (see §2.3.3).

relationships (Smith & Levy, 2008a, 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, there are a number of methodological reasons to revisit these results.

First, none of these previous studies has attempted a quantitative measure of superlinearity, relying instead on visual impression of the fitted curves. For instance, Goodkind and Bicknell (2018) and Wilcox et al. (2020) used nonlinear models to qualitatively confirm that the relationship looked linear before using linear models for interpretation.

Second, there is considerable variability between individuals in reading times and other psychometric measures of language processing (see Farmer et al., 2012). While GAMs allow the fitting an overall effect while controlling for arbitrary nonlinear by-subject effects, previous studies have either not controlled for such effects, (Smith & Levy, 2013; Wilcox et al., 2020; Hofmann et al., 2022),[5] or assumed they were just constant offsets (Goodkind & Bicknell, 2018).

Third, all previous studies make strong assumptions about variance. Nearly all earlier studies have assumed that variance is constant, and normally distributed. A noteworthy exception is Hofmann et al. (2022), who used a Gamma-distributed response distribution, which instead encodes the assumption that variance increases proportional to the square of the predicted reading time value. Smith and Levy (2013) also mention that their results are robust to switching to an assumption of Gamma-distributed response, though they do not report results of this modelling choice. As far as we are aware, no previous study has explored the form of the effect surprisal has on variance in processing time.

Fourth and finally, many of the earlier studies that examined the shape of the linking function directly using GAMs, notably including Smith and Levy (2008a, 2013), used surprisal estimates from trigram language models, which are far from current state-of-the-art. Modern pre-trained LMs allow unprecedentedly accurate prediction of words in context (see e.g., Brown et al., 2020; Floridi & Chiriatti, 2020). While questions remain about the similarity between even the best modern LM's predictions and those of humans, numerous studies in this area have found that higher quality LMs (those better able to predict test data) make better predictors of processing difficulty (Frank, 2009; Fossum & Levy, 2012; Goodkind & Bicknell, 2018; Wilcox et al., 2020).[6] Additionally, recent work comparing architectures has found that surprisal estimates from Transformer-based LMs (Vaswani et al., 2017) tend to be the best predictors of psychometric measures (Hao et al., 2020; Merkx & Frank, 2021; Laverghetta et al., 2022).[7] Only one recent

[5] Smith and Levy (2013) did examine the nonlinear effect of surprisal on fixation time for eye-tracking data, fitting nonlinear GAMs for each subject separately, but not as random effects in a common model, and not for self-paced reading data, due to lack of a sufficient data to fit such models.

[6] However, some very recent work has begun to argue the opposite—that higher perplexity LMs or those using only limited context may be better psychometric models (e.g., Kuribayashi et al., 2022; Oh & Schuler, 2023a, 2023b). We will return to this topic in §2.6.

[7] Note, these studies mostly implicitly assume a linear relationship, using χ^2 or linear models' difference in log

published study—Wilcox et al. (2020)—has fit nonlinear GAMs of the linking function using surprisals from a modern Transformer-based LM (GPT-2 Radford et al., 2019).[8] While they found evidence broadly in favor of a '(near-)linear' linking function, they did not control for by-subject differences. Also, the surprisals they use are from versions of GPT-2 trained on much smaller datasets than the standard pretrained versions, and they do not provide the model with access to context outside of the current sentence. We will compare their results with ours in §2.6.

2.3.2 Theoretical arguments for linearity

A number of lines of work have given theoretical arguments in favor of a linear linking function between processing time and surprisal. Hale (2001) gave the original suggestion that processing effort was proportional the log ratio of prefix probabilities,[9] which is equal to surprisal:

$$\begin{aligned} \text{Time}(w_n) &\propto \log \frac{p(w_{1:n-1})}{p(w_{1:n})} \\ &= \log \frac{1}{p(w_n \mid w_{1:n-1})} = s(w_n) \end{aligned} \tag{2.9}$$

R. Levy (2005, §2.2.1), showed that the surprisal of a word is equal to the relative entropy between distributions over structures (such as parses, or meanings) before and after observing the word,

$$s(w_n) = D_{\text{KL}}(p(\cdot \mid w_{1:n}) \,\|\, p(\cdot \mid w_{1:n-1})) \tag{2.10}$$

assuming (crucially) that the structures consist at least in part of the words themselves. This provides an additional justification for surprisal theory, linking the processing difficulty of a word to a quantification of the amount by which the comprehender's beliefs must be updated to account for the observation. The relative entropy between such distributions appears in a number of theoretical analyses of algorithm runtime in Bayesian statistics, notably in the analysis of rejection sampling (Freer et al., 2010, and §2.2.3.1 above) and importance sampling (Agapiou et al., 2017; Chatterjee & Diaconis, 2018; Sanz-Alonso, 2018). However, in both cases the relationship between relative entropy and algorithm cost (number of samples needed) is exponential rather than linear. We are not aware of the analysis of any algorithm that leads to a linear relationship.

Other arguments for the linear linking function come from work which models the comprehender as a rational agent managing the cost of perceptually discriminating between possible

likelihood to assess psychometric predictive power.

[8] In recent unpublished work, Shain et al. (2022) conduct a new large-scale study of the linking function using multiple LMs, including modern pretrained Transformer-based models, using nonlinear continuous-time deconvolutional regressive neural networks (CDRNNs; Shain & Schuler, 2022), rather than GAMs. We discuss their results and preliminarily compare with ours in appendix A.4.

[9] Hale assumed prefix probabilities according to a probabilistic context-free grammar Earley parser, but this is not crucial to the intuition.

alternatives, or preparing resources (Smith & Levy, 2008a, 2008b, 2013; Bicknell & Levy, 2010, 2012). We will not review these arguments here; see R. Levy (2013) for more detail. In the context of our discussion, the important thing about all such arguments is that they are *computational-level* (in the sense of Marr, 1982). That is, they show that—subject to certain constraints—an optimal information processor would have cost that is linear in surprisal. However, none of these arguments provides a concrete algorithm for achieving this optimal behaviour in practice.

2.3.3 Superlinearity in surprisal theory

A number of earlier theoretical proposals have assumed a superlinear linking function between surprisal and processing time. For instance, Narayanan and Jurafsky (2004) conjectured that reading time is inversely proportional to incremental probability—that is, exponential in surprisal.

$$\text{Time}(w_n) \propto \frac{1}{p(w_n \mid w_{1:n-1})} = e^{s(w_n)} \tag{2.11}$$

Their justification for this linking function is based on a similar intuition to that of Hale (2001), but without assuming the logarithmic relationship. We note this relationship is also the one implicitly assumed by studies using linear models of log-transformed reading times (as in Boston et al., 2008; Roark et al., 2009; J. Mitchell et al., 2010; Aurnhammer & Frank, 2019; Merkx & Frank, 2021; Oh et al., 2022; Oh & Schuler, 2023a, 2023b).

Although much subsequent work has assumed a linear linking function, some of the earliest work in surprisal theory (R. Levy, 2005, §2.8.8) provided an argument for a *nonlinear* function, motivated by the uniform information density hypothesis (UID; see Jaeger, 2006; R. Levy & Jaeger, 2006). While the argument itself does not suggest an algorithm, and thus is not relevant to the present discussion, Meister et al. (2021) followed up on the suggestion, experimenting with a linking function of the form

$$\text{Time}(w_n) \propto (s(w_n))^k \tag{2.12}$$

where the parameter k was fit empirically. They report that their results are consistent with a somewhat superlinear linking function (k slightly larger than 1), when using surprisal estimates from high-quality pre-trained Transformer-based LMs.[10]

Models of sentence processing within the ACT-R framework (adaptive control of thought–rational; Anderson & Lebiere, 1998) also make claims about the relationship between the statistical

[10]Cf. Brothers and Kuperberg (2021) who recently presented evidence for a *sub*linear linking function, using cloze-probabilities (Taylor, 1953), not LMs, to estimate surprisal. Note however, cloze probabilities are in practice impossible to estimate for high-surprisal items (see R. Levy, 2008a; Smith & Levy, 2011), and LM surprisals generally give an empirically better fit to psychometric data (Hofmann et al., 2022). Recent investigation in (Shain et al., 2024, SI Appendix 1) found that the use of Cloze norming rather than language models for probability estimates is fully responsible for the sublinear relationship observed in Brothers and Kuperberg (2021)'s data.

properties of words and incremental processing times. In this framework, an item (such as a word) is recalled in an amount of time that is a function of its *activation*, A, as Fe^{-fA}, where $F > 0, f \geq 1$ are parameters. The activation, in turn, is assumed to model the log-odds of the item being needed (Anderson, 1991b, simplifying slightly). In accounts of sentence processing within this framework (such as R. L. Lewis & Vasishth, 2005; Jäger et al., 2015; Engelmann, 2016; Nicenboim & Vasishth, 2018; Vasishth et al., 2019; Engelmann et al., 2019; Vasishth & Engelmann, 2021; Dotlačil, 2021), the latency formula is taken as an assumption of the model, rather than being explicitly motivated by the intrinsic properties of an algorithm. It is worth noting, however, that the original work proposing this formula did in fact provide a way the formula could be related to the runtime of a serial search algorithm, which we discuss below in §2.6.3. Transforming the ACT-R latency formula from its usual form given above into a statement about surprisal rather than log odds[11] gives the following superlinear function of surprisal.

$$\text{Time}(w_n) = F(e^{s(w_n)} - 1)^f \tag{2.13}$$

When $f = 1$, as is often assumed, the latency formula then becomes simply the statement that retrieval time increases exponentially in surprisal.

Finally, other recent empirical work which may suggest superlinearity comes from van Schijndel and Linzen (2021) and subsequently Arehalli et al. (2022) who look at reading times in garden-path sentences. They fit linear models of the relationship between surprisal and reading time, and find that these models consistently underpredict the amount to which humans slow down in the critical region. This work is framed as challenging the assumption that reading time can be predicted solely based on incremental surprisal, but an additional interpretation of their results may be that the linking function is superlinear.[12] Results such as these also highlight the importance of using data with a broad range of surprisal values, since the items with high surprisal will be the most useful in distinguishing whether the shape of the linking function is linear or superlinear.

2.4 Empirical study

In the preceding sections, we argued that no existing theory of sentence processing provides an algorithmic explanation for processing scaling surprisal, and that a natural class of algorithms that do scale in surprisal are those based on sampling. However, these algorithms predict processing times that are superlinear in surprisal, in contrast to most of the existing literature on surprisal theory, which proposes the relationship is linear and generally assumes constant variance. Addi-

[11] Via the identity $\log \text{odds}(\cdot) = -\log(e^{-\log p(\cdot)} - 1)$. We believe we are the first to note this way of relating ACT-R's latency formula with surprisal theory.

[12] Note this interpretation does not necessarily contradict their framing, provided the human slowdowns they observe are larger than even the best-fit superlinear linking function could predict—see §2.6.

tionally, we identified a number of potential problems with earlier empirical analyses which found evidence of a linear linking function. All together, this motivates a re-examination of the empirical relationship, which we present in this section.

We use generalized additive models to predict reading times on the Natural Stories corpus (Futrell et al., 2021), using surprisal estimates from a variety of pre-trained language models, including modern Transformer-based models. In our modelling we control for nonlinear by-subject differences, and allow the effect of surprisal on variance in reading time to be fit empirically. We give a quantitative assessment of the superlinearity of the effect surprisal has on reading time and on variance in reading time.

2.4.1 Language models

To get estimates of incremental surprisal values, we use causal[13] language models (LMs)—statistical models of the probability of words given previous context. An LM M gives an estimate of surprisal as $s_M := -\log p_M(w_n \mid w_{1:n-1})$. We obtain surprisal estimates from a collection of LMs, listed in table 2.1. These include the following pre-trained Transformer-based LMs: Transformer-XL (TXL; Dai et al., 2019), GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021), GPT-J (Wang & Komatsuzaki, 2021), and GPT-3 (Brown et al., 2020). We also include two older, non-Transformer-based LMs: an LSTM-based model (Gulordava et al., 2018) and a Kneser-Essen-Ney smoothed 5-gram model (both from Boyce & Levy, 2020).

Context amount One of the main benefits of modern LMs is their ability to incorporate information from large amounts of previous context when making predictions. Different models allow differing amounts of preceding context (table 2.1, second column), and for the most accurate estimates of next-token probability, we provide each LM as many previous tokens as it can use. Since all ten stories in the corpus are between 1024 and 2048 GPT tokens in length, this means GPT-Neo, GPT-J and GPT-3 models will always have access to all preceding context in the story when making their predictions. For comparison, we also compute surprisals for each Transformer-based LM when provided only the tokens within the same sentence as the current token. In discussing results below, when we need to distinguish between the surprisals estimated from the same LM with differing amounts of context, we will refer to "within sentence" versus "maximum"-context surprisals. Restricting the amount of context can have a noticeable deleterious effect on language modelling accuracy.[14]

[13] We only consider unidirectional or causal LMs: models which predict words given previous context, without access to future context. Bidirectional or masked LMs are less appropriate for modelling incremental processing.

[14] Note however that some recent work has suggested that restricting context can increase psychometric predictive power: See discussion in §2.6.1.

model	max context (tokens)	number of parameters	pre-training data amount	log PPL
5-gram	5		90Mtok	6.4
LSTM	NA		90Mtok	4.9
Transformer-XL	NA	88M	100Mtok	4.2
GPT-2	1024	124M	40GB	3.4
GPT-2 large	1024	774M	40GB	3.0
GPT-2 XL	1024	1.5G	40GB	2.9
GPT-Neo	2048	2.7G	800GB	2.8
GPT-J	2048	6G	800GB	2.6
GPT-3 Ada	2048	*350M	300Gtok	3.0
GPT-3 Curie	2048	*6.7G	300Gtok	2.6
GPT-3 Davinci	2048	*175G	300Gtok	2.3

Table 2.1: Language Models used in this study, along with their max context size, number of trainable parameters, amount of pretraining data, and log perplexity score on Natural Stories corpus. For OpenAI GPT-3 models estimates (marked *) are deduced from evaluations (Gao, 2021).

Model quality To quantify language model accuracy we use *perplexity*—the standard measure of how well an LM predicts a test corpus. The logarithm of perplexity is the mean surprisal, the average uncertainty per word.

$$\mathrm{PPL}_M(w_{1:N}) = \left[\prod_{n=1}^{N} \frac{1}{p_M(w_n \mid w_{1:n-1})}\right]^{\frac{1}{N}}$$
$$\log \mathrm{PPL}_M(w_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} s_M(w_n)$$

A lower perplexity language model is one which can more accurately predict tokens given previous context. Note, the perplexity of two models is not directly comparable unless they have the same vocabulary. All eight GPT-type models we use are directly comparable.[15] The remaining three models (the LSTM, n-gram, and Transformer-XL) are not. For this reason, while we will use perplexity values for all models in discussion and figures to follow, we will only make direct comparisons of the GPT models.

2.4.2 Corpus

For our empirical analysis we use the Natural Stories corpus (Futrell et al., 2021), an English-language corpus which was released with self-paced reading time (RT) psychometric data. The

[15] All use the byte-level BPE tokenization scheme of GPT-2 (Radford et al., 2019).

corpus consists of 10 stories, of about 1000 words each. Each story is a modified version of publicly available text, edited to contain "many rare or marked syntactic constructions, ...while maintaining a high degree of overall fluency and comprehensibility." The relatively high concentration of rare constructions makes this corpus particularly appropriate for our study, since the difference between a linear and a superlinear linking function may only be appreciable in the high end of the surprisal range. Reading times released with this corpus were gathered from 181 native speakers, with each word in the corpus read by a median of 84 reading participants.

To allow inspection of the full text of the corpus, annotated with LM surprisals and reading times, we provide an interactive utility, linked in appendix A.5.

2.4.3 Generalized additive models

We fit GAMs to model the effect of surprisal on reading time. In particular, we use Gaussian location-scale mixed models (Rigby & Stasinopoulos, 2005; Wood et al., 2016) which allow us to model surprisal's nonlinear effect on mean RT, while also modelling its nonlinear effect on variance in RT, rather than assuming variance is constant or has a particular parametric relationship to the mean.

For each LM's set of surprisals, we fit a model we will call the **nonlinear GAM**, which predicts reading time, and variance in reading time (in the form of log standard deviation), each as an overall nonlinear function of surprisal, controlling for nonlinear by-subject variation and control predictors. It is these nonlinear GAM fits which we will use to interpret the relationship between surprisal and reading time. We also fit a minimally-different control model for each LM's surprisals, which we will call the **linear control GAM**, in which overall and by-subject effects of surprisal (for predicting both reading time and variance in reading time) are forced to be linear.

2.4.3.1 Model specification

In specifying the nonlinear GAMs, we include the following terms for the effect of surprisal and control predictors. To model the linking function we are interested in, we include a smooth term for the overall nonlinear effect of surprisal. To control for possibly nonlinear individual deviations from the overall curve, we include a by-subject factor-smooth interaction term. We also include a tensor product term for the nonlinear interaction between log-frequency and word length (following Smith & Levy, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020). Finally we include versions of all three above terms but for the previous word, to control for spillover effects (following Goodkind & Bicknell, 2018, 2021; Meister et al., 2021).

To predict variance (precisely, log standard deviation) in reading time, we include the same terms as above, though only for the current word, since there is no a priori reason to expect spillover in variance. So that the resulting overall curve fit by the model can be interpreted simply, we choose a relatively low number ($k = 6$) for the basis dimension, effectively limiting the maximum wiggliness

of the fitted curve.

For the linear control GAMs, we use the same model specifications as for the nonlinear GAMs above, but with the main surprisal smooth and factor-smooth interaction terms replaced with a linear parametric term and linear by-subject random effects (likewise for the previous word, and for the effect on variance). To differ only minimally from the nonlinear GAMs, we allow the terms for the interactions between length and frequency to remain nonlinear similar to the approach taken in Goodkind and Bicknell, 2018.

We give further details and discussion of the specification of GAMs in appendix A.3.[16]

2.5 Results

Figure 2.2 displays our main results, showing the relationship between surprisal and human reading time for each LM and context amount. Each curve represents the nonlinear GAM's fitted effect of surprisal on mean RT (top two rows, solid coloured lines), or on log standard deviation in RT (bottom two rows, dashed coloured lines). In each small plot, the linear linking function predicted by the corresponding linear control GAM is underlaid as a black dotted line. Density plots at the bottom of each plot for the mean effect show the distribution of that LM's estimated surprisal values. The curves for LMs with maximum context are plotted in blue (first and third rows); within-sentence context in red (second and fourth rows). LMs are ordered left-to-right by decreasing perplexity, given maximum context.

We first examine the effect of surprisal on RT (top set of plots). For all language models, reading time generally increases with surprisal. Impressionistically, better LMs (as measured by perplexity) appear to exhibit a superlinear relationship between surprisal and reading time, with higher quality LMs exhibiting more strongly superlinear curves (see below for quantification of this claim). By contrast, lower quality LMs (including the n-gram, LSTM, Transformer-XL), and models with only within-sentence context, tend to exhibit closer to linear relationships—or even *sub*linear relationships at high surprisal values (see §2.6). The slopes fit by the linear control GAMs are positive for all models.

Examining the relationship between surprisal and variance (as log standard deviation; bottom set of plots), we see a similar pattern. Variance in RT appears to generally increase with surprisal, with a few exceptions among the models with only access to within-sentence context. And for the linear controls, we generally see a positive slope for all fitted lines, similarly to the slopes fit by these control models for the effect on RT.

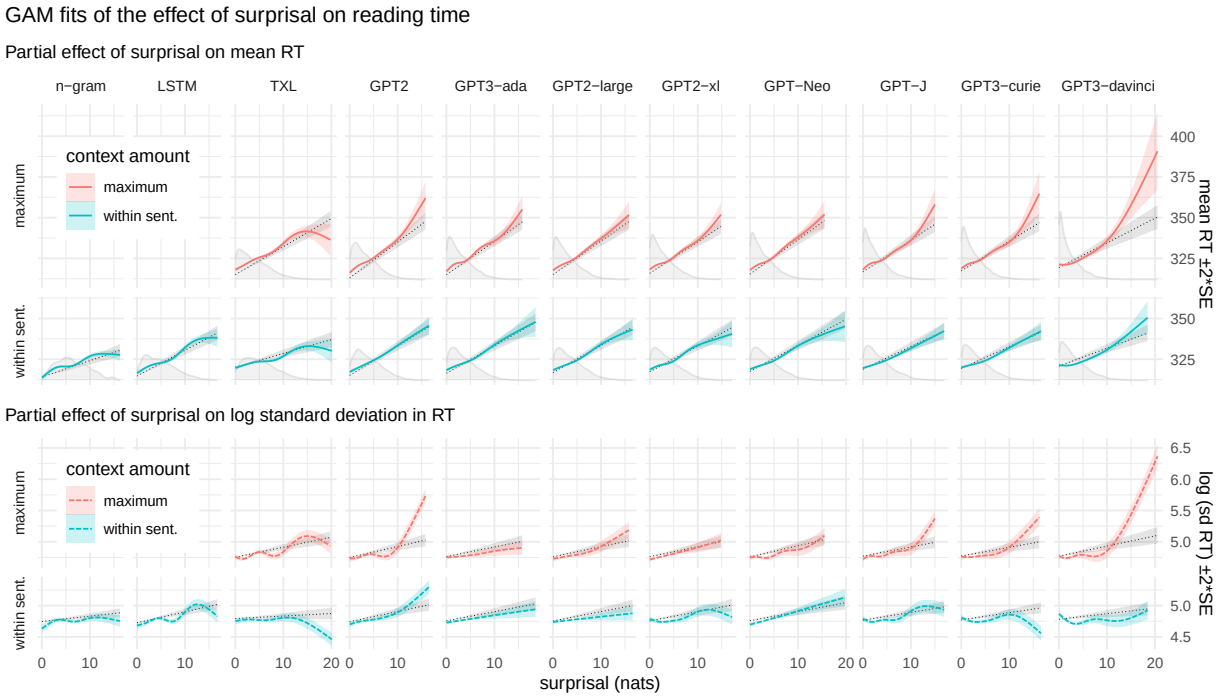


Figure 2.2: The effect of surprisal on self-paced reading time. Coloured lines are the fitted effects from the nonlinear GAMs, dotted black lines beneath are from the corresponding linear control GAMs. **Top two rows**: effect of surprisal on mean RT, with density plots of surprisal underlaid at the 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). Shaded regions represent 95% CIs.

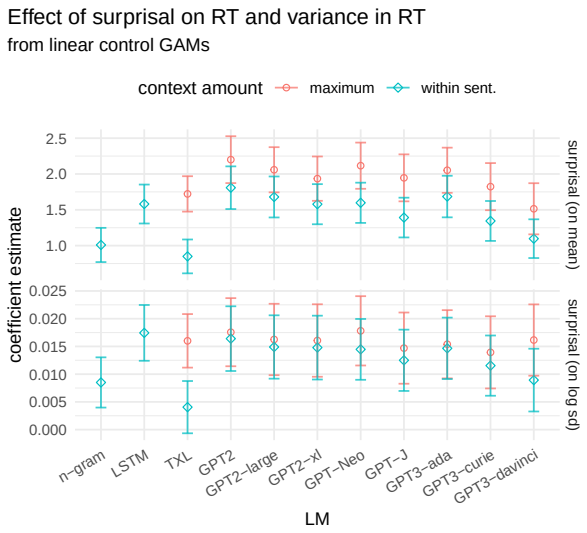


Figure 2.3: Coefficient estimates (with 95% CI) for the main effect of surprisal on RT and log standard deviation in RT, as fit by the linear control GAMs. For all LMs, both coefficients are positive, and significant ($p < 0.05$)—with the exception of the variance effect for Transformer-XL constrained to within-sentence context.

2.5.1 Quantifying the direction of the effect

To establish the overall direction of the effect, as well as replicate earlier work which used linear models for the effect on RT (though not variance), we will start by examining the slopes fit by our linear control GAMs. We use these models to get a quantitative interpretation of the overall direction of these effects, before introducing our superlinearity measure to examine the shape of the curve in the next subsection. Figure 2.3 provides the coefficients for the effect of surprisal. Each point describes the slope of the relationship between surprisal and RT (top) or log standard deviation in RT (bottom), with bars indicating 95% confidence intervals.

We observe that surprisal has a positive effect on RT for all LMs, consistent with the findings of the large number of previous studies of this relationship. This is also true for variance in RT: As surprisal increases, variance in reading time also increases, for all LMs and context amounts.[17] This is noteworthy, given that previous work has nearly universally assumed that variance is constant. Incidentally, we also note a general trend that the effect of surprisal on mean RT is larger when using LMs with access to full previous context compared to restricting to only within-sentence context,[18] though this is not true for the effect on variance in RT (with the exception of Transformer-XL).

[16]Scripts for data preprocessing and reproducing all results and figures will be made available in supplementary material.

[17]These coefficients are all significantly different from zero (at the 0.05 level), with the sole exception being Transformer-XL when only given within-sentence context, for which the coefficient is positive but not significant.

[18]However, this difference is only significant for TXL, GPT-Neo and GPT-J (at the 0.05 level)—for all the other models the difference is just barely beneath this threshold for significance.

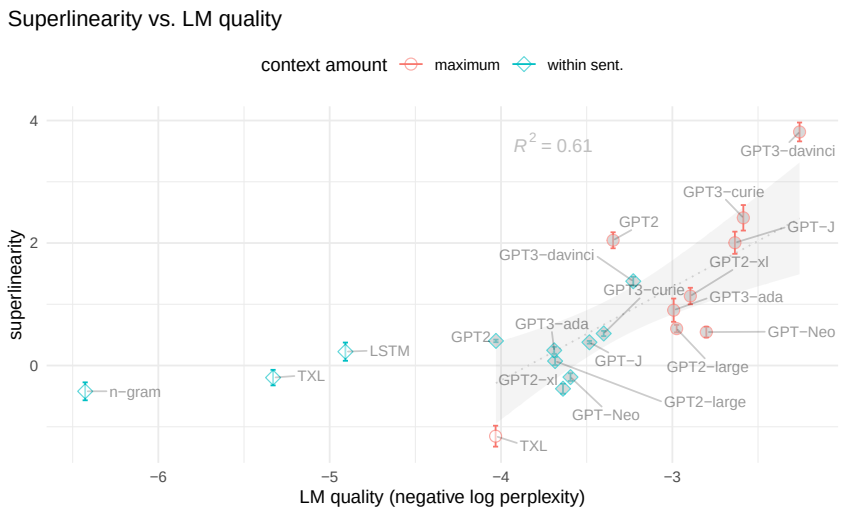


Figure 2.4: Superlinearity, measured as the amount by which the slope of the nonlinear GAMs' predictions at high surprisal exceeds that at lower surprisal, versus model quality (as negative log perplexity). The effect of surprisal on reading time is more superlinear for better LMs, as demonstrated by a best-fit regression line (dashed line with 95% CI shaded and correlation coefficient R^2 printed above). Note only GPT-based models (filled grey) are directly comparable by perplexity, hence the line describing this trend is fit on only those points.

2.5.2 Quantifying superlinearity

To quantify the observation that the relationship seems more superlinear for better quality LMs, we define a simple descriptive value which we will call *superlinearity*. This value is computed as follows: (i) split the surprisal range into two equal intervals, (ii) find the slope of the best linear approximation to the curve in each interval, and (iii) take the difference between these two slopes. A curve which bends upward will have positive superlinearity; one which bends downward will have negative superlinearity. For a relationship which is overall increasing[19] positive superlinearity indicates that the curve is increasing superlinearly in a global sense, though it may not be locally monotonic.

Figure 2.4 presents superlinearity plotted against LM quality (as negative log perplexity, so that higher values represent better LMs). Points for GPT-based models—which share a common tokenization scheme and vocabulary and are thus directly comparable by perplexity—are filled in grey, and a weighted linear regression fit on these points is displayed as a dashed line, with correlation coefficient printed above, and 95% CI shaded.

We see a clear correlation between an LM's quality and the superlinearity of the effect on RT. This correlation is evident visually, and is attested by the correlation coefficient $R^2 = 0.61$. This

[19]Note that this definition of superlinear doesn't imply increasing—a U-shaped curve would be superlinear. This is a reason for the previous analysis showing all effects were increasing.

provides a quantitative confirmation of our claim that the better the LM, the more superlinear the effect of surprisal on reading time.

2.5.3 Controls

In our modelling we chose to fit the effect of surprisal on variance, unlike previous work, which has often assumed constant variance. To check whether the superlinearity we see in the relationship with mean RT is dependent on this modelling choice, we fit models which assume constant variance. For this control, we assume a normally-distributed dependent variable and identity link (as is standard, following Smith & Levy, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020).[20] We found the relationships between surprisal and RT predicted by these models were similar to the results reported above. They exhibited increasing nonlinearity with increasing LM quality (plots from these models, and further details, are in appendix A.7.1).

In our models, we controlled for spillover effects by including predictors for one previous word (following e.g., Goodkind & Bicknell, 2018, 2021; Meister et al., 2021). However, other studies (including Smith & Levy, 2013) have argued for using up to 3 previous words. To understand whether this choice is likely to have influenced our general results, we include additional analyses in appendix A.7, examining autocorrelation in residuals and fitting models with predictors for three previous words, rather than one. We find there is little evidence to suggest that additional spillover predictors would have a large effect on our main qualitative results.

In order to understand the degree to which our results are dependent on nonlinear by-subject effects we include, we experimented with fitting models as above, but in which we removed the terms controlling for by-subject effects. We found that this modification resulted in predicted relationships that were similar in shape, but with much wider confidence intervals. This suggests that controlling for by-subject variation in this data gives us higher power to detect population-level nonlinear effects. This control is also useful for comparing our results with previous literature which did not include by-subject random effects (e.g. Fernandez Monsalve et al., 2012; Smith & Levy, 2013; Wilcox et al., 2020; Hofmann et al., 2022). Not controlling for by-subject variation may be one reason why such studies did not find evidence of a nonlinear effect.

As is readily evident in the density plots of surprisal values (plotted in fig. 2.2, top two rows), the overwhelming majority of words have relatively low surprisal. This is especially true for the lowest-perplexity LMs. To check that the shape of the curves we see are not being determined by a few high-surprisal outliers, we carried out two controls. First, we carried out a cross-validation, refitting GAMs for each of the LMs on 6 folds of the data.[21] We found that the degree of superlinearity

[20] The assumption of constant variance could also be relaxed by only partially, by assuming a specific parametric relationship between mean and variance. See details in appendix A.3.5.

[21] We also note that the evaluation technique used to fit GAMs is designed to control against such sensitivity to outliers (see discussion in Wood, 2011).

in the results was consistent across folds, confirming that the results are not driven by a small number of outliers (see appendix A.7). Second, focusing on the most superlinear GAM, which also has the most drastically skewed distribution of surprisals (GPT-3 Davinci), we performed a hand-inspection of the highest-surprisal words, and found that most occur within the kinds of rare syntactic examples that Natural Stories was designed to contain, but otherwise seem plausible in context, and therefore do not seem to be outliers in any way which should have required their removal from our data (see appendix A.6 for a complete list of these words in context and further discussion). We then re-fit GAMs with the highest surprisal items removed. We found that superlinearity was somewhat reduced (due to truncating the range of surprisals), but curve remained superlinear.

2.6 Discussion

In the first part of this paper, we investigated the runtime characteristics of inference algorithms that iteratively sample from the prior—a natural example of a broad class of algorithms whose runtime scales with surprisal. As we showed, simple examples of such algorithms predict that both runtime and variance in runtime increase with surprisal, the former superlinearly. In the second part, we carried out an empirical study to test these predictions, finding that for one widely-studied dataset the empirical relationship between surprisal and processing time is broadly consistent with these predictions when using the best-available LMs to estimate surprisal. In this section we discuss the implications of these results.

The correlation we observe between LM quality and superlinearity suggests that one reason why a superlinear relationship has not been detected in earlier work may simply be due to the use of surprisal estimates from earlier language models, which were less accurate. For example, as discussed in §2.3.1, Smith and Levy (2008a, 2013) found empirical support for the linear linking function, using a trigram model to estimate surprisal. Our results confirm their finding for this type of LM, showing no evidence of superlinearity for the n-gram model. Wilcox et al. (2020) also presented evidence of a linear linking function, using some higher quality LMs and multiple datasets, including the Natural Stories corpus. However, their highest-quality LM was a GPT-2 model trained on much smaller datasets than the pretrained GPT-2 model we use,[22] and they estimate surprisals using only within-sentence context. Both choices likely mean less accurate predictions in general (higher perplexity), although they do not report perplexity values. As our results demonstrate, using LMs restricting to only within-sentence context, and using higher-perplexity LMs in general, tends to reduce the superlinearity of the relationship.

[22]They use versions of GPT-2 trained on multiple different datasets, with the best model they use being trained on 42 million tokens, compared to the ∼40GB (roughly 10 billion tokens) of training data for the GPT-2 model which we use.

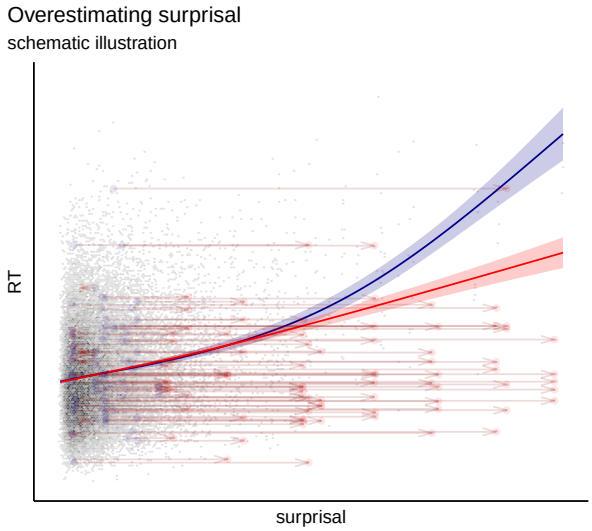


Figure 2.5: Diagram illustrating schematically how a superlinear relationship may look linear if some surprisal values are systematically overestimated: When a subset of points are moved higher in the surprisal range (indicated by arrows), the best-fit curve becomes less superlinear (blue to red).

This tendency is consistent with the following interpretation, illustrated schematically in fig. 2.5. The blue curve represents the best-fit curve for reading time as a function of hypothetical 'true' surprisal, and the red curve represents the best-fit curve after raising the surprisal values assigned to a subset of observations (while keeping their reading times the same). A lower quality (higher perplexity) language model will tend to overestimate surprisal in general (since log perplexity is simply average surprisal). If an LM consistently overestimates surprisals compared to humans in such a way, we would expect the resulting best-fit linking function to be lower than it should be at the higher end of surprisal range, due to these items with low reading time being (wrongly) assigned high surprisal.[23] As illustrated in the diagram, such underestimation (moving these points rightward) results in changing the best-fit curve from superlinear (blue), to linear (red). This is what we see in our results; the lower quality LMs display less superlinear relationships (or even sublinear ones in some cases, especially those restricted to only within-sentence context). Under this interpretation, the superlinearity we observe in our results stems from our using more accurate surprisal estimators and, in particular, models which can make best use of large amounts of previous context to accurately predict words.

An additional factor that may explain why superlinearity has not been observed in previous studies that fit GAMs to describe this relationship is that most did not control for by-subject

[23]One way this may occur for an LM with restricted access to context, for instance, is when it it consistently assigns high surprisal to see some uncommon words in a text where, given the context, they are not surprising to humans, who have a good model of the topic being discussed.

variation (Smith & Levy, 2013; Wilcox et al., 2020; Hofmann et al., 2022), or assumed that such variation could be modeled by a constant offset (Goodkind & Bicknell, 2018). As described in the previous section, our experiments lesioning the by-subject random effects from our GAMs resulted in models which were much less confident about the shape of the curve, even for the more accurate LMs.

As mentioned in §2.3.3, a recent line of work introduced in van Schijndel and Linzen (2021) has examined garden path effects, where humans show increased processing difficulty at the point in a sentence where temporary structural ambiguities are resolved in favor of the less expected alternative. Van Schijndel and Linzen (2021) and Arehalli et al. (2022) argue that the degree of slowdown that occurs in humans exceeds that which can be predicted by linear linking function. We propose that intuitively, a superlinear linking function (such as those we see in our results) should be able to predict a larger slowdown the than a linear one, and thereby at least partially explain the human slowdown observed in their study. However, in the current study, our focus is on determining the best-fit form of the linking function broadly. We don't necessarily predict that the general superlinear trend we see in our results (for GPT-3 Davinci, for instance) should be sufficient to entirely explain the human reading times on particular sentences, where many other factors specific to that particular sentence may influence human reading times. However, with proper controls, examining the degree to which a superlinear linking function can explain human processing on particular grammatical constructions (including garden path sentences) is a promising direction for future work.

2.6.1 Language model perplexity and quality as psychometric models

In this work, we use pre-trained LMs as the best-available approximators of the true predictability of individual words—the quantity which should describe the behaviour of an optimally rational comprehender. The interpretation of our results relies on the assumption that more accurate LMs provide better estimators of human surprisal, at least for those words which drive the superlinear fit of our GAMs. As discussed above, this assumption is supported by recent literature (Goodkind & Bicknell, 2018; Wilcox et al., 2020; Hao et al., 2020; Merkx & Frank, 2021; Laverghetta et al., 2022). Very recently, however, another line of work has emerged arguing that, to the contrary, lower perplexity LMs sometimes provide *poorer* fits to psychometric data. Building on a preliminary observation in Oh et al. (2022), Oh and Schuler (2023a) present a study of three different families of Transformer-based LMs (GPT-2, GPT-Neo, and OPT; S. Zhang et al., 2022), finding that the lower-perplexity LMs in each family tend to have poorer psychometric predictive power. In related work, Kuribayashi et al. (2022) report that for GPT-2 and LSTM models, psychometric predictive power increases as access to context is restricted, in English and Japanese. This improvement in psychometric predictive power continues even for extremely severe restrictions such as limiting

context to just one previous word.

These studies raise two important problems to be explored in future work. First, it is important to understand which subsets of words drive the two effects (psychometric power and superlinearity) and how much they overlap. If the words driving the decrease in psychometric power are not the same as those driving the superlinearity effect, then these studies and our own may be complementary. For example, Oh and Schuler (2023a) show that named entities and predicative adjectives are among the classes of words most responsible for the decreasing psychometric predictive power. Intuitively, better LMs may underestimate how surprising these items are to people because the LMs are trained on superhuman quantities of data. It is possible for a model to find such words much less surprising than humans, while improving the psychometric fit of other classes of words, such as function words. If the latter classes of words are those most critical for superlinearity, then both effects could very well hold. Determining whether this is or is not the case requires a detailed sensitivity analysis that carefully matches datasets, LMs, and analytical models. We leave this to future work.

A second, and more important, question is whether these recent results are an artefact of using linear models to study the relationship between surprisal and processing time. Our analyses above show that the lower-perplexity a model is, the greater the advantage of a superlinear linking function over a linear one. Studies such as Kuribayashi et al. (2022) and Oh and Schuler (2023a) make use of linear linking functions,[24] showing that lower perplexity LMs predict psychometric results more poorly. However, if the true relationship between surprisal and processing time is nonlinear, then the seeming decrease in psychometric predictive power that they report might even be related to the increasing superlinearity that we observe. A large-scale examination of the relationship between LM perplexity and psychometric predictive power using nonlinear regression models such as GAMs would provide a useful contribution to more fully understand the potential three-way relationship between LM accuracy, psychometric predictive power, and superlinearity.

2.6.2 A particle filter model

To our knowledge, the only explicit sampling-based model of incremental sentence processing to date is the approach presented in R. Levy et al. (2008). Their model uses particle filtering, a standard sequential Monte Carlo (SMC) technique based on importance sampling (Doucet et al., 2001; Doucet & Johansen, 2008). The parsing algorithm estimates the posterior distribution $p(z_n \mid w_{1:n})$ with a collection of K weighted particles (partial parses). Each of these particles is first

[24] Though this picture is complicated by differing choices on whether to log-transform the reading times before fitting models (as discussed above): we do not transform, nor do Kuribayashi et al., while Oh and Schuler do. Note, Shain et al. (2022) also observe that GPT-2 performs better than GPT-3 and GPT-J overall, though their study is aimed at determining the shape of the linking function, not the relationship between perplexity and psychometric power—see appendix A.4 for further discussion.

obtained by sampling from the prior $p(z_{n-1} \mid w_{1:n-1})$. Then each particle is updated according to an incremental transition distribution $p(z_n \mid z_{n-1})$, and weighted proportional to how likely it is to explain the next observation (word): $p(w_n \mid z_n)$.[25] Because their algorithm uses a fixed number of particles (the beam width, K), the number of samples drawn is identical at every word. Thus, this algorithm's runtime does not directly depend on surprisal in the way that the algorithms that we examined above do.

However, R. Levy et al. offer an analysis of processing difficulty which can be related indirectly to the present work. Rather than relating difficulty to runtime via expected number of samples, they relate processing difficulty at a particular word to the probability of failure at that word—that is, the probability that none of the particles in the beam can be extended to explain that word. They estimate this quantity by running the particle filter multiple times and counting the proportion of runs where the set of particles contains no successful parses.

This probability of failure is directly related to our analysis in §2.2.3.1, where runtime is inversely proportional to the probability of success (one minus the probability of failure). In the particle filter, the probability of success at step n is the probability that at least one particle contains a successful parse for w_n. If the particles are sampled from the *exact* posterior $\Pr(\cdot \mid w_{1:n-1})$, the number of such samples required for an accurate approximation of the posterior $\Pr(\cdot \mid w_{1:n})$ scales as $e^{s(w_n)} = 1/\Pr(w_n \mid w_{1:n-1})$.[26] In the particle filtering setup, which estimates the posterior distribution using importance sampling from an *approximate* prior, the expected number of samples required to integrate w_n is at least $e^{s(w_n)}$.[27] This suggests that a modified version of the particle filtering model, where variable numbers of samples were drawn until some desired number of successful parses were obtained, would have runtime that scaled naturally in surprisal. Examples of this type of modified approach to particle filtering include adaptive beam width algorithms (such as Fox, 2003; Buys, 2018; Elvira et al., 2017), which allow the number of particles (K) to vary at each step in order to maintain a criterion such as a bound on probability of error, or uncertainty of the model. Such algorithms could potentially be natural for use in models of sentence processing, and would have the property that higher surprisal words would require (exponentially) more samples.

[25] The algorithm is recursive, so the representation of the prior $p(z_{n-1} \mid w_{1:n-1})$ is itself an estimate of the posterior from the prior step, computed using samples from $p(z_{n-2} \mid w_{1:n-2})$, etc.

[26] This can be seen by first recalling that surprisal equals the relative entropy between prior and posterior (R. Levy, 2005)—again, assuming that the full parses consist at least in part of the words themselves. Then, note that in importance sampling, the number of samples required for accurate estimation scales as the exponent of precisely this relative entropy (see Chatterjee and Diaconis, 2018, Thm. 1.2, also discussed in Agapiou et al., 2017; Sanz-Alonso, 2018).

[27] Given the approximate prior makes predictions that are on average no better than the true prior, the expected number of samples will be no smaller than the expected number from the true prior.

2.6.3 Deterministic search algorithms

Besides nondeterministic sampling algorithms, we identified a related class of deterministic algorithms whose runtime scales in surprisal: those involving probability-ordered search.[28] In particular, probabilistic pruning (where only the high prior-probability parses are kept) has the potential to predict a monotonic increasing relationship with surprisal. Such methods (like beam search; Y. Zhang & Clark, 2008), have seen extensive use in parsing literature (see e.g., Jurafsky, 1996; Roark et al., 2009; Bouchard-Côté et al., 2009; Vieira & Eisner, 2017; Meister, Cotterell, & Vieira, 2020; Meister, Vieira, & Cotterell, 2020), yet as far as we are aware, there are no results relating these specific algorithms' time complexity to surprisal or incremental probability.

As noted above in §2.3.3, one simple and specific deterministic algorithm which can predict runtime increasing as a function of surprisal is the serial search mechanism assumed in the rational analysis of memory and ACT-R literature (Anderson, 1990; Anderson & Lebiere, 1998). The formula for reaction time in this framework was originally derived under the assumption that items in memory are considered in order of decreasing need probability. If each item requires a fixed amount of time, the runtime is simply the ordinal position of the item in a probability-ordered list.[29] Using this argument, along with the assumption that item need-odds are power-law distributed,[30] Anderson and Lebiere (1998) derived the latency formula linking (log) odds to run time exponentially as Fe^{-fA}. As noted above, this can be restated as $F(e^{s(w_n)} - 1)^f$—a superlinear function of surprisal (eq. 2.13). For this derivation, and a similar one assuming Pareto-distributed probabilities rather than odds, see (appendix A.8).

The upshot of this analysis (independent of the specifics of the ACT-R framework) is that the runtime of simple probability-ordered search makes a concrete prediction about the linking function with surprisal. And, this prediction is similar to the predictions of sampling algorithms we have discussed. However, unlike the sampling-based mechanisms we explored, a deterministic ranked-search mechanism such as this cannot predict nonzero variance in any intrinsic way.[31]

Conclusion

In this chapter, we have considered inference algorithms that involve iteratively sampling from a prior, and proposed that such mechanisms provide a plausible framework for formalizing theories

[28] This is not necessarily a separate class of algorithms in any discrete sense, but rather may potentially be viewed as a special subset of sampling algorithms, since any deterministic algorithm can be framed as sampling from delta functions.

[29] The original argument (Anderson, 1990, ch. 2) predated ACT-R . A modified version for ACT-R, which is stated in terms of activation rather than need probability is given in Anderson and Lebiere (1998, app. 3B).

[30] This assumption is very similar to our assumption that item weights are Pareto-distributed, in our analysis in §2.2.3.2.

[31] In the ACT-R framework, in practice, a noise term is added to the basic latency formula, but this is not motivated by the deterministic search algorithm used to derive the basic formula.

of incremental processing, since their complexity naturally depends on the predictability of their input. Analyzing simple representative examples of this class of algorithms, we found that the number of samples required scales superlinearly as a function of surprisal, with variance also increasing. In our empirical study of human reading times we found evidence of a linking function consistent with these predictions, when using surprisal estimates of the most accurate modern LMs.

Acknowledgements

We thank our anonymous reviewers for their detailed comments and suggestions. We also thank Roger Levy, Cory Shain, Jakub Dotlačil, Michaela Socolof, Benjamin LeBrun, and the members of the Montréal Computational and Quantitative Linguistics Lab (MCQLL), and MIT Probabilistic Computing Project (ProbComp) for feedback.

This research was enabled in part by resources provided by Mila (mila.quebec), Calcul Québec (calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca). We also gratefully acknowledge the support of the Canada CIFAR AI Chairs Program, the Centre for Research on Brain Language and Music (CRBLM.ca), the Natural Sciences and Engineering Research Council of Canada, and the Fonds de recherche du Québec - Société et culture (FRQSC).

Note introducing chapter 3

As outlined in chapter 1, there are two basic assumptions required in order for standard surprisal theory (hypothesis 1.1) to be equivalent to the claim that processing cost scales with the size of the Bayesian belief update, quantified as divergence from prior to posterior (hypothesis 1.4). One assumption is that this divergence is always equal to surprisal; the other assumption is that the linking function between surprisal and processing cost is linear. In the manuscript presented above in chapter 2, we followed previous literature in assuming the former assumption, and investigated the latter, presenting arguments for and evidence of a nonlinear linking function.

In the manuscript presented in this next chapter, we shift our focus to question the assumption that KL and surprisal are always equivalent, arguing that there are cases in which we can expect surprisal and KL to differ substantially. We motivate typographical errors as an exemplary test case, and investigate human processing cost on such items in a self-paced reading time study.

3

When unpredictable does not mean difficult to process

The defining claim of surprisal theory consists of the hypothesis that the effort required to process a word is proportional to its surprisal—the negative log of its probability, in context. A primary justification for surprisal theory relies on a proof that surprisal is equivalent to the Kullback-Leibler (KL) divergence between prior and posterior distributions, a quantification of information gain under the perspective of processing as incremental probabilistic inference. However, this proof holds only with a critical assumption: that structures in the space of possible interpretations are deterministically related to the observable words. In this work we propose that, during reading, minor typographical errors provide a useful example of the type of situation where the surprisal theory's crucial assumption is likely to be violated: they can be very unlikely in their precise form, but may not cause a commensurate amount of effort to process. Thus these examples form an ideal testing ground for the predictions of surprisal theory versus what we refer to as KL theory—the hypothesis that processing effort scales with KL, even when this is not equal to surprisal. We present a self-paced reading time study to estimate human processing cost on typographical errors in controlled environments, and find evidence that the general pattern of human processing effort follows the predictions of KL theory, rather than surprisal. We validate these results against surprisal estimates from large language models.

3.1 Introduction

A key aspect of expectation-based theories of processing is the idea that the processing cost of a word is a reflection of the information gained upon observing it. Surprisal theory captures the intuition that a word which closely matches the comprehender's expectations requires very little effort to process, and that the less a word corresponds to these expectations, the more work must be done in order to incorporate the observation (Hale, 2001; R. Levy, 2008a). This intuition is formalized within the perspective of processing as incremental probabilistic inference, where the comprehender's interpretation of an utterance can be represented by a distribution over the space of possible structures or meanings. In this setting, the effort necessary to process a word can be quantified by the amount by which this distribution changes when the word is observed. This change is measured as the Kullback-Leibler (KL) divergence, also known as relative entropy, between prior and posterior distributions. Provided this divergence equals surprisal, as is assumed in previous work (following R. Levy, 2008a), this information-cost interpretation provides a justification for surprisal theory's central hypothesis.

Yet, as discussed in chapter 1, surprisal's equivalence to this KL divergence does not hold in a general setting. Equivalence between these two quantities requires assuming that the latent structures can be deterministically mapped onto observable words. In a more general setting, where the relationship is nondeterministic, surprisal merely provides an upper bound on KL, and may be arbitrarily larger than it. This fact prompts two questions: In what type of situation, if any, can we expect that surprisal would differ from KL to a meaningful extent? And, in a situation where they are not equivalent, should we expect that processing cost scales with surprisal or KL?

If processing cost reflects information gain (a perspective that forms a main motivating justification for surprisal theory), this implies that cost should be different for items which result in different sized KL divergences, even if they have similar surprisal values. In this work we propose that minor typographical errors provide an exemplary test-case where we may expect KL to differ systematically and substantially from surprisal. 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. In this work we design a data set of words in controlled contexts designed to make the words' meaning highly expected or highly unexpected, and carry out a self-paced reading time study on these words, both with and without typographical errors, to investigate the effect on processing cost, and determine whether this effect is better explained by surprisal or KL. We find that the patterns of human processing cost do not behave as predicted by surprisal, but rather follow what would be expected under a KL theory of processing cost.

3.1.1 Motivation: surprisal versus KL divergence

As discussed in chapter 1, the equivalence between surprisal and KL divergence does not hold in a general setting without making any assumptions about the relationship between the observed word $\breve{w}$ and the latent structure Z. Rather, in general, surprisal $s(\breve{w}) := -\log p(\breve{w})$ only provides an upper bound on the divergence $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)$, as given by the following equation (repeating eq. 1.6).

$$s(\breve{w}) = D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) + R(\breve{w}) \tag{3.1}$$

This equation describes a partition of surprisal into two nonnegative components, the first being the relative entropy, and the second being $R(\breve{w}) := \mathbb{E}_{p_{Z|\breve{w}}}\left[-\log p(\breve{w} \mid z)\right]$, the expected negative log likelihood under the posterior—which we will refer to as the *reconstruction information*. In order for surprisal to be universally equivalent to KL, this second term must always be zero, which requires that the likelihood $\mathrm{lik}_{\breve{w}}(z) := p(\breve{w} \mid z) = 1$ everywhere in the support of the posterior. This is certainly the case if the latent Z is assumed to range over structures that deterministically map to observable words, as is explicitly assumed by R. Levy (2008a, §2.1) when giving the proof of surprisal's equivalence to KL.[1]

However, it is plausible to consider that in general the relationship between latent structures and observable words may not always be deterministic, and thus surprisal may not equal KL—in particular, the case of typographical errors or other malformed input provides a potential example. So, in a situation where they are not equivalent, the question becomes which of the two quantities drives processing cost. Explicitly, we can contrast general surprisal theory, with our KL theory:

Surprisal hypothesis

$\mathrm{cost}(\breve{w})$ increases as a function of $s(\breve{w}) = D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) + R(\breve{w})$

KL hypothesis

$\mathrm{cost}(\breve{w})$ increases as a function of $D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) = s(\breve{w}) - R(\breve{w})$

The former is the perspective taken in the literature on surprisal theory, by definition. Yet, as argued above, the latter may fact be more clearly motivated from the perspective of cost as information gain, an original justification of surprisal theory (R. Levy, 2013).

[1] Specifically, he stipulates that Z ranges over structures "each consisting at least partly of surface strings to be identified with serial linguistic input," implying $p(\breve{w} \mid z) = p(\breve{w} \mid z, \breve{w}) = 1$ everywhere in the support of $p_{Z|\breve{w}}$, and thus $R(\breve{w}) = \mathbb{E}_{p_{Z|\breve{w}}}\left[-\log p(\breve{w} \mid z)\right] = \mathbb{E}_{p_{Z|\breve{w}}}\left[-\log 1\right] = 0$.

3.2 Typographical errors as a case study

To understand how the KL hypothesis differs from surprisal theory it will be necessary to look at constructions where the predictions are likely to differ most drastically. Typographical errors, particularly when they are relatively minor, and occur on highly predictable words, provide an interesting case study, since it is intuitively plausible that despite a word having a meaning that is very expected in context, any particular typo of that word will be unpredictable (high surprisal), while carrying very little information about the meaning of the utterance (low KL).

For example, consider the following sentence. *After tripping over the rug in front of everyone, the student felt deeply embarrassed.* In the context, the final word, *embarrassed*, is predictable, and should require relatively little effort to process. Now consider a simple letter-transposition error, swapping an *s* and an *a*, resulting in a non-word, *embarrsased*. In the same context, this observation contributes roughly the same information in terms of how it will cause a person who is reading for comprehension to adjust their understanding of the intended meaning of the sentence. That is, the posterior distribution on meaning for the utterance after observing the word should be similar with or without the typo. Thus, intuitively, a KL theory of processing cost should predict similar difficulty for these two cases.

Traditional descriptions of surprisal theory, conceived of explicitly within a probabilistic grammar as in Hale (2001) and R. Levy (2008a), predict potentially infinite surprisal for a malformed or out-of-vocabulary word: If the grammar cannot generate such an observation, it has zero probability, and therefore infinite surprisal. In general, suppose there is a joint distribution described by the probabilistic graphical model $Z \rightarrow W$ where Z ranges over latent structures, and W is observations (words) generated from such structures via an emission or yield function described by the family of conditional distributions $\{p(w \mid z)\}_{z \in \mathcal{Z}}$. The surprisal of a particular observation $\breve{w}$ is the negative log of the expectation of the likelihood $p(\breve{w} \mid z)$ under the prior over Z:

$$s(\breve{w}) \coloneqq -\log p(\breve{w}) = -\log \int_{\mathcal{Z}} p(\breve{w}, z)\, \mathrm{d}z = -\log \mathop{\mathbb{E}}_{pz}[p(\breve{w} \mid z)] \tag{3.2}$$

In a traditional probabilistic grammar with smoothing (so that the model assigns nonzero probability to all possible observations), a nonword typo should have finite but still very high surprisal, because its expected likelihood would be near zero (any particular typo being unlikely). Importantly, this remains true even in an accurate noisy-channel model of surprisal with a likelihood that models true production-error probability: the probability of a particular typo like *embarrsased* will necessarily be very small, even if it is much more likely than other non-word strings, for the simple reason that the probability mass assigned to malformed versions of the intended word, must

context After tripping over the rug in front of everyone, she quickly got up, but her cheeks turned red and she felt deeply `[target]` as she walked carefully back to her chair.

target

Condition 1. `expected`	embarrassed
Condition 2. `unexpected`	innovative
Condition 3. `expected_typo`	embarrsased
Condition 4. `unexpected_typo`	innovaitve

Figure 3.1: Example context and set of targets for each of the four conditions. Condition 1: `expected`—the target word is very predictable given the pre-target context; condition 2: `unexpected`—the target word has a meaning that is unexpected in the context; condition 3: `expected_typo`—the target is the same as in the `expected` condition, but with a typographical error introduced; and condition 4: `unexpected_typo` likewise.

be spread across multiple possible ways this malformation might be realized.[2] Yet, a KL theory of processing cost predicts a typo's having low cost whenever it doesn't result in a large change in the interpreted meaning. Mathematically, this is expressed in the quantity $R(\breve{w}) = \mathbb{E}_{p_{Z|\breve{w}}}\left[-\log p(\breve{w} \mid z)\right]$ which will cancel most of the bits of surprisal in such a case. This matches intuition: Under the posterior, the expected likelihood of the typo remains small. Observing the non-word *embarrsased* contributes to an understanding of the meaning of the utterance, but it is indeed not the most predictable way of expressing this meaning!

The ideal situation in which to distinguish whether effort is driven by KL or surprisal would be one where the KL is identical across conditions, but surprisal is manipulated. For this purpose, comparing processing effort on a highly predictable word, with and without a typographical error, provides precisely this kind of situation.

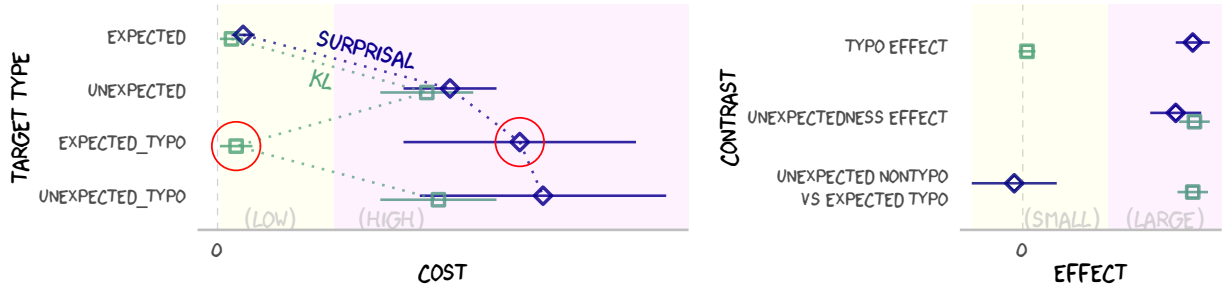
3.2.1 Experiment

In order to investigate processing on such words, we created a data set of example sentences containing with target words in identical contexts, for each of four conditions, either an expected or unexpected meaning, and with or without a typo: $\{\text{expected}, \text{unexpected}\} \times \{\text{nontypo}, \text{typo}\}$, as illustrated with an example in fig. 3.1.

3.2.1.1 Predictions of surprisal and KL

Figure 3.2 provides schematic sketches of the patterns predicted under surprisal theory versus KL theory. KL predicts a zig-zag pattern in cost across the four conditions: For a word expressing an expected meaning, KL theory predicts processing to be easy, whether or not it contains a

[2] This type of accuracy might plausibly be expected from the best modern language models, based on the typos and malformed words encountered in training data.



(a) Sketch of predicted processing cost in each condition. (b) Sketch of predicted contrasts of interest.

Figure 3.2: Sketches of the predictions of KL (▫) vs surprisal (◇) theory about processing cost. **Left (a):** An `expected` word will be low cost under surprisal or KL theory, and likewise an `unexpected` should be be high. Surprisal should also be high for a typo whether it is on an expected or unexpected word, and KL should be high for an unexpected word, whether or not it has a typo. The place where predictions differ is for the `expected_typo` condition (circled). **Right (b):** Sketch of comparisons between conditions implied by fig. 3.2a. Surprisal theory predicts a positive typo effect and likewise a positive unexpectedness effect. By contrast, KL theory distinguishes between words with typos versus word with unexpected meanings, with only the latter having a large effect on cost. When comparing an unexpected word without a typo to an expected word with a typo, the difference should be large and positive according to KL theory, while the difference in surprisal should be small or even negative.

typo, and conversely, a word with an unexpected meaning is predicted to be effortful. Under surprisal theory, an expected word should be low cost, but the other three conditions should all be substantially higher cost, potentially with typos causing even higher cost than simply unexpected words. Figure 3.2a shows these patterns in expected for cost for each of the target type conditions under surprisal theory versus KL theory.

Of particular interest to us are the following questions relating to contrasts between our conditions, in terms of their predicted effect on human reading time response under surprisal theory versus KL theory.

(**A**) What is the general effect on processing cost when encountering typo versus nontypo, marginalizing over the expectedness of the word? [Call this the **typo effect**.] Surprisal theory predicts this to be large, if typos are unpredictable. KL theory on the other hand, predicts it to be small or nonexistent, if the typos are minor enough to not cause a similar change in belief as the corresponding word without a typo.

(**B**) What is the general effect of encountering unexpected versus expected meaning marginalizing over whether or not there is a typo? [Call this the **unexpectedness effect**.] Surprisal theory predicts this to be large. KL theory does likewise.

(**C**) What is the difference between an unexpected word without a typo, versus an expected word with a typo? [This is unexpectedness effect versus typo effect.] This contrast compares

two conditions where the typo effect and unexpectedness effect work against each other, measuring the extent to which the unexpectedness effect overpowers the typo effect. KL theory predicts this difference to be large, whereas under surprisal theory the prediction is less clear, but it would be plausible for it to be small or even negative.

Figure 3.2b illustrates the predictions about these three effects of contrasts between conditions, for surprisal (the first large and positive, the third lower or even negative) versus KL (the first near zero, the second two large and positive).

3.2.1.2 Target specification

In order to isolate the effects we are interested in, in designing the materials for this experiment, we matched target words across conditions for length and frequency. To reduce the potential of a floor effect on reading time, we chose to use only relatively long words as targets. We also chose to control for frequency by using only relatively common words, due to empirical evidence that frequency (unigram predictability) may have an effect on processing cost independent to contextual predictability (Goodkind & Bicknell, 2021; Shain, 2024).

In choosing the kind of typographical errors to introduce in our materials, our goal is not to introduce nonwords that cannot be deciphered or lead to confusion with competing word(s). Rather, we are interested in typos that result in nonwords that are easy-to understand and unambiguous, since these may still have high surprisal, but plausibly not cause much increased effort under KL theory. For this reason, we chose to focus solely on typographical errors introduced by the transposition of two adjacent letters, in the middle of the word, since these type of errors are among the most easily understood/corrected (Andrews, 1996; Perea & Lupker, 2003; Johnson et al., 2007; Johnson, 2009; Lupker et al., 2008; Huang & Staub, 2021).

To validate the predictions of surprisal theory for our target words, we obtained contextual surprisal estimates from causal language models (LMs). Rather than choosing a single model of surprisal, we use surprisal estimates from a large number of pretrained models, to see how the pattern in our results may differ between less capable smaller models versus larger modern LMs that are more plausibly able to capture the actual probability of typos. We can then compare these surprisal values to the results of the human reading time experiment to determine whether processing effort tracks closer to the prediction of surprisal theory or KL.

3.3 Methods

We generated a corpus of short texts containing target words with meanings that were either very expected or very unexpected, with and without typos, in controlled contexts. On this corpus we conducted a self-paced reading time experiment to assess human processing cost. We also gathered surprisal estimates from language models on these same materials. We then fit separate statistical

models for these two different dependent variables, to compare the patterns of human processing cost versus surprisal.

3.3.1 Materials

Participants read 51 short texts across 4 experimental conditions (from a dataset of 204 unique stimuli). Each text contained a single target word whose processing cost was of interest, preceded by a prefix of 10 or more words, and followed by a suffix of 3 or more words to finish the sentence. The target word differed across four conditions, which we will refer to as `expected`, `unexpected`, `expected_typo`, and `unexpected_typo`. Each context was designed to make the `expected` word be very predictable in the target location given the prefix. The `unexpected` word was chosen to have a meaning that would be be highly unlikely in the context but not ungrammatical. The `expected_typo` and `unexpected_typo` targets were created by adding typographical errors to the the respective non-typo words, as described below. For all items, the post-target context was designed to be as natural as possible while working grammatically with both the expected and unexpected word.

Figure 3.1 shows an example item (a context with targets for each of the four conditions). All stimulus items are listed in appendix B.5. All target words (expected and unexpected) were chosen from among the top 5000 most frequent words in the Corpus of Contemporary American English (COCA; Davies, 2008), with a median length of 10 letters. Within each item, the `expected` word and the `unexpected` word were chosen to be matched as closely as possible for frequency (in COCA) and length (number of characters). The `expected_typo` and `unexpected_typo` targets were generated by transposing two adjacent characters in the corresponding non-typo word. All such transpositions were of medial characters; no transpositions involved the initial two characters, nor the final character.

3.3.2 Experiment design

Participants 118 participants were recruited on the Prolific platform. Participants took a median of 18½ minutes to complete the study, with a reward per approved participant of £2.85 (average reward rate: £9.24/hr). All participants were native speakers of English, most located in the United States.

Procedure The self-paced reading experiment was implemented using the PennController for IBEX (Zehr & Schwarz, 2018) and was hosted on PCIbex Farm.[3] Each trial started with a single asterisk displayed alone at the centre of the screen, which the participant could navigate past to start the self-paced reading when ready, by pressing the space bar. After this primer, the stimulus text would be initially presented as a sequence of underscores the length of each word in the text, and the participant could press the space bar to reveal each word one at a time in sequence, with

[3] Experiment code and demo are available at farm.pcibex.net/r/KOqOiK.

the previous word reverting to an underscore as soon as the participant advanced to the next. The time interval between presses was recorded as reading time (RT) in milliseconds. After completing a self-paced reading sentence, the screen would clear, and a comprehension question would be displayed, with four answer choices, one correct and three incorrect, presented in random order. Once the participant selected an answer, the next experimental trial would be presented, starting with the primer asterisk. Comprehension questions were identical across conditions, querying information from the pre-target part of the text, so that the answer did not depend on the target word. Accuracy on these questions was used as an attention check, and accuracy below 80% was a criterion for participant data exclusion. In our analysis we used the reading times on nontarget words to control for participant baseline reading speed, allowing us to focus on collecting RTs on target items from our four conditions of interest, without a global control condition.

We used a Latin square design for our study. Each participant was assigned to one of four groups, with the conditions in each item randomized once across groups, so that all participants saw each of the 51 items exactly one time, with a roughly even balance of conditions across items. The order of items was randomized per participant. Before starting on the experimental trials, each participant was given detailed instructions, which mentioned that the sentences they were to read may contain typographical errors, and four practice self-paced reading trials were presented following the same procedure described above, each followed by a comprehension question.

All self-paced reading stimuli and corresponding comprehension questions, for experimental and practice trials, are provided in appendix B.5 (tables B.1 and B.2).

Data exclusion The mean comprehension score accuracy was 90.4%, with a median of 92.2%. All data from any participant who scored scored less than 80% on the comprehension questions was excluded. This resulted in 14 participants being excluded, with 104 participants remaining. Reading time data from all words in the self-paced reading sentences (target words and context words) were used in analysis; reading time on comprehension questions was not recorded. We followed common practice (Jegerski, 2013; Marsden et al., 2018; Nicklin & Plonsky, 2020; Futrell et al., 2021; Harrington Stack et al., 2018; Burchill & Jaeger, 2024) in excluding as outliers any RTs faster than 100 ms (including one negative value, due to an apparent software error). We also excluded any RTs slower than 5000 ms.[4] This RT outlier exclusion step removed only about 0.2% of remaining RTs.

After exclusions, the data consisted of a total of 128,179 RTs, 5,290 of which were on target words, with a median of 24.5 RTs per target word.

[4]We chose *a priori* a more inclusive upper outlier bound than the 2000 ms threshold used in much previous literature (e.g., the same references cited above), since we are interested particularly in high surprisal items, and due to the strong skew and kurtosis of RT data meaning classifying high RT values as outliers is more likely to be unwarranted. However preliminary experiments re-running our analyses with the less inclusive upper bound did not result in any appreciable change in results.

3.3.3 Language model surprisal estimates

To compare against the patterns of human processing cost, we gathered estimates of surprisal of the target words in context, using a collection of pretrained causal language models (LMs). Given input tokens $w_{1:n-1}$, the logits of the final hidden state of a causal LM M give a direct estimate of surprisal of a token w_n as $-\log p_M(w_n \mid w_{1:n-1})$. For a multiple-token words, we used the sum of subword token surprisal values, as licensed by the chain rule.[5]

We obtained surprisal estimates from the following pre-trained Transformer-based LMs: GPT-2, GPT-3, GPT-Neo, GPT-NeoX, OPT, OLMo, Llama-2, Llama-3, Mistral and Mixtral (Radford et al., 2019; Brown et al., 2020; Black et al., 2021; Black et al., 2022; S. Zhang et al., 2022; Groeneveld et al., 2024; Touvron et al., 2023; Llama team, 2024; Jiang et al., 2023, 2024). Surprisal values from the proprietary GPT-3 model were obtained using OpenAI's API; all others were computed using the model implementations provided in the Huggingface Transformers library (Wolf et al., 2020). For further details see appendix B.1.

3.4 Results

Figure 3.3 displays the empirical means of human reading time response (left subplot) and surprisal (right subplot), with bootstrapped 99% confidence intervals. In the left subplot, the horizontal axis represents reading slowdown associated with the target word. Slowdown is calculated as log reading time relative to participant mean—so that a positive value indicates reading time that is slower than average for a particular participant, and a negative value indicates faster. For each target word, reading times were aggregated over a 3-word region of interest consisting of the target word and the two subsequent words (a common strategy in reading time studies; see e.g., Burchill & Jaeger, 2024; Huang et al., 2024).[6]

In the plot of mean reading times (fig. 3.3, left), we see a striking zig-zag pattern, as predicted by KL theory: For `expected` and `expected_typo` conditions, reading time on the target region is near to or slightly faster than average, while for `unexpected` and `unexpected_typo`, reading time is substantially slower than average. Looking at the plot of LM surprisal estimates (fig. 3.3, right), we see that the mean surprisal estimates pattern as we anticipated: The `expected` words are very predictable, according to every LM, with surprisal values about or below 1 nat for all language models (that is, these words are assigned probability at least roughly one in three), and the other three conditions are all vastly higher surprisal. In particular, the mean surprisal for `expected_typo` is high—above

[5] So, for example, if target word w_n = `" sweetheart"` is broken into as two tokens by the LM's tokenizer, surprisal is computed as $-\log p_M($`" sweetheart"` $\mid w_{1:n-1}) = -\log p_M($`" sweet"` $\mid w_{1:n-1}) - \log p_M($`"heart"` $\mid w_{1:n-1},$ `" sweet"`$)$.

[6] In post-hoc analyses, we explored other window-sizes, and found that a 3-word window indeed leads to the clearest distinction between expected vs unexpected conditions—further analysis is given in appendix B.2.1.

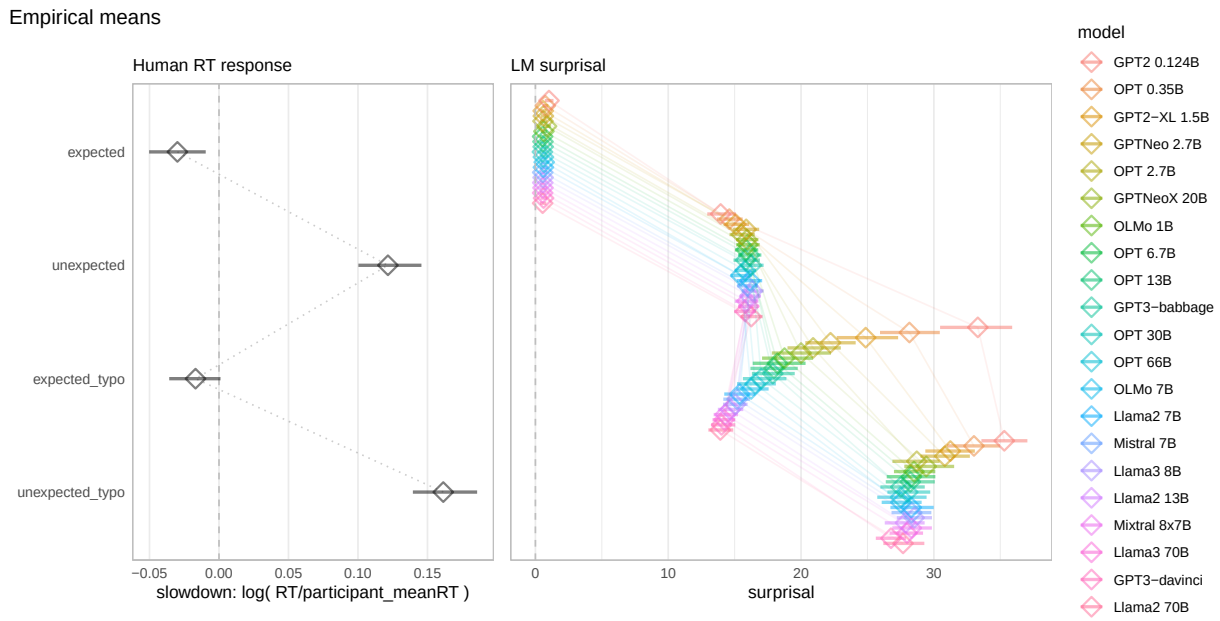


Figure 3.3: Empirical means of human reading time response and mean LM surprisal, across the four experimental conditions. Diamonds mark mean values, with horizontal lines indicating 99% CIs. **Left**: Reading time response represented on the horizontal axis as "slowdown", the log RT on the target region time relative to the participant's overall mean log RT. **Right**: Horizontal axis is surprisal; each LM is plotted in a separate color. LMs are ordered by their mean surprisal on the `expected_typo` condition, and LM names include number of parameters (where available).

14 nats for all language models (probability below 10^{-6}).[7] This plot is organized to highlight the general pattern across all language models. An alternative view of this data, plotted in groups of LMs to allow easier inspection of the patterns for each individual LM, is given in appendix B.2.2.

3.4.1 Regression analysis

To quantify more precisely the differences across our four experimental conditions, we fit mixed-effects linear regressions to predict human processing time and each LM's surprisal, separately. Specifically, we used Bayesian multilevel regression models, as used in other recent studies in this area (e.g., Cutter et al., 2022; Huang et al., 2024; For an overview of Bayesian regression models and data analysis techniques in psycholinguistics and cognitive science, see Nicenboim et al., 2024).[8] Fitting regression models allows us to examine the effect of condition on RT and surprisal, and to directly contrast conditions, while controlling for variation across participants

[7] Raw surprisal values are not comparable across different language models with different tokenization schemes and vocabulary sizes, but it remains clear that these values can reasonably be called 'high', in particular by comparison to the `expected` condition.

[8] We also fit frequentist regressions for RT and each LM's surprisal, using the same structure as the Bayesian regressions. The frequentist models and results are described in appendix B.4. Results were roughly equivalent to the Bayesian regressions—both in terms of overall interpretation as well as in terms of the specific coefficient estimates—indicating our interpretation of this data is not due to our choice of regression model.

and items. We fit one regression for human RT (formula 3.1), and a one for surprisal from each language model separately (formula 3.2), using the probabilistic programming language Stan (Carpenter et al., 2017), via frontend `brms` (Bürkner, 2017). Each regression included a fixed effect for the experimental condition (`target_type`), the predictor of interest, a four-level factor (`expected`, `unexpected`, `expected_typo`, `unexpected_typo`). In all regressions we also included the following control variables: target word number in sentence, and target length in characters (log transformed and centred).[9] For the regressions of surprisal, for which there was only one datapoint per item per condition, we included only by-item random intercepts. For the RT regression, we included by-item and by-subject random effects (both slopes and intercepts) to control for variation across individuals, and differences between experimental items, respectively. We did not include random effects for the other predictors that were not of theoretical interest (length and word number). RT values were aggregated over a three-word window starting at the target word.[10]

In fitting the regression models of RT, we used a log-shift transform on the dependent variable, subtracting off a global lower bound from all RT values, and then log-transforming (as advocated by Burchill & Jaeger, 2024, who found it produces more consistent fits in regression models of human reaction times compared to other common transformations). We set this global shift value to be the minimum RT value in the post-exclusion data (minus 1ms, to avoid undefined values an the minimum RT). This choice can be justified over the non-shifted log transform by the observation that reading times have a soft lower bound, with the relevant psychometric quantity being the amount by which reading time exceeds this lower bound. Refitting the regressions without this shift confirmed that this choice did meaningfully affect the interpretation of our results.

Priors For these regression models, we use the weakly informative priors (Lemoine, 2019; McElreath, 2020) given in table 3.1, consistent with the following intuitions. Under the sum-coding which we used for the four-level condition factor, the intercept is the grand mean; this is likely to be somewhere within a reasonable range of surprisal values (unlikely to be higher than 50 nats) or RTs (unlikely to be higher than 2000 ms). The difference between conditions is likely to be within a similar range. The standard deviations of the random slopes and intercepts is unlikely to be more than 30 nats or 400 ms, and the standard deviation of residuals is most likely smaller

[9] In exploratory data analysis, the relationship between word number and RT was plausibly linear; likewise for log-length, and likewise for the relationship of each as predictors of each LM's surprisal. We do not include random slopes for these predictors, in the interest of focusing the model complexity on the effects of theoretical interest. Preliminary experimentation with adding random slopes by participant did not affect the resulting interpretation.

[10] As noted above, we noticed in data exploration that faster readers seemed to show more spillover. A more involved regression model would allow the response metrics to differ between participants, either based on their reading speed or fit empirically (or even more computationally sophisticated regressions models that model the response continuously in time, as introduced by Shain & Schuler, 2018; Shain, 2021). We choose to use linear regression with this simple mean across 3-word window as the response, for simplicity and interpretability.

```
log(RT-shift) ~ target_type + (target_type | subj) + (target_type | item) + len + wordnum
```

Formula 3.1: Formula for the regression of RT (log-shift transformed), predicted as a function of condition (`target_type`), with by-participant and by-item random slopes and intercepts, and additional fixed effects for the length of the target word and its position in the sentence.

```
surprisal ~ target_type + (1 | item) + len + word_num
```

Formula 3.2: Formula for the regression of surprisal, predicted as a function of condition (`target_type`), with by-item random intercept, and additional fixed effects as above.

class	prior for LM surprisal	prior for log RT
Intercept	$\mathcal{N}(\mu = 5, \sigma^2 = 30)$	$\mathcal{N}(\mu = 6, \sigma^2 = 1)$
Coefficients	$\mathcal{N}(\mu = 0, \sigma^2 = 20)$	$\mathcal{N}(\mu = 0, \sigma^2 = 1)$
Standard deviation (random effects)	$\mathcal{N}(\mu = 0, \sigma^2 = 20)$	$\mathcal{N}(\mu = 0, \sigma^2 = 1.5)$
Standard deviation (residuals)	$\mathcal{N}(\mu = 0, \sigma^2 = 30)$	$\mathcal{N}(\mu = 0, \sigma^2 = 2)$

Table 3.1: Priors for regressions for surprisal (separately for each LM) and human reading time (with log-shift transform).

than about 50 nats or 800 ms. We allowed correlation between random slopes and intercepts, with a prior on correlation between slope and intercept being uniform between -1 and 1 (as is common practice in Bayesian multilevel models McElreath, 2020; Nalborczyk et al., 2019, §5.2.4; Nicenboim et al., 2024).

Note that the use of (weakly) informative priors may be most important in situations of data sparsity, where they can mitigate some potential problems such a reducing type I errors (Lemoine, 2019). Due to the relatively large amount of data, our choices of prior distributions likely did not have any meaningful effect on the results (confirmed by comparison with equivalent frequentist models which are reported in appendix B.4).

Regression fitting details In fitting the regression models, we used four independent chains of Markov chain Monte Carlo (MCMC), with 5,000 iterations each, with the first half of the samples in each chain discarded as warmup, for a total of 10,000 post-warmup draws. For our regressions all estimates had $\hat{R} < 1.01$, indicating successful convergence.[11]

3.4.1.1 Analyzing contrasts of interest between conditions

While the general trends in our results are relatively clear in the empirical plots above (fig. 3.3), analysis via the regression models allows us to better control for by subject and by item variability, estimate marginal mean effects (which we compute with the `emmeans` package Lenth, 2024) and

[11] For these Bayesian regression models the *potential scale reduction factor*, $\hat{R}$, compares between-chain to within-chain variance; a value below about 1.1 can be taken to indicate that the chains have converged to the posterior distribution (see e.g., Gelman et al., 2014; Bürkner, 2017; Nalborczyk et al., 2019).

examine post hoc contrasts to answer our specific research questions. An estimated marginal mean for a given effect represents the effect value, marginalizing over the values of other predictors (see Sonderegger, 2023, ch. 7; Lenth, 2024). In our case, for the regression model of RT, the marginal mean effect of a given contrast represents the slowdown effect for an average speaker and item, according to the regression model. Likewise, for each regression of LM surprisal, the estimated marginal mean is the effect for an average item.

Figure 3.4 displays three contrasts between conditions, one to address each of the three research questions outlined above (§3.2.1.1 A, B and C), according to the regressions of human RT (left), and LM surprisal (right). These marginal mean effect contrasts are analogous in interpretation to contrast-coded fixed effects: They measure how much the difference between conditions has on the response (human RT slowdown, or LM surprisal).[12] On the vertical axis is the contrast of interest, and on the horizontal axis is the estimated marginal mean effect for that contrast (in units of log ms for the RT regression, and nats for the surprisal regressions). On the horizontal axis is the regression models estimated effect.

These results confirm the main way in which human RTs and LM surprisal behave similarly, the 'unexpectedness effect' (middle row) is positive across the board: Words with unexpected meanings took longer to read, and had larger surprisal. Yet, looking at the 'typo effect' (top row), we see evidence of one important way in which LM surprisal estimates behaved very differently from human reading times: The typo effect on RT was very small or perhaps negligible, whereas the typo effect on surprisal is large, and is of a similar magnitude to the unexpectedness effect—or even larger than it, depending on the LM. In the third row we have the contrast where the typo effect and unexpectedness effect work in opposite directions (`unexpected` vs `expected_typo`). Here we see that for human RTs, this is similar in size to the unexpectedness effect, as predicted, whereas for surprisal, it is much smaller, near zero or even negative for all but a few of the largest and most recent LMs.

3.4.1.2 Consistency within items and participants

The pattern we observe in these marginal mean effects on contrasts in RT response are also mirrored within items and within participants. The consistency across participants can be seen by conditioning on each individual participant in the experiment, and estimating the marginal mean effect across items for that participant. Likewise, consistency across items can be inspected by conditioning on each individual item in our experiment in turn and computing the estimated marginal mean across participants. A figure showing all of these conditional estimated marginal effects (one for each item and one for each experiment) side by side in small subplots is given in

[12] Here we analyze the regression-estimated contrasts between conditions. For a plot of marginal mean effect estimates for each of the four conditions individually—corresponding to the empirical plot in fig. 3.3—see appendix B.3 (fig. B.3).

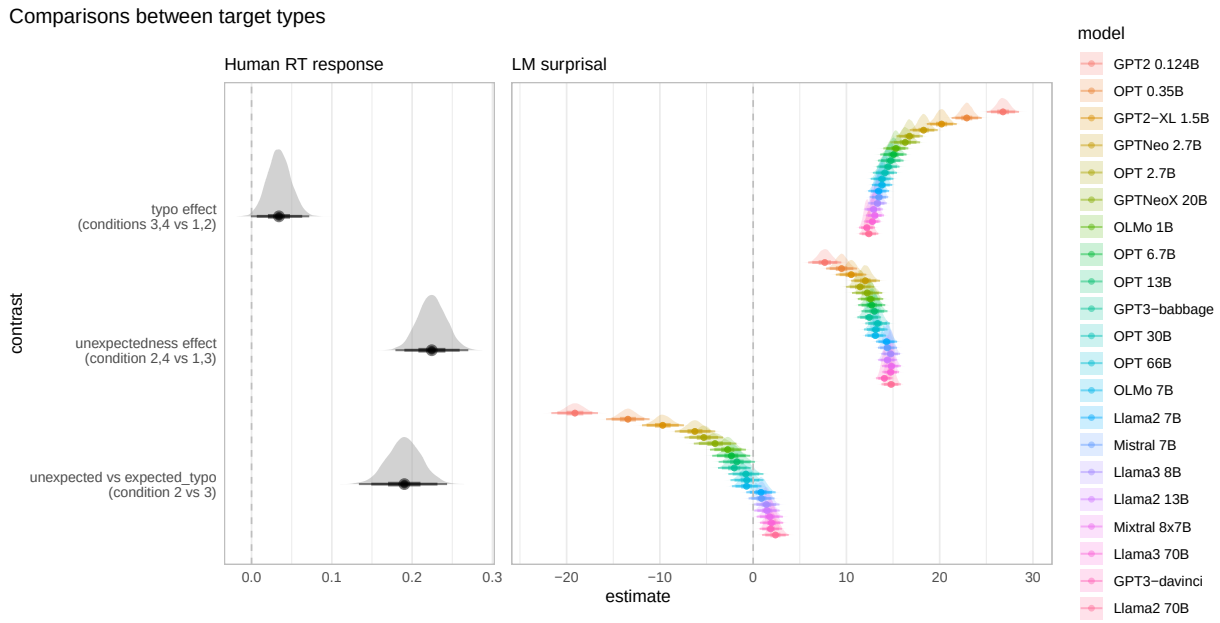


Figure 3.4: Contrasts between conditions' effect on RT (left subplot) and surprisal from each LM (right subplot). The three contrasts of interest, plotted on the vertical axis, top to bottom, correspond with our research questions A, B and C, respectively. The posterior over the regression model-predicted contrast of interest is represented with a shaded density plot. Dots mark the median estimates, with horizontal bars indicating credible intervals (at the levels of 0.66, 0.95, and 0.99). As in the previous plot, language models are ordered by their mean surprisal on the `expected_typo` condition.

appendix B.3.2, showing a general tendency for each item and each participant to mirror the overall pattern seen in fig. 3.4.

With a similar intent, instead of comparing each of these conditional effect plots individually, we can alternatively use the regression model's distribution over items (or participants), to get an estimate for such a conditional effect for a *typical* single item (or participant). These conditional plots are presented in fig. 3.5, for a typical item, on the left, and for a typical participant, on the right.

These results confirm that the general pattern we see in the overall marginal mean effects are present at the level of individual items and participants: The typo effect is generally small or nonexistent, and the unexpectedness effect is larger (and when these two effects work in opposite directions, the unexpectedness effect overwhelms the typo effect). While there is substantially lower uncertainty about this general pattern when controlling for by-subject and by-item variability together, it is notable that the pattern is also present at the within-group level.

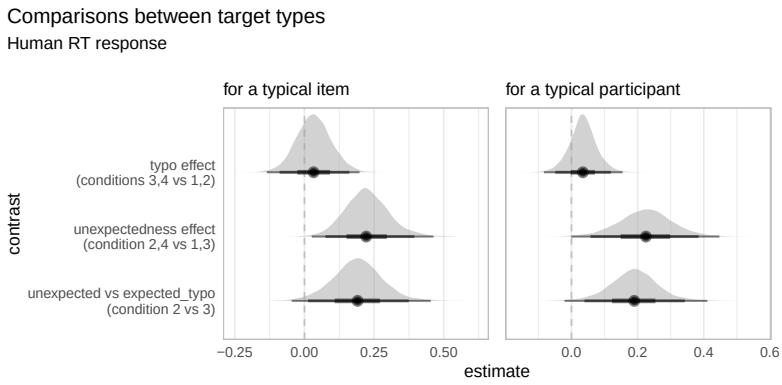


Figure 3.5: Average marginal contrasts conditioning on a typical item, marginalized over participants (left), or conditioning on a single typical participant and marginalizing over items (right).

3.5 Discussion

The results of this study demonstrate that typographical errors form an exemplary situation where surprisal alone is not an adequate predictor of human processing, systematically overestimating the cost, when compared to items of similar surprisal that are not typos. These results can be interpreted as confirmation that there are real-world situations where the equivalence between surprisal and KL does not hold. One of the central motivations for surprisal theory relies on its equivalence to KL, so this study provides evidence that the theory may need to be modified. Adopting the hypothesis that processing cost is driven by KL, which only sometimes is equivalent to surprisal, rather than by surprisal directly, provides a more general theory of language processing under uncertain input.

While it is not the main focus of our study, it is also notable that there is marked variation across the language models in the pattern of surprisal across our four conditions. In the empirical means of surprisal, displayed in fig. 3.3, this variation is most striking in the `expected_typo` condition (and to a lesser extent `unexpected_typo`), when compared with the same models' surprisal on the `unexpected` condition: Those models that assign the highest surprisal values to `expected_typo` words are the smaller/less recent LMs like the GPT2, and those that assign somewhat lower (but still high) surprisal are generally the larger/more recent models like the largest versions of GPT-3 and Llama. Recall that, in the plots, language models are sorted top to bottom by their decreasing mean surprisal on the target words in the `expected_typo` condition, but the resulting order ends up being generally from smaller/older models to larger/more recent. This trend demonstrates that as language models improve in their accuracy at capturing the statistical patterns in their training data, their surprisal values for typographical errors are becoming less astronomically high—but remain large, even for the best LMs. This is consistent with the intuition that a good model of the statistics of language data would be one which encodes accurately not just the probability of

well-formed words, but also the likelihood of corrupted versions due to mistakes such as typos (recall that surprisal can be seen as the negative log of the expected likelihood, under the prior distribution; eq. 3.2). Better language models behave as if they represent a more accurate likelihood function, in this way, providing better estimates of the true probability of these surface forms, increasing their probability to what may be presumed to be more accurate estimates. Yet it remains the case, even for the best LMs, that the surprisal of these typographical errors does not pattern at all similarly to human reading time. Instead, human reading time behaves as should be expected if processing cost is explained by the more direct measure of information gain provided by KL, rather than surprisal.

Limitations and further directions

In this chapter I have argued that surprisal cannot explain the ease with which humans can process words with typographical errors, and that if information gain is instead operationalized more directly—as KL—it can explain our results. However while we have validated the failure of surprisal to match human reading times in our experiment, using estimates of surprisal from LMs, we have not provided a direct estimator of KL, to validate our assertion that it predicts human processing cost better than surprisal. The implementation of estimators of this information theoretic quantity is material for future work. For instance, one way this could be achieved by using strings to representing the intended utterance, and defining a likelihood function (or a family of such functions) for mapping intended meanings into observations, in order to introduce typographical errors in proportion with empirical facts about error production, such as from actual typographical error statistics (for instance, using TypeRacer data as in R. Chen et al., 2021; or annotation of hand-corrected errors, as in Geertzen et al., 2014; Hahn et al., 2019). The study presented in this work did not investigate the extent to which participants' rapid processing on words containing errors involved their actually visually perceiving the errors, and correcting them, or simply not perceiving them at all. While in some sense this may not matter for the high-level point that the processing cost for such items was lower than predicted by their surprisal, it is potentially important to understand whether the observed effect would remain in a modified version of the experiment designed to control for whether the errors were in fact perceived.

Summary

From the perspective of language processing as incremental inference, the computational cost associated with a word can be measured as the amount by which it requires the comprehender to change their expectations about the meaning. Surprisal theory makes the claim that this cost can be quantified by the negative log probability. However, a main motivation for this hypothesis relies on the assumption that there is no nondeterminism in the mapping of intended meanings

to observable words. Relaxing this assumption gives rise to a novel prediction: when the actually observed word is malformed, but is interpreted as expressing an expected meaning, surprisal should substantially overestimate processing cost. In this situation, cost would be better quantified as the KL divergence between prior and posterior. Typographical errors form one example of a situation where this is plausible, and in this study, we investigated such examples, collecting self-paced reading times to measure human processing cost, and using large language models to estimate surprisal.

The main takeaway can be put succinctly as follows: In terms of surprisal, a typographical error cannot generally be distinguished from a word with an unexpected meaning, but these are intuitively very different things for humans. This intuition can be formalized with KL theory, with predictions that are borne out in the results of our reading-time study.

Note introducing chapter 4

As outlined in the Overview, the manuscript presented in the following chapter departs from the focus of the rest of the dissertation on incremental comprehension to present a study on another aspect of the relationship of language processing to the statistics of language use. It examines the connection between the linguistic structure by which a sentence's meaning is composed, as represented by dependency trees, and statistical dependencies between words in context, as quantified by pointwise mutual information.

As with the previous two manuscripts, the empirical study presented in the following work makes use of pretrained language models to estimate the probability of words in context. Note however, in this case we are interested in the amount by which a given word's probability of occurring is influenced by another word in the same sentence, conditioned on all the surrounding context, both preceding and following. Estimates of such conditional probabilities allow us to compute estimates of pointwise mutual information between two words, given the particular surrounding context. For this reason the language models used in this study are bidirectional models, trained on masked language modelling tasks (e.g., BERT; Devlin et al., 2019) rather than the unidirectional or causal models (e.g., GPT-2; Radford et al., 2019), which provide estimates of the probability of words given only the preceding context, as was appropriate for the studies of incremental processing presented in the previous chapters. Thus, in the following chapter, the term language model (LM) is used exclusively to refer to bidirectional models unless explicitly specified otherwise.

4

Linguistic dependencies and statistical dependence

Published as Hoover et al. (2021)

Are pairs of words that tend to occur together also likely to stand in a linguistic dependency? This empirical question is motivated by a long history of literature in cognitive science, psycholinguistics, and natural language processing. In this work we contribute an extensive analysis of the relationship between linguistic dependencies and statistical dependence between words. Improving on previous work, we introduce the use of large pretrained language models to compute contextualized estimates of the pointwise mutual information between words (CPMI). For multiple models and languages, we extract dependency trees which maximize CPMI, and compare to gold standard linguistic dependencies. Overall, we find that CPMI dependencies achieve an unlabelled undirected attachment score of at most ≈ 0.5. While far above chance, and consistently above a non-contextualized PMI baseline, this score is generally comparable to a simple baseline formed by connecting adjacent words. We analyze which kinds of linguistic dependencies are best captured in CPMI dependencies, and also find marked differences between the estimates of the large pretrained language models, illustrating how their different training schemes affect the type of dependencies they capture.

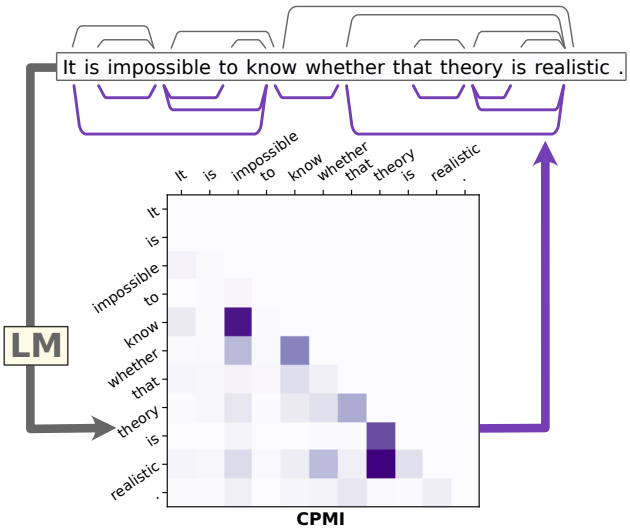


Figure 4.1: We use models pretrained on *masked language modelling* objectives to extract trees which maximize contextualized pointwise mutual information (CPMI) between words, to examine how linguistic dependencies relate to statistical dependence.

4.1 Introduction

A fundamental aspect of natural language structure is the set of *dependency relations* which hold between pairs of words in a sentence. Such dependencies indicate how the sentence is to be interpreted and mediate other aspects of its structure, such as agreement. Consider the sentence: *Several ravens flew out of their nests to confront the invading mongoose*. In this example, there is a dependency between the verb *flew* and its subject *ravens*, capturing the role this subject plays in the flying event, and how it controls number agreement. All modern linguistic theories recognize the centrality of such word-word relationships, despite considerable differences in detail in how they are treated (for a review of linguistic dependency grammar literature see de Marneffe & Nivre, 2019).

In addition to linguistic dependencies between words, there are also clear and robust statistical relationships. A noun like *ravens* is likely to occur with a verb like *flew*. In short, the presence or absence of certain words in certain positions in a sentence is informative about the presence or absence of certain other words in other positions. This raises the question: Do words that are strongly statistically dependent tend to be those related by linguistic dependency (and vice versa)? In everyday language, a sentence like the example above is probably more likely than *Several pigs flew out of their nests to confront the invading shrubbery*, despite this second example being syntactically identical to the first.

The long tradition of both supervised and unsupervised learning of grammars and parsers in computational linguistics suggests a strong link between dependency structure and statistical

dependence. Works such as Magerman and Marcus (1990) and de Paiva Alves (1996) introduced the use of pointwise mutual information (PMI) as a measure of the strength of statistical dependence between words, for the purpose of inferring linguistic structures from corpus statistics. The link between PMI and linguistic dependency has been studied and affirmed in Futrell et al. (2019). They show that for words linked by linguistic dependencies, the estimated mutual information between POS tags (and distributional clusters) is higher than that between non-dependent word pairs, matched for linear distance.

In this work, we dig further into the question of the correspondence between statistical and linguistic dependencies using modern pretrained language models (LMs) to compute estimates of conditional PMI between words given context, which we term *contextualized pointwise mutual information* (CPMI). For each sentence we extract a *CPMI dependency tree*, the spanning tree with maximum total CPMI, and compare these trees with gold standard linguistic dependency trees.[1]

We find that CPMI trees correspond better to gold standard trees than non context-dependent PMI trees. However our analysis shows that CPMI dependencies and linguistic dependencies correspond only roughly 50% of the time, even when we introduce a number of strong controls. Notably, we do not see better correspondence when we examine CPMI trees inferred by models that are explicitly trained to recover syntactic structure during training. Likewise, we see no increase in correspondence when we calculate CPMI over part-of-speech (POS) tags, a control designed to examine a less fine-grained statistical dependency than that between actual word forms. In fact, CPMI arcs broadly correspond to linguistic dependencies slightly less often than a simple baseline that just connects all and only adjacent words. We see similar overall unlabeled undirected attachment score (UUAS) when evaluated across a variety of pretrained models and different languages. However, a close analysis shows noteworthy differences between the different LMs, in particular revealing that BERT-based models are markedly more sensitive to adjacent words than XLNet. These differences yield insights about how different LM pretraining regimes result in differences in how the models allocate statistical dependencies between words in a sentence.

4.2 Background

Pointwise mutual information (PMI; Fano, 1961) is commonly used as a measure of the strength of statistical dependence between two words. Formally, PMI is a symmetric function of the probabilities of the outcomes x, y of two random variables X, Y, which quantifies the amount of

[1] We release our code at https://github.com/mcqll/cpmi-dependencies.

information about one outcome that is gained by learning the other.

$$\operatorname{pmi}(x; y) \coloneqq \log \frac{p(x, y)}{p(x)p(y)} = \log \frac{p(x \mid y)}{p(x)} \tag{4.1}$$

In our case, the observations are two words in a sentence (drawn from discrete random variables indexed by position in the sentence, ranging over the vocabulary). PMI has been used in computational linguistic studies as a measure of how words inform each other's probabilities since Church and Hanks (1989).[2]

Much earlier work on unsupervised dependency parsing (e.g., Van der Mude & Walker, 1978; Magerman & Marcus, 1990; Carroll & Charniak, 1992; Yuret, 1998; Paskin, 2001) used techniques involving maximizing estimates of total pointwise mutual information between heads and dependents, or maximizing the conditional probability of dependents given heads (these two objectives can be shown to be equivalent under certain assumptions; see §C.3). While such PMI-induced dependencies proved useful for certain tasks (such as identifying the correct modifier for a word among a selection of possible choices; de Paiva Alves, 1996), purely PMI-based dependency parsers did not perform well at the general task of recovering linguistic structures overall see discussion in Klein and Manning, 2004.

The recent advent of pretrained contextualized LMs (such as BERT, XLNet; Devlin et al., 2019; Z. Yang et al., 2019) provides an opportunity to revisit the relationship between PMI-induced dependencies and linguistic dependencies. These networks are pretrained on very large amounts of natural language text using masked language modelling objectives to be accurate estimators of conditional probabilities of words given context, and thus are natural tools for investigating the statistical relationships between words.

4.3 Contextualized PMI dependencies

Linguistic dependencies are highly sensitive to context. For example, consider the following two sentences: *I see that the crows retreated*, and *The mongoose pursued by crows retreated*. In the first there is a dependency between *retreated* and *crows*, and in the second there is not. However, PMI between two words in a sentence is strictly independent of the other words in that sentence.

Here we define *contextualized pointwise mutual information* (CPMI) as the conditional PMI given context, which we estimate using pretrained contextualized LMs. A contextualized LM M provides an estimate for the probability of words given context, which we use to define CPMI_M

[2] They used the term *word association*, which had a more subjective meaning in the psycholinguistic literature, to refer specifically to PMI.

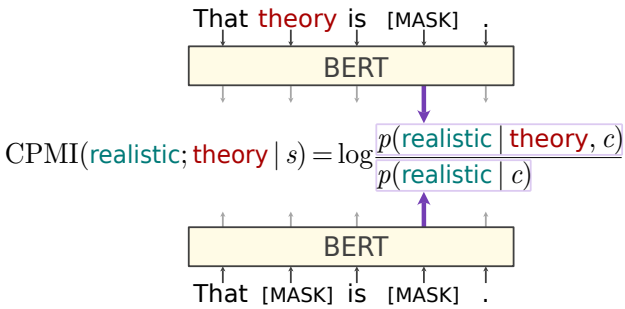


Figure 4.2: Diagram illustrating using BERT to compute the probability of *realistic* with and without masking *theory*, to obtain a CPMI score between those two words in the sentence $s =$ *That theory is realistic*.

between two words w_i and w_j in a sentence W as

$$\text{CPMI}_M(w_i; w_j) = \log \frac{p_M(w_i \mid W_{-i})}{p_M(w_i \mid W_{-i,j})} \tag{4.2}$$

where the W_{-i} is the sentence with word w_i masked, and $W_{-i,j}$ is the sentence with words w_i, w_j masked. To demonstrate the computation of this quantity, Figure 4.2 illustrates how BERT is used to obtain a CPMI score between the words *theory* and *realistic* in the sentence *That theory is realistic*.

4.3.1 Dependency tree induction

Given a sentence, we compute a matrix consisting of the CPMI between each pair of words. We then symmetrize this matrix by summing across the diagonal, so that we have a single score for each pair of words (omitting this step led to extremely similar results).[3] We then extract tree structures which maximize total CPMI. Since natural language dependencies are overwhelmingly projective (see Kuhlmann, 2010) we extract maximum projective spanning trees using the dynamic programming algorithm from J. M. Eisner (1996) and J. Eisner (1997).[4] Results for dependency trees alternatively extracted without the projectivity constraint, using Prim's maximum spanning tree (MST) algorithm (Prim, 1957), are similar, and results using both algorithms are provided in §C.4 for comparison. For further details on the extraction of CPMI dependencies, see §C.1.3.

4.4 Evaluating CPMI dependencies

In this section, we analyze the degree to which CPMI-inferred dependencies from pretrained LMs resemble linguistic dependencies.

[3]Note that while theoretically CPMI should be symmetric, nothing in the pretraining of the LMs we use enforces this identity (see §C.1.3.2 for details).

[4]Eisner's algorithm recovers the optimal projective *directed* dependency structure from a weighted ordered graph, but with a symmetric weight matrix, the output dependency trees may be interpreted as undirected.

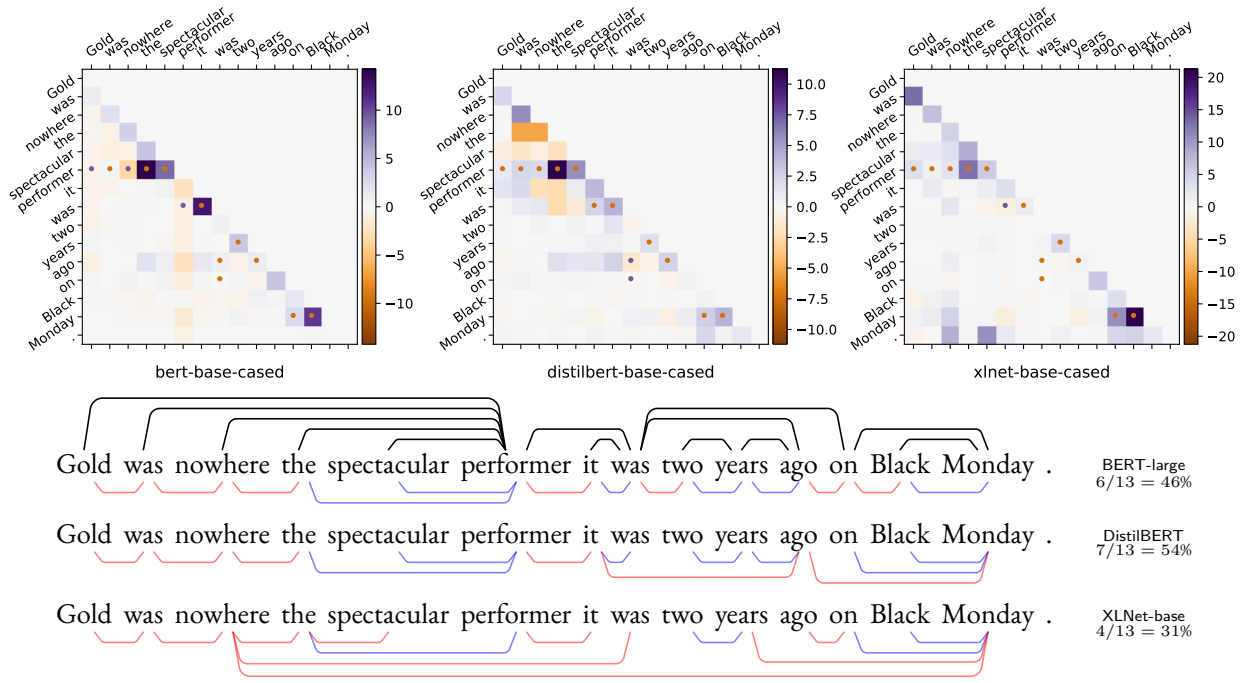


Figure 4.3: **Top:** CPMI matrices for an example sentence, from BERT, DistilBERT, XLNet. Gold dependencies are marked with a dot. **Bottom:** Resulting projective MST parses for the three models. Gold dependency parse above in black, CPMI dependencies below, blue where they agree, and red when they do not. The unlabeled undirected attachment score (UUAS) is given at right. Further examples provided in appendix, Figure C.7.

4.4.1 Method

We use gold dependencies for sentences from the Wall Street Journal (WSJ), from the Penn Treebank (PTB) corpus of English text hand-annotated for syntactic constituency parses (Marcus et al., 1994), converted into Stanford Dependencies (de Marneffe et al., 2006; de Marneffe & Manning, 2008).[5] We evaluate all extracted dependency trees on the full development split (WSJ section 22, consisting of 1700 sentences). For comparison with other work in unsupervised grammar induction, we also report results on the WSJ10 (all 389 sentences of length ≤ 10 from section 23, the test split, as used in, e.g., S. Yang et al., 2020) in §C.4.1.

To compare results across languages we use the Parallel Universal Dependencies treebanks subset of Universal Dependencies (Nivre et al., 2020, p. v2.7). These consist of 1000 sentences translated into 20 languages.

Pretrained contextualized LMs We compute CPMI scores using a number of transformer-based pretrained LMs for English (BERT, XLNet, XLM, BART, DistilBERT; Devlin et al., 2019; Z. Yang et al., 2019; Conneau & Lample, 2019; M. Lewis et al., 2020; Sanh et al., 2019). For

[5]We use Stanford CoreNLP v3.9.2 to convert.

other languages (and English) we use pretrained multilingual BERT base; see C.4.2 for details. All pretrained contextualized LMs we use are provided by Hugging Face transformers (Wolf et al., 2020).

Syntactically aware models We likewise compute CPMI estimates using models explicitly designed to have a linguistically-oriented inductive bias, by taking syntax into account in their training objectives and architecture. Following Du et al. (2020), we include two pretrained versions of an ordered-neuron LSTM (Shen et al., 2019)—a language model designed to have a hierarchical structural bias. The first (ONLSTM) is pretrained on raw text data, the second (ONLSTM-SYD) is pretrained on the same data but with an additional auxiliary objective to reconstruct PTB syntax trees. As a control, we also include a vanilla LSTM model. All three models are trained on the PTB training split. Example parses extracted from these models are given in the appendix (Figure C.9). We extract CPMI estimates from these models similarly to the above, but we condition only on preceding material, since these LSTM-based models operate left-to-right. See §C.1.2 for details.[6]

Noncontextualized PMI control We also compute a non-contextualized PMI estimate using a pretrained global word embedding model (Word2Vec; Mikolov, Sutskever, et al., 2013), to capture word-to-word statistical relationships present in global distributional information, not sensitive to the context of particular sentences. This control is calculated as the inner product of Word2Vec's target and context embeddings, $\text{pmi}_{\text{w2v}}(w_i; w_j) \coloneqq \mathbf{w}_i^\top \mathbf{c}_j$, since its training objective is optimized when this quantity equals the PMI plus a global constant (as explained in O. Levy & Goldberg, 2014; Allen & Hospedales, 2019). Details are given in §C.1.1.

Baselines A random baseline is obtained by extracting a parse for each sentence from a random matrix (so each pair of words is equally likely to be connected). We also include a 'connect-adjacent' baseline—degenerate trees formed by simply connecting the words in order—a simple, strong, and linguistically plausible baseline for English.

In addition to these baselines, we will compare unlabelled undirected accuracy score (UUAS) with that reported for the Dependency Model with Valence (DMV; Klein & Manning, 2004), a classic dependency parsing model. Note, importantly, the DMV is not fully unsupervised, as it relies on gold POS tags, but it is still a useful benchmark, with UUAS 54.4% on the entire WSJ corpus, and 63.7% on WSJ10 (as reported in Klein & Manning, 2004, Fig. 3).

	all	len = 1 prec. | rec.	len > 1 prec. | rec.
random	.22	.49 | .34	.08 | .10
connect-adjacent	**.49**	.49 | **1**	– | 0
Word2Vec	.39	**.61** | .59	.19 | .19
BERT base	.46	.57 | .72	.27 | .21
BERT large	.47	.55 | .81	.24 | .13
DistilBERT	.48	.57 | .72	**.32** | .24
Bart large	.38	.52 | .64	.16 | .13
XLM	.42	.60 | .64	.23 | .22
XLNet base	.45	.59 | .66	.29 | **.25**
XLNet large	.41	.59 | .61	.23 | .22
vanilla LSTM	.44	.54 | .70	.26 | .19
ONLSTM	.44	.55 | .71	.27 | .19
ONLSTM-SYD	.45	.55 | .71	.27 | .19

Table 4.1: Total UUAS for max-CPMI trees (projective). Overall scores in the first column (over all arcs in the corpus, precision = recall), followed by precision and recall for adjacent words in the second and third columns, and likewise for nonadjacent words in the final two columns. Compare with an overall UUAS of **.544** originally reported in Klein and Manning (2004) for the DMV on the WSJ corpus.

4.4.2 Results

Example CPMI dependencies and extracted projective trees are given in Figure 4.3, with gold dependencies for comparison. Table 4.1 gives the UUAS results.[7] Overall UUAS is given in the first column. The remaining columns give the UUAS for the subset of edges of length 1 and longer, in terms of precision and recall respectively.[8] Table 4.2 gives overall UUAS from multilingual BERT for a selection of languages from the PUD treebanks (for full results see Table C.5, Figure C.6).

The overall results show broadly that CPMI dependencies correspond to linguistic dependencies better than the noncontextual PMI-dependencies estimated from Word2Vec. However, across the models, and across languages, UUAS in general is in the range 40–50%. Degenerate trees formed by connecting words in linear order (the connect-adjacent baseline) achieve similar UUAS. Additionally, for the ONLSTM models, which have a hierarchical bias in their design, we

[6]Note that results of the (ON)LSTM models are not directly comparable to the transformer-based models, as these models are trained on much less data.

[7]The overall UUAS constitutes both precision and recall, since the number of gold edges and CPMI edges are the same: for a sentence of length n, the denominator is $n - 1$.

[8]For the connect-adjacent baseline, note: for length 1, the recall score is perfect, because all gold arcs of length 1 are predicted correctly by this trivial baseline; for the length > 1 subset, precision is undefined since there are no predicted edges of length > 1, and recall is 0.

language	rand.	connect-adj.	**BERT** base
Chinese	.23	.45	.40
Czech	.25	.48	.48
English	.22	.42	.43
French	.23	.45	.47
German	.22	.42	.46
Korean	.28	.58	.49
Polish	.27	.54	.52
Russian	.26	.51	.51
Spanish	.23	.45	.48
Turkish	.27	.55	.48

Table 4.2: Total UUAS for selected languages from the multilingual Parallel UD dataset, for CPMI dependencies extracted from from BERT (base multilingual cased). See full results in Table C.5.

see that accuracy of the CPMI-induced dependencies is the essentially the same with or without the auxiliary syntactic objective. Overall accuracy for both syntactically aware models is the same as for the vanilla LSTM. Further analysis of these results is in §4.6.

4.5 Delexicalized POS-CPMI dependencies

In this second experiment we estimate CPMI-dependencies over part-of-speech (POS) tags, rather than words. In the unsupervised dependency parsing literature there is an ample history of approaches making use of gold POS tags (see e.g., Bod, 2006; Cramer, 2007; Klein & Manning, 2004). Additionally, a traditional objection to the idea of deducing dependency structures directly from co-occurrence statistics, beyond data sparsity issues, is the possibility that "actual lexical items are too semantically charged to represent workable units of syntactic structure" (as phrased by Klein & Manning, 2004, p.3). That is, perhaps words' patterns of co-occurrence contain simply too much information about factors irrelevant to dependency parsing, so as to drown out the information that would be useful for recovering dependency structure. According to this line of thinking, we might expect linguistic dependency structure to be better related to the statistical dependencies between the *categories* of words, rather than lexical items themselves. Thus a version of CPMI calculated over POS tags would be predicted to achieve higher accuracy than the CPMI calculated over lexical item probabilities above.

A straightforward but unfeasible way to investigate this idea would be to obtain contextualized POS-embeddings by re-training all the LMs from scratch on large delexicalized corpora only consisting of POS tags. Instead, for efficiency, follow LM probing literature (Hewitt & Manning, 2019) and train a small POS probe on top of a pretrained LM, which estimates the probability of the POS tag at a given position in a sentence. After training this probe, we can extract a POS-based CPMI score between words. We define this POS-CPMI analogously to CPMI, but using

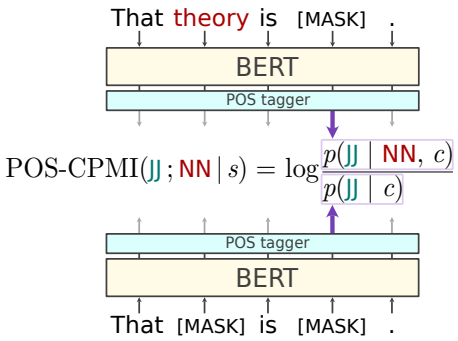


Figure 4.4: Diagram illustrating using BERT to compute the POS-CPMI score between the POS tags of the two words, *theory* (a noun, NN) and *realistic* (and adjective, JJ) in the sentence $s =$ *That theory is realistic*.

conditional probabilities of POS tags, rather than word tokens:

$$\text{POS-CPMI}_M(\pi_i; \pi_j) = \log \frac{p_{M_{\text{POS}}}(\pi_i \mid W_{-i})}{p_{M_{\text{POS}}}(\pi_i \mid W_{-i,j})} \tag{4.3}$$

where π_i, π_j are the gold POS tags of w_i, w_j in sentence W, and M_{POS} is the contextualized LM M with a pretrained POS embedding network on top. This is illustrated in Figure 4.4. We then extract POS-CPMI dependencies to compare to gold dependencies.

4.5.1 Method

We implement a POS probe as a linear transformation on top of the final hidden layer of a fixed pretrained LM. We train two versions of this probe: one trained simply to minimize cross entropy loss (simple POS probe), the other trained using the information bottleneck technique (following Tishby et al., 2000; Li & Eisner, 2019), to maximize accuracy while minimizing extra information included in the representation (IB POS probe). Using LMs BERT and XLNet (both base and large, each), we train each type of probe, to recover PTB gold POS tags. All eight probes achieve between 92% and 98% training accuracy.

We extract parses from POS-CPMI matrices just for CPMI (described above in §4.4). Below, we refer to the estimates extracted using the simple POS probe as simple-POS-CPMI, and those extracted using the IB POS probe as IB-POS-CPMI.

4.5.2 Results

Using the POS-CPMI dependencies does not result in higher accuracy. This provides evidence that the correlation between linguistic dependencies and CPMI dependencies is not merely artificially low due to distracting lexical information.

Table 4.3 shows the UUAS of the simple-POS-CPMI and IB-POS-CPMI trees. Compared

		all	len = 1	len > 1
			prec. | rec.	prec. | rec.
simple-POS	**BERT** base	.48	.56 | .79	.32 | .19
	BERT large	.45	.53 | .75	.27 | .16
	XLNet base	.36	.55 | .56	.17 | .17
	XLNet large	.32	.56 | .51	.14 | .15
IB-POS	**BERT** base	.41	.58 | .65	.20 | .18
	BERT large	.41	.55 | .69	.18 | .14
	XLNet base	.40	.55 | .60	.22 | .20
	XLNet large	.36	.56 | .56	.16 | .16

Table 4.3: Total UUAS for POS-CPMI using the simple POS probe and IB POS probe, from BERT and XLNet models. Overall results are in the first column, remaining columns break down results by arc length and recall and precision as in Table 4.1.

to the lexicalized CPMI trees discussed in the previous section, for BERT models, the simple-POS-CPMI dependencies have rather comparable overall UUAS, while for XLNet it is markedly lower. For both models, IB-POS-CPMI dependencies have lower UUAS. While these results are somewhat mixed, it is clear that, in our experimental setting, POS-CPMI dependencies correspond to gold dependencies no more than the CPMI dependencies do, performing at best roughly as well as the connect-adjacent baseline.

4.6 Analysis

In this section we outline main takeaways from a more detailed examination of the results from §§4.4–4.5, including additional analysis in §C.1.4.

UUAS is higher for length 1 arcs Breaking down the results by dependency length, Figure C.1 (in appendix) shows the recall accuracy of CPMI dependencies, grouped by length of gold arc. Length 1 arcs have the highest accuracy, and longer dependencies have lower accuracy. This trend holds for CPMI from all LMs. For BERT large, in particular, arcs of length 1 have recall accuracy of 80%, while longer arcs are near random. For XLNet, this trend is less pronounced.

No relation label has high UUAS In Figure 4.5, recall accuracy is plotted against gold dependency arc label.[9] When examining all lengths of dependency together (left) recall accuracy would seem to be correlated with mean arc length. But, filtering out all the gold arcs of length 1 (49% of arcs), we see that there is not a strong overall effect of arclength on mean accuracy for lengths > 1.

For most dependency labels, CPMI accuracy from each of the models is above the random baseline, but at or below to the connect-adjacent baseline. Exceptions to this trend include de-

[9]For descriptions of labels see the Stanford Dependencies manual (de Marneffe & Manning, 2008)

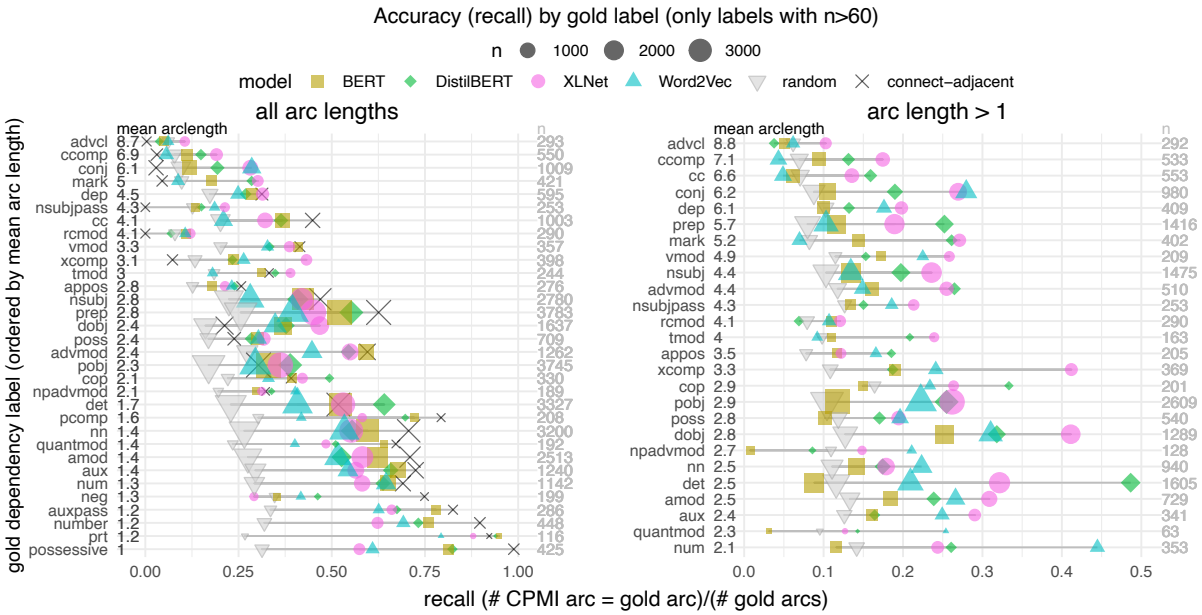


Figure 4.5: Plots of CPMI dependency recall accuracy versus gold edge relation (on the vertical axis, ordered by mean arc length). Only dependency relations of which there are more than 60 observations are included. **Left**: Including dependency arcs of all lengths. **Right**: Including only arcs between nonadjacent words. The connect-adjacent baseline predicts no such edges. Notice that the correlation with mean length disappears when excluding the length 1 arcs.

pendency labels dobj (direct object), xcomp (which connects a verb or adjective to the root of its clausal complement). For wordpairs in these relations, CPMI estimates (XLNet in particular) achieve higher accuracy than the baselines. However, even in these cases, CPMI dependencies do not perform at a level that could be considered successful for an unsupervised parser. This is contrary to what would be expected if CPMI-dependencies were in a strong correspondence with linguistic dependencies, even if this only held for certain types of linguistic dependency.

When considering arcs of length > 1, there is no dependency arc label which has UUAS above 0.5 from any of the models. More complete results including the other models not shown in Figure 4.5 are given in Table C.1 (in appendix).

UUAS is not correlated with LM performance Figure 4.6 shows per-sentence UUAS plotted against log pseudo-perplexity (PPL) for BERT and XLNet models (results are similar for other models; see §C.1.4.3, Figure C.2). These results show that correspondence between CPMI-dependencies and linguistic dependencies isn't higher on sentences on which the models are more confident.

We also examined the accuracy of CPMI dependencies during training of BERT (base uncased) from scratch. Figure C.4 (in appendix) shows the average perplexity of this model at checkpoints during training, along with average UUAS of induced CPMI structures. UUAS reaches its highest

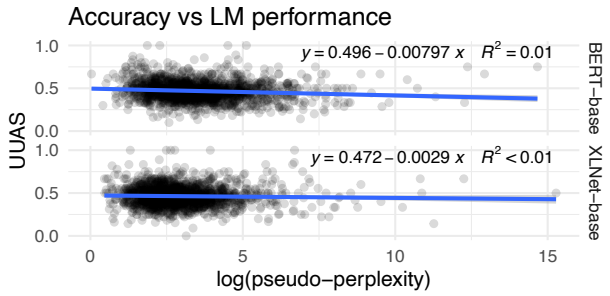


Figure 4.6: Per-sentence accuracy (UUAS) against log psuedo-perplexity. Each dot represents one sentence. Fitting a linear regression, the coefficient of determination R^2 is very close to 0 for all models (here BERT and XLNet are shown; other models are in Figure C.2)

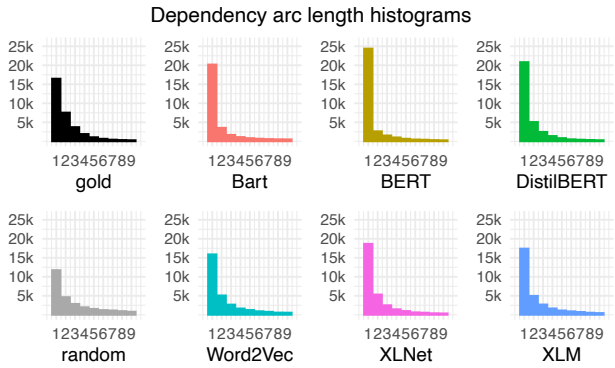


Figure 4.7: Histograms of arc length. Note, 49% of the gold arcs are length 1, whereas all of the CPMI dependencies had a higher proportion. BERT (base), in particular has 72%. For Word2Vec (which does not have access to word order), 47% are length 1. For the connect-adjacent baseline (not shown) the histogram is trivial: all arcs are length 1.

value before perplexity plateaus.

We should also stress that, throughout this paper, UUAS is not a measure of LM quality. Rather, it simply measures how well patterns of statistical dependence captured by the LM align with linguistic dependencies. Better alignment may not be related to better language modelling.

Dependencies differ between LMs Dependency structures extracted from the different pre-trained LMs show roughly similar overall UUAS, though the models agree with each other on only 25–48% of edges. They agree with the noncontextualized word embedding model Word2Vec at just slightly lower rates (21–27%), while agreeing with the linear baseline at higher rates (34–57%). See §C.1.4.1 and for these details.

In particular, CPMI dependencies from all the models connect adjacent words more often than the gold dependencies do, but this effect is much more pronounced for BERT models than for XLM, and XLNet models (Figure 4.7). A possible reason for this difference lies in the way these models are trained. XLNet is trained to predict words according to randomly sampled chain

rule decompositions, enforcing a bias to be able to predict words in any order, including longer dependencies. XLNet's probability estimates for words may therefore be sensitive to a larger set of words, rather than mostly the adjacent ones. Whereas BERT, trained with a less constrained masked LM objective, has probability estimates that are evidently more sensitive to adjacent words.

4.7 Related work

Probing pretrained embeddings In the past few years, a substantial amount of literature has emerged on probing pretrained language models (in the sense of e.g. Conneau et al., 2018; Manning et al., 2020), wherein a presumably weak network (a *probe*) is trained to extract linguistic information (in particular, dependency information, in e.g. Hewitt & Manning, 2019; K. Clark et al., 2019) from pretrained embeddings. Extracting CPMI-dependencies differs from training a dependency probe in that it is entirely unsupervised, and is motivated by a specific hypothesis—about the relationship linguistic dependencies have with statistical dependence.

Nonparametric probing A number of other recent works have taken an unsupervised approach to investigating syntactic structure encoded by pretrained LMs, largely focusing on self-attention weights (e.g. Mareček & Rosa, 2018, 2019; Kim, Choi, et al., 2020; Kim, Li, & Lee, 2020; Htut et al., 2019). Very recently, T. Zhang and Hashimoto (2021, concurrent with this paper) examined conditional dependencies implied by masked language modelling using a nonparametric method similar to our CPMI, using BERT to estimate Conditional PMI (and Conditional MI) between words. They extract maximum spanning trees, and report UUAS on WSJ dependency data. Their results are similar to those reported here: namely, scores are much higher than a chance baseline, but close to a connect-adjacent baseline. While their numerical results are similar, their interpretation differs somewhat. Given our analysis, we find less reason for optimism about the prospects of unsupervised dependency parsing directly from probability estimates by pretrained LMs.

Perturbation impact The experiments in the current paper extracting CPMI can be seen as an application of the token perturbation approach of Z. Wu et al. (2020). They describe general nonparametric method to examine the *impact*, $f(w_i, w_j)$, of a word w_j on another word w_i in the sentence, where f is some difference function between the embedding of w_i (masked in the input) with and without the word w_j also being masked. In their experiments, they use two examples of impact-measuring functions (see Z. Wu et al., 2020, §2.2). The first, the *Dist* metric, is simply Euclidean distance between embeddings. The second, the *Prob* metric, is defined as $f(w_i, w_j) = p(w_i \mid W_{-i}) - p(w_i \mid W_{-i,j})$, using the masked LM's probability estimates (notation as defined in §4.3). The latter impact metric is quite similar to CPMI, the difference being only that *Prob* impact is the difference in probabilities, while CPMI is the difference in log probabilities.

Table 4.4 compares the reported UUAS of maximum projective spanning trees from CPMI

matrices, to those from *Dist* impact matrices on the English PUD data set. They do not report UUAS for the *Prob* metric or release code for it, but mention that it is significantly outperformed by the *Dist* method. Z. Wu et al. (2020, p. 1) note that their "best performing method does not go much beyond the strong right-chain baseline." While it may be seen as an application of perturbed masking technique, CPMI is motivated as a method to test a specific hypothesis about the relationship between linguistic and statistical dependence. Extracting matrices using another impact metric (such as Euclidean distance between embeddings, *Dist*) may indeed achieve higher attachment scores, as Z. Wu et al. (2020) demonstrate, but this does not bear on the hypothesis we focus on in this paper.

4.8 Discussion

In this paper we explored the connection between linguistic dependency and statistical dependence. We contribute a method to use modern pretrained language models to compute CPMI, a context-dependent estimate of PMI, and infer maximum CPMI dependency trees over sentences.

We find that these trees correlate with linguistic dependencies better than trees extracted from a noncontextual PMI estimate trained on similar data. However, we do not see evidence of a systematic correspondence between dependency arc label and the accuracy of CPMI arcs, nor do we see evidence that the correspondence increases when using models explicitly designed to encode linguistically-motivated inductive biases, nor when estimated between POS embeddings instead of word forms. Overall, CPMI-inferred dependencies correspond to gold dependencies no more than a simple baseline connecting adjacent words. This is our first main takeaway: statistical dependence (as modelled by these pretrained LMs) is not a good predictor of linguistic dependencies. Second, our analysis shows that CPMI trees extracted from different LMs differ to an extent that is perhaps surprising, given the similarity in spirit of their training regimes. The difference in accuracy when broken down with respect to linear distance between words offers information about the ways in which these models' inductive and structural biases inform the way they perform the task of prediction. BERT aligns better overall, but this is driven by its being more like the linear baseline. For longer arcs, XLNet aligns a bit better with linguistic structure. Compared to BERT, XLNet can be seen as imposing a constraint on the language modelling objective by forcing the model to have accurate predictions under different permutation masks.

Generalizing this observation, we ask whether linguistic dependencies would correspond to the patterns of statistical dependence in a model trained with a language modelling loss while concurrently minimizing the amount of contextual information used to perform predictions. Finding ways of expressing such constraints on the amount of information used during prediction, and verifying the ways in which this can affect our results and LM pretraining in general constitutes material for future work.

connect-adj. baseline	.42
CPMI (proj.) BERT base multilingual cased	.43
right-chain baseline	.40*
***Dist* impact** (proj.) BERT base uncased	.52*

*As reported in Z. Wu et al. (2020, Table 2)

Table 4.4: UUAS on English PUD, for CPMI (from Table 4.2), compared to Z. Wu et al. (2020)'s results. Note: the baselines aboves are theoretically identical, discrepancy may be due to data processing differences.

Acknowledgements

We thank anonymous reviewers, in particular the reviewer who alerted us to the work of Z. Wu et al. (2020), and also Richard Futrell for helpful discussions and feedback. We also gratefully acknowledge support from the Centre for Research on Brain, Language & Music, the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche du Québec, Nature et Technologies and Société et Culture, and the Canada CIFAR AI Chairs Program.

5

Discussion and conclusion

This dissertation has been broadly aimed at understanding ways that theories of language structure and processing relate to the statistical patterns of language use. The first three chapters explored a computational theory of the effort involved in sentence processing, viewing it as an incremental inference task. Starting from a theoretical justification at the core of existing work in expectation-based models of processing cost, in the first chapter I presented a hypothesis that I refer to as divergence theory, which can be seen as a generalization of the influential hypothesis known as surprisal theory, derived by relaxing two standardly-held assumptions. Then, within this framing, the following chapters presented studies which examined the potential benefits of relaxing these assumptions. In the last chapter, I presented a study of the way statistical dependence between words can be compared to the word-to-word dependencies that describe grammatical structure. Throughout this work, I have framed questions about language processing and structure in terms of the information-theoretic quantities of surprisal, pointwise mutual information, and KL divergence, and used pretrained large language models, as the best-available statistical models of word-probability in context, in order to estimate these information theoretic quantities.

5.1 Summary and general discussion

Incremental processing difficulty as an information cost

Chapter 1 developed divergence theory—the hypothesis that processing cost can be measured in terms of information gain, quantified with divergence between probability distributions. In this application, the relevant distributions in whose divergence we are interested are those representing the degree of belief about the intended meaning of an utterance within a Bayesian inference setting.

We explore Kullback-Leibler (KL) divergence (also known as relative entropy) as quantification of the amount of information that is gained in moving from one such belief distribution to another, given an observation. Divergence theory hypothesizes that processing cost increases as a function of this quantity. This hypothesis derives from the intuition that processing cost represents the effort involved in computing a representation of the posterior distribution over the meaning upon making an observation, and is supported by results from the literature on the computational complexity of sampling algorithms for approximate inference, which have been shown to scale in such a quantity.

In its most general form, divergence theory hypothesizes that processing cost for an observed word $\breve{w}$ be measured by the divergence between the true posterior distribution, $p_{Z|\breve{w}}$, and some proposal distribution $q_{Z;\breve{w}}$ that may be used to approximate this posterior. Choosing KL divergence in particular as the quantification of belief update size, divergence theory becomes the hypothesis that $\text{cost}(\breve{w}) = f\left(D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}})\right)$ (hypothesis 1.5). This divergence can be decomposed as in the following equation (repeating eq. 1.8).

$$D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}}) = \overbrace{\log \frac{1}{p(\breve{w})}}^{s(\breve{w})} - \left(\overbrace{\mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{1}{p(\breve{w} \mid z)}\right]}^{R(\breve{w})} + \overbrace{\mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{q(z;\breve{w})}{p(z)}\right]}^{D_{q_{Z;\breve{w}}}} \right) \tag{5.1}$$

The term $R(\breve{w})$—which I refer to as the *reconstruction information*—quantifies the amount of uncertainty (expected surprisal) remaining in the observation under the posterior distribution over Z. This 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. The final term, $D_{q_{Z;\breve{w}}}$—which I refer to as the *proposal advantage*—quantifies how much closer $q_{Z;\breve{w}}$ is to the posterior than the prior is. If we assume that the distribution we use for $q_{Z;\breve{w}}$ is simply equal to the prior, then this final term vanishes, and we have the special case of KL theory from the prior—the hypothesis that $\text{cost}(\breve{w}) = f\left(D_{\text{KL}}(p_{Z|\breve{w}} \,\|\, p_Z)\right) = f\left(s(\breve{w}) - R(\breve{w})\right)$ (hypothesis 1.4), that is, that the processing cost for an observation is directly measured as a function of the size of the Bayesian belief update it incurs.

The hypothesis that cost scales as a function of KL divergence reduces completely to standard linear surprisal theory if two assumptions are made: first, that $R(\breve{w}) + D_{q_{Z;\breve{w}}} = 0$ for all observations and all contexts (implying that KL is always equal to surprisal), and second, that the linking function f is linear. This framing naturally leads to the question of whether these assumptions are justified. The empirical studies in chapters 2 and 3 were aimed at questioning these assumptions, with each study finding evidence in favor of generalizing surprisal theory, by relaxing the respective assumption.

Arguments and evidence for a superlinear linking function

Chapter 2 took up the question of the linking function between surprisal and processing cost. In order to focus on this question within the framework of KL theory, I followed all previous literature in this area in explicitly assuming that KL was equivalent to surprisal (deferring the question about whether this assumption is always merited).

This work was aimed at the question of finding a class of algorithms that could explain general surprisal theory, looking at the complexity of algorithms that sample from a prior distribution in order to approximate the posterior. This investigation into sampling algorithms was motivated by the need for an algorithmic theory to implement the relationship that the computational-level surprisal theory (or more generally divergence theory) predict: It is well-documented that less-expected words take longer to process, but no known parsing algorithm has computational complexity that scales in surprisal. If a processing algorithm is conceived of as a mechanism for building a representation of posterior given an observation, the natural way in which computational cost might be related to surprisal 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 meanings with high prior probability are those which sample from the prior. Conducting an analysis of some simple fundamental examples of such algorithms revealed that they predict runtime to increase in surprisal superlinearly. This is also the case for more sophisticated algorithms based on importance sampling, where the number of samples required scales exponentially in KL, and therefore in surprisal, under the standard assumption of their equivalence.

In the second part of chapter 2, we conducted an empirical study, using nonlinear regression models to predict human reading times from surprisal estimates from a number of contemporary large language models. For the most accurate language models we found evidence that reading time increases superlinearly in surprisal, consistent with the predictions of sampling-based algorithms.

Evidence that words may be surprising but not difficult to process

Chapter 3 focused on questioning the assumption that the relationship between latent structure and observed words is deterministic. This assumption implies the equivalence of surprisal and KL divergence. We argued that one type of situation where this assumption is likely not to be warranted is when the precise form of the observed word is the result of some kind of production error, such as a typographical error or misprint in written text. In such cases, the degree to which this observation is unpredictable, as quantified by surprisal, may substantially exceed the degree to which it is informative, as quantified the KL divergence between Bayesian belief distributions.

We conducted a human reading time experiment on text containing minor typographical errors, as a case study to compare the predictions of surprisal versus KL. We found that human processing

effort on these items did not behave as predicted by surprisal theory, but was consistent with the qualitative predictions of divergence theory. In particular, 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. These results confirmed that there are situations in which estimates of surprisal alone are not adequate as a predictor of human processing effort, and suggest that an estimate of KL may be better suited to this task.

Comparing statistical and linguistic dependencies between words

Putting aside the question of incremental processing cost, another area where the distributional patterns of language use may relate to the way in which language processed is in the types of relationships that words have to one another. Chapter 4 presented a standalone study of the connection between two ways words may depend on one another: grammatically and statistically. This work was motivated by the question of whether words that stand in linguistic dependency relationship with each other tend to also be dependent on each other in terms of their co-occurrence frequency. We extracted tree structures which maximize contextualized pointwise mutual independence (CPMI) between words, using a pretrained bidirectional language models to estimate the probability of words in context, as well as a non-contextualized word-embedding baseline, and compared these resulting tree structures against linguistic dependency structures.

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 contextualized language models, compared than the non-contextualized baseline. 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 dependency trees by maximizing CPMI is at best only roughly 50% accurate, and 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 de-lexicalizing the CPMI metric. 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.

5.2 Future directions

5.2.1 Further developing divergence theory

As outlined in chapter 1, standard surprisal theory can be derived within the general framing of processing cost as quantified by information gain by making a number of independent assumptions. The studies of processing effort presented in chapters 2 and 3 were framed at questioning two of these assumptions of standard surprisal theory, while maintaining its core motivation. The empirical study in chapter 2 investigated the relationship between surprisal and reading time, motivated by the runtime of sampling-based algorithms, which imply a superlinear relationship between KL divergence and processing cost. This exploration of superlinearity in surprisal theory focused on questioning the assumption of a linear linking function, while maintaining the other standard assumption—that surprisal is equal to KL divergence. Then, chapter 3 focused on examining the ramifications of relaxing this assumption of equivalence, looking in particular at typographical errors as a source of exemplary situations where a word may being unpredictable but not incur a large belief-update cost. The empirical study of constructed typographical errors found a substantial difference between human processing cost and the predictions of language models surprisal estimates. Human reading time patterned in a way which could not be explained by standard surprisal theory, but would be expected under divergence theory. Taken as a pair, a potential limitation of these two studies is that each of the two assumptions are relaxed independently. The first examines potential nonlinearity in surprisal theory, while the second examines generalizing surprisal theory to divergence theory, without having anything to say about the linking function per se. Future work would take aim at determining the form of the linking function directly in the context of divergence theory, relaxing these two assumptions simultaneously.

Additionally, as described in chapter 1, the more general version of divergence theory presented as hypothesis 1.5 proposes that cost be measured with the divergence of the posterior from a 'proposal' distribution, rather than from the naïve prior (see table 1.1). From this perspective, the studies and explorations presented in this dissertation have all assumed that this proposal distribution was equal to the prior. Investigating the implications of relaxing this final assumption, allowing the use of a proposal distribution, forms an important direction for future work, with the exploration of plausible families of proposal distributions being an important component in the design of inference algorithms for sentence processing, given the intuition that one can often do better than the naïve prior in practice.

Implementing estimators of KL divergence

Another crucial direction for future research should be the design and implementation of estimators of KL divergence which take into account the potentially nondeterministic ways in which

the linguistic structures of interest to a comprehender are related to the linguistic inputs they may observe. The derivation of divergence theory and the empirical results presented in this dissertation provide evidence of and explanation for how the patterns of surprisal may be expected to systematically differ from those of human processing cost. However, as mentioned above, while estimates of the surprisal of surface forms may be computed using the variety of increasingly accurate modern language models, to fully assess the theory as an alternative to surprisal theory requires implementing an estimator of KL divergence between distributions over the latent structures of interest. To achieve this will require constructing a model of the likelihood function, which scores latent representations for a given observation. Current large language models may provide useful models of the prior distribution, to be used alongside such a likelihood function to compute estimates of KL.

Applying divergence theory to explain other types of effects

As an example of a type of situation in which KL can be expected to differ substantially from surprisal, the study in chapter 3 focused on typographical errors. I chose to focus on typos for two reasons: They are relatively understudied despite their ubiquity, and they provide a relatively simple and convenient setting in which nondeterminism in the relationship between intended meanings and observable forms is immediately relevant. However, in future research there is no reason to limit the focus to orthographic effects. Any number of constructions where processing is known to proceed with relatively little impediment despite surprisal being high provides an inherent conundrum for standard surprisal theory, that may potentially be amenable to explanation in terms of divergence theory. As described in §1.3.3, the various types of constructions known broadly as *grammatical illusions* provide potential examples of such situations where the mismatch between literal observation and inferred meaning occurs at a more abstract level than orthography. This term (or similar ones, such as linguistic illusion, semantic illusion, or illusion of acceptability) is used to refer to a variety of constructions that have been documented to be generally judged acceptable when they are encountered, despite being structurally malformed or having compositional meaning that is impossible or absurd (as studied in, e.g., Wason & Reich, 1979; Barton & Sanford, 1993; O'Connor, 2015; Kelley, 2018; Wellwood et al., 2018; Paape et al., 2020; Muller & Phillips, 2020; Y. Zhang, Ryskin, & Gibson, 2023; Y. Zhang et al., 2024).

As one concrete example, 'depth-charge illusions' are a specific kind of grammatical illusion that may provide a line of research worth exploring in future work from the perspective of divergence theory. The box below briefly outlines a preliminary motivation for this application.

Depth-charge illusions

As mentioned in §1.3.3, the following is the canonical example of a depth-charge sentence.

No head injury is too trivial to ignore. (Wason & Reich, 1979)

Upon hearing or reading this sentence, most people agree that it is acceptable and perfectly reasonable, however upon careful examination of the literal meaning of the sentence it becomes eventually apparent that it is in fact pragmatically absurd, though this often takes prompting and considerable reflection to notice (see review in, e.g., Paape et al., 2020). [To understand the literal meaning, it is helpful compare with the non-illusory sentence *No head injury is too trivial to attend to* (noting *ignore* is essentially the negation of *attend to*).] The illusion gets its evocative name from a comparison of the time it takes to realize the error to the time before explosion of a delayed-release bomb.

Such sentences provide an exemplary situation where a very implausible word does not induce processing difficulty. During processing of this sentence, at the point after hearing the prefix $\check{c} =$ *No head injury is too trivial to...*, we may plausibly suppose a typical comprehender would be relatively confident that the completion of the sentence would contribute to the overall meaning that "even trivial head injuries should be paid attention to." Then, upon observing $\breve{w} =$ *ignore*, there is apparently little change to the belief about the meaning being conveyed, leading to the well-documented illusion effect. In a sense, a crucial bit[a] of the meaning of this last word seems to be ignored or misinterpreted. The small KL between prior and posterior should, then, provide an explanation of the relative ease of processing such sentences, where surprisal could not. Indeed surprisal from modern LMs does not predict human judgments well on such semantically illusory sentences (as explored in recent work by Y. Zhang, Gibson, & Davis, 2023). In a setting where this type of sentence is uttered as an error, we can think of it as the semantic analogue of a minor typographical error. In both cases the observed utterance is malformed, with little effect on interpretation, and in both cases a theory of processing cost measured with KL divergence may be better suited to explain human behaviour than standard surprisal theory is.

[a] Here I intend the colloquial meaning of 'bit,' not the unit of binary information. But, depending on the structure by which meanings are represented, transforming 'not attend' $\mapsto$ 'attend' may well be equivalent to the flipping of one binary bit in the meaning representation.

5.2.2 Applications of CPMI in incremental processing

In addition to the directions for further study with respect incremental processing cost outlined above, the approach to quantification of word-to-word statistical dependence explored in chapter 4

also presents opportunities for application and further research. In particular, a better understanding of the statistical relations between words may be helpful in the modelling of online sentence processing, and vice versa, tying together these two aspects of the work in this dissertation. For instance, Futrell and Levy (2017) and Futrell et al. (2020) have explored information-theoretic explanations for so-called locality effects (see; Gibson, 2000; R. L. Lewis & Vasishth, 2005), theorizing a pressure for words of high pointwise mutual information to be close to each other in the string, under a model of processing where there is uncertainty about past material (due to memory or attention constraints). From this perspective, CPMI estimates may be expected to have direct implications for models of human reading behaviour, large absolute-value CPMI being potentially predictive of human eye movement during reading. If, upon reading a word, the uncertainty about of a previous word is drastically changed, this may be expected to induce a regressive eye movement. Very recently, this line of inquiry has been validated against multilingual eye-tracking data: Following methods similar to ours to compute CPMI from masked language models, Wilcox et al. (2024) found that positive CPMI was indeed predictive of regressive saccades, across multiple languages.

5.3 Conclusion

This dissertation has studied aspects and limitations of the ways that the statistics of word occurrence in context can inform our understanding of the way that language is processed. The main focus of this work has been to motivate and explore the idea that processing cost may be explained by modelling comprehension as incremental inference. In particular this has led to proposed modifications to and generalizations of the traditional view that processing cost scales proportional to surprisal. In exploring the ramifications of this generalization, which I have called divergence theory, I presented studies which provide evidence of a nonlinear relationship between surprisal and processing cost, and suggest that cost may be better quantified using the KL divergence between belief distributions, a more direct measure of the information gained about meaning. Together, these results motivate further exploration of incremental inference algorithms and inference-based models of human sentence processing.

Glossary

Definitions and notation for information theoretic and probabilistic quantities

surprisal of outcome x of a discrete random variable X:

$$s(x) \coloneqq \log \frac{1}{p(x)} = -\log p(x)$$

When referring to the surprisal of an observed word $\breve{w}$ (outcome of random variable W) given context c (previous words), I generally elide this conditioner, for brevity, writing $s(\breve{w}) \coloneqq -\log p(\breve{w} \mid c)$. Where necessary, it may explicitly denoted as $s(\breve{w} \mid c)$.

entropy (expected surprisal) of random variable Z:

$$\mathrm{H}(Z) \coloneqq \mathop{\mathbb{E}}_{p_Z}\left[\log \frac{1}{p(z)}\right]$$

- entropy of Z upon observing fixed outcome $\breve{w}$ (of random variable W):

$$\mathrm{H}(Z \mid \breve{w}) \coloneqq \mathop{\mathbb{E}}_{p_{Z \mid \breve{w}}}\left[\log \frac{1}{p(z \mid \breve{w})}\right]$$

- the **conditional entropy** of Z conditioned on random variable W:

$$\mathrm{H}(Z \mid W) \coloneqq \mathop{\mathbb{E}}_{p_{Z,W}}\left[\log \frac{1}{p(z \mid w)}\right]$$

entropy reduction upon observing $\breve{w}$ (can be negative):

$$\begin{aligned}\mathrm{ER}(\breve{w}) &\coloneqq \mathrm{H}(Z) - \mathrm{H}(Z \mid \breve{w}) \\ &= \mathop{\mathbb{E}}_{p_Z}\left[\log \frac{1}{p(z)}\right] - \mathop{\mathbb{E}}_{p_{Z \mid \breve{w}}}\left[\log \frac{1}{p(z \mid \breve{w})}\right]\end{aligned}$$

reconstruction information* of observed $\breve{w}$ under posterior $p_{Z|\breve{w}}$:

$$R(\breve{w}) \coloneqq \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[-\log p(\breve{w} \mid z)\right]$$

*Unlike other terms listed here, which are standard names for quantities in information theory, the term 'reconstruction information' is introduced in this dissertation. It quantifies the expected surprisal of the observation $\breve{w}$ conditioned on $z \sim p_{Z|\breve{w}}$.

Note that this quantity is nonnegative, and takes on its minimum value of zero iff the likelihood $p(\breve{w} \mid z) = 1$ everywhere in the support of the posterior. Conversely, note that $R(\breve{w})$ takes on its maximum value, equal to $s(\breve{w})$, in the case that the posterior equals the prior, which occurs iff the likelihood is constant everywhere in the support of the prior.

KL divergence (aka relative entropy) between two probability distributions $p \ll q$:

$$D_{\mathrm{KL}}(p \,\|\, q) \coloneqq \mathop{\mathbb{E}}_{p}\left[\log \frac{\mathrm{d}p}{\mathrm{d}q}\right]$$

with Radon-Nikodým derivative $\frac{\mathrm{d}p}{\mathrm{d}q}$. Or, with densities $p(z)$ and $q(z)$:

$$= \mathop{\mathbb{E}}_{p}\left[\log \frac{p(z)}{q(z)}\right]$$

- specifically, KL between posterior and prior:

$$D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, p_Z) = s(\breve{w}) - R(\breve{w})$$

- specifically, KL between posterior and proposal $q_{Z;\breve{w}}$ with $p_{Z|\breve{w}} \ll q_{Z;\breve{w}}$:

$$D_{\mathrm{KL}}(p_{Z|\breve{w}} \,\|\, q_{Z;\breve{w}}) = s(\breve{w}) - R(\breve{w}) + \mathop{\mathbb{E}}_{p_{Z|\breve{w}}}\left[\log \frac{p(z)}{q(z;\breve{w})}\right]$$

χ^2 **divergence** between two probability distributions $p \ll q$:

$$D_{\chi^2}(p \,\|\, q) \coloneqq \mathop{\mathbb{E}}_{q}\left[\left(\frac{\mathrm{d}p}{\mathrm{d}q} - 1\right)^2\right] = \mathop{\mathbb{E}}_{q}\left[\frac{(\mathrm{d}p - \mathrm{d}q)^2}{\mathrm{d}q}\right] = \mathop{\mathbb{E}}_{p}\left[\frac{\mathrm{d}p}{\mathrm{d}q}\right] - 1$$

Note that $D_{\mathrm{KL}}(p \,\|\, q) \leq \log(1 + D_{\chi^2}(p \,\|\, q))$ (see Gibbs & Su, 2002, Thm. 5).

cross entropy of q with respect to p:

$$\mathrm{H}(p, q) \coloneqq \mathop{\mathbb{E}}_{p}\left[\log \frac{1}{q(z)}\right]$$

pointwise mutual information between observations z, w of discrete random variables Z, W (can be negative):

$$\begin{aligned}\operatorname{pmi}(z, w) \coloneqq \log \frac{p(z, w)}{p(z)p(w)} &= \log \frac{p(z \mid w)}{p(z)} &&= \log \frac{p(w \mid z)}{p(w)} \\ &= s(z) - s(z \mid w) &&= s(w) - s(w \mid z)\end{aligned}$$

mutual information (expected pointwise mutual information) between two random variables Z and W:

$$\mathrm{I}(Z; W) \coloneqq D_{\mathrm{KL}}(p_{Z,W} \,\|\, p_Z \cdot p_W) = \mathop{\mathbb{E}}_{p_{Z,W}} \left[\log \frac{p(z, w)}{p(z)p(w)}\right]$$

(note, mututal information is equivalent to expected KL between posterior and prior)

$$= \mathop{\mathbb{E}}_{p_{Z,W}} \left[\log \frac{p(z \mid w)}{p(z)}\right] = \mathop{\mathbb{E}}_{p_W} \left[D_{\mathrm{KL}}(p_{Z|w} \,\|\, p_Z)\right]$$

(and likewise equivalent to expected entropy reduction)

$$= \mathrm{H}(Z) - \mathrm{H}(Z \mid W) = \mathop{\mathbb{E}}_{p_W} \left[\mathrm{H}(Z) - \mathrm{H}(Z \mid w)\right] = \mathop{\mathbb{E}}_{p_W} \left[\mathrm{ER}(w)\right]$$

Bibliography

Abney, S. P., & Johnson, M. (1991). Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, *20*(3), 233–250. https://doi.org/10.1007/bf01067217 (cit. on p. 39).

Adelman, J. S., & Brown, G. D. A. (2008). Modeling lexical decision: The form of frequency and diversity effects. *Psychological Review*, *115*(1), 214–227. https://doi.org/10.1037/0033-295X.115.1.214 (cit. on p. 14).

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., & Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, *32*(3). https://doi.org/10.1214/17-STS611 (cit. on pp. 26, 46, 61).

AI@Meta. (2024). *Llama 3 model card*. Meta. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (cit. on p. 165).

Allen, C., & Hospedales, T. M. (2019). Analogies explained: Towards understanding word embeddings. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (pp. 223–231, Vol. 97). PMLR. http://proceedings.mlr.press/v97/allen19a.html (cit. on pp. 90, 184).

Anderson, J. R. (1990, January). *The adaptive character of thought*. Psychology Press. https://doi.org/10.4324%2F9780203771730 (cit. on pp. 10, 62).

Anderson, J. R. (1991a). The place of cognitive architectures in a rational analysis. In *Architectures for intelligence* (pp. 1–24). Psychology Press. (Cit. on p. 10).

Anderson, J. R. (1991b). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485. https://doi.org/10.1017/S0140525X00070801 (cit. on pp. 10, 48).

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Psychology Press. https://doi.org/10.4324/9781315805696 (cit. on pp. 47, 62, 164).

Andrews, S. (1996). Lexical Retrieval and Selection Processes: Effects of Transposed-Letter Confusability. *Journal of Memory and Language*, *35*(6), 775–800. https://doi.org/10.1006/jmla.1996.0040 (cit. on p. 71).

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 (cit. on pp. 14, 48, 59).

Attias, H. (1999). A variational baysian framework for graphical models. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in Neural Information Processing Systems* (Vol. 12). MIT Press. Retrieved June 27, 2022, from https://proceedings.neurips.cc/paper/1999/hash/74563ba21a90da13dacf2a73e3ddefa7-Abstract.html (cit. on p. 22).

Attneave, F. (1959). *Applications of information theory to psychology: A summary of basic concepts, methods, and results*. Henry Holt. https://lccn.loc.gov/59008712 (cit. on pp. 2, 13, 27, 28).

Aurnhammer, C., & Frank, S. L. (2019). Comparing gated and simple recurrent neural network architectures as models of human sentence processing. *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, 112–118. http://hdl.handle.net/2066/213724 (cit. on pp. 30, 44, 47).

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201 (cit. on p. 12).

Bahdanau, D., Cho, K., & Bengio, Y. (2016, May 19). *Neural Machine Translation by Jointly Learning to Align and Translate*. (Cit. on p. 3).

Baldi, P. (2002). A computational theory of surprise. In M. Blaum, P. G. Farrell, & H. C. A. van Tilborg (Eds.), *Information, Coding and Mathematics: Proceedings of Workshop honoring Prof. Bob McEliece on his 60th birthday* (pp. 1–25). Springer US. https://doi.org/10.1007/978-1-4757-3585-7_1 (cit. on pp. 19, 29).

Baldi, P., & Itti, L. (2010). Of bits and wows: A Bayesian theory of surprise with applications to attention. *Neural Networks*, *23*(5), 649–666. https://doi.org/10.1016/j.neunet.2009.12.007 (cit. on p. 19).

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 (cit. on pp. 37, 38).

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 1–21. https://doi.org/10.2307/2983610 (cit. on p. 14).

Barton, S. B., & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & Cognition*, *21*(4), 477–487. https://doi.org/10.3758/BF03197179 (cit. on p. 105).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01 (cit. on p. 173).

Baum, C., & Veeravalli, V. (1994). A sequential procedure for multihypothesis testing. *IEEE Transactions on Information Theory*, *40*(6), 1994–2007. https://doi.org/10.1109/18.340472 (cit. on p. 14).

Bengio, Y., Ducharme, R., & Vincent, P. (2000). A neural probabilistic language model. *Advances in Neural Information Processing Systems*, *13*. Retrieved May 15, 2024, from https://proceedings.neurips.cc/paper_files/paper/2000/hash/728f206c2a01bf572b5940d7d9a8fa4c-Abstract.html (cit. on p. 3).

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137–1155. Retrieved May 15, 2024, from https://jmlr.csail.mit.edu/papers/v3/bengio03a.html (cit. on p. 3).

Berwick, R. C., & Weinberg, A. S. (1982). Parsing efficiency, computational complexity, and the evaluation of grammatical theories. *Linguistic Inquiry*, *13*(2), 165–191. http://www.jstor.org/stable/4178272 (cit. on p. 39).

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1168–1178. https://www.aclweb.org/anthology/P10-1119 (cit. on pp. 44, 47).

Bicknell, K., & Levy, R. (2012). Word predictability and frequency effects in a rational model of reading. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, *34*, 126–131. https://cogsci.mindmodeling.org/2012/papers/0035/ (cit. on pp. 44, 47).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning* (1st ed.). Springer. Retrieved June 10, 2022, from https://link.springer.com/book/9780387310732 (cit. on p. 22).

Blachman, N. (1968). The amount of information that y gives about X. *IEEE Transactions on Information Theory*, *14*(1), 27–31. https://doi.org/10.1109/tit.1968.1054094 (cit. on p. 33).

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021, October 6). *GPT-Neo: Large scale autoregressive language modeling with meshtensorflow* (Version v1.1.1). https://doi.org/10.5281/ZENODO.5551208 (cit. on pp. 49, 74, 165).

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., & Weinbach, S. (2022, May). GPT-NeoX-20B: An open-source autoregressive language model. In A. Fan, S. Ilic, T. Wolf, & M. Gallé (Eds.), *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*

(pp. 95–136). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.9 (cit. on pp. 74, 165).

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773 (cit. on pp. 19, 22).

Bod, R. (2006). An all-subtrees approach to unsupervised parsing. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 865–872. https://doi.org/10.3115/1220175.1220284 (cit. on p. 92).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022, July 12). *On the opportunities and risks of foundation models*. arXiv: 2108.07258 [cs]. (Cit. on p. 3). Note: Authored by the Center for Research on Foundation Models (CRFM) at the Stanford Institute for Human-Centered Artificial Intelligence (HAI).

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1). https://doi.org/10.16910/jemr.2.1.1 (cit. on pp. 30, 39, 44, 47).

Bouchard-Côté, A., Petrov, S., & Klein, D. (2009). Randomized pruning: Efficiently calculating expectations in large dynamic programs. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2009/file/e515df0d202ae52fcebb14295743063b-Paper.pdf (cit. on p. 62).

Boyce, V. (2022, March 27). *Amaze-natural-stories*. Retrieved September 24, 2022, from https://github.com/vboyce/amaze-natural-stories (cit. on p. 146).

Boyce, V., & Levy, R. (2020, September). A-maze of Natural Stories: Texts are comprehensible using the Maze task [Potsdam, Germany]. https://amlap2020.github.io/a/154.pdf (cit. on p. 49).

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 (cit. on pp. 14, 37, 38, 44, 47).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T.

Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (cit. on pp. 3, 37, 40, 45, 49, 74).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023, April 13). *Sparks of artificial general intelligence: Early experiments with GPT-4*. arXiv: 2303.12712 [cs]. (Cit. on p. 3).

Burchill, Z. J., & Jaeger, T. F. (2024). How reliable are standard reading time analyses? Hierarchical bootstrap reveals substantial power over-optimism and scale-dependent Type I error inflation. *Journal of Memory and Language*, *136*, 104494. https://doi.org/10.1016/j.jml.2023.104494 (cit. on pp. 73, 74, 76).

Bürkner, P.-C. (2017). **brms**: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01 (cit. on pp. 76, 77, 169).

Burnham, K. P., & Anderson, D. R. (Eds.). (2004). *Model Selection and Multimodel Inference*. Springer. https://doi.org/10.1007/b97636 (cit. on p. 17).

Buys, J. (2018). *Incremental generative models for syntactic and semantic natural language processing* [Doctoral dissertation, University of Oxford]. Retrieved June 14, 2022, from https://ora.ox.ac.uk/objects/uuid:a9a7b5cf-3bb1-4e08-b109-de06bf387d1d (cit. on p. 61).

Cai, W. (2014). Making comparisons fair: How LS-means unify the analysis of linear models. *Proceedings of the SAS Global Forum*, *Paper SAS060-2014*, 1–22. https://support.sas.com/resources/papers/proceedings14/SAS060-2014.pdf (cit. on p. 171).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). **Stan**: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01 (cit. on p. 76).

Carroll, G., & Charniak, E. (1992). *Two experiments on learning probabilistic dependency grammars from corpora* (AAAI technical report No. WS-92-01). AAAI. https://aaai.org/Library/Workshops/1992/ws92-01-001.php (cit. on p. 87).

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289 (cit. on p. 3).

Chater, N., Crocker, M. J., & Pickering, M. J. (1998, November 19). The rational analysis of inquiry: The case of parsing. In M. Oaksford & N. Chater (Eds.), *Rational models of*

cognition (pp. 441–468). Oxford University Press. https://doi.org/10.1093/oso/9780198524151.003.0020 (cit. on pp. 11, 33, 34).

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, *10*(7), 335–344. https://doi.org/10.1016/j.tics.2006.05.006 (cit. on pp. 1, 11).

Chatterjee, S., & Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, *28*(2). https://doi.org/10.1214/17-aap1326 (cit. on pp. 25, 46, 61).

Chen, R., Levy, R., & Eisape, T. (2021). On factors influencing typing time: Insights from a viral online typing game. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*. Retrieved September 8, 2023, from https://escholarship.org/uc/item/8bb8v4g3 (cit. on p. 81).

Chen, Y. (2005). Another look at rejection sampling through importance sampling. *Statistics & Probability Letters*, *72*(4), 277–283. https://doi.org/10.1016/j.spl.2005.01.002 (cit. on p. 26).

Chopin, N., & Papaspiliopoulos, O. (2020, October 2). *An introduction to sequential Monte Carlo* (1st ed.). Springer. https://doi.org/10.1007/978-3-030-47845-2 (cit. on pp. 25, 41).

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *27th Annual Meeting of the Association for Computational Linguistics*, 76–83. https://doi.org/10.3115/981623.981633 (cit. on p. 87).

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276–286. https://doi.org/10.18653/v1/W19-4828 (cit. on p. 97).

Clark, T. H., Meister, C., Pimentel, T., Hahn, M., Cotterell, R., Futrell, R., & Levy, R. (2023). A cross-linguistic pressure for uniform information density in word order. *Transactions of the Association for Computational Linguistics*, *11*, 1048–1065. https://doi.org/10.1162/tacl_a_00589 (cit. on p. 12).

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. https://doi.org/10.1016/j.cognition.2008.04.004 (cit. on p. 1).

Collins, M., & Roark, B. (2004). Incremental parsing with the perceptron algorithm. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 111–118. https://doi.org/10.3115/1218955.1218970 (cit. on p. 37).

Conneau, A., Kruszewski, G., Lample, G., Barrault, L., & Baroni, M. (2018). What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), 2126–2136. https://doi.org/10.18653/v1/P18-1198 (cit. on p. 97).

Conneau, A., & Lample, G. (2019, December). Cross-lingual language model pretraining. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32 (NeurIPS 2019)* (pp. 7057–7067). https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html (cit. on p. 89).

Costa, F. (2003). Towards incremental parsing of natural language using recursive neural networks. *Applied Intelligence*, *19*(1/2), 9–25. https://doi.org/10.1023/a:1023860521975 (cit. on p. 39).

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley. https://doi.org/10.1002/047174882x (cit. on p. 17).

Cramer, B. (2007). Limitations of current grammar induction algorithms. *Proceedings of the ACL 2007 Student Research Workshop*, 43–48. https://www.aclweb.org/anthology/P07-3008 (cit. on p. 92).

Cronbach, L. J. (1953, November). *A consideration of information theory and utility theory as tools for psychometric problems* (Technical report No. 1, contract N60ri-07146, Office of Naval Research). University of Illinois. Retrieved May 14, 2024, from https://apps.dtic.mil/sti/citations/tr/AD0025723 (cit. on p. 33).

Cutter, M. G., Filik, R., & Paterson, K. B. (2022). Do readers maintain word-level uncertainty during reading? A pre-registered replication study. *Journal of Memory and Language*, *125*, 104336. https://doi.org/10.1016/j.jml.2022.104336 (cit. on p. 75).

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/p19-1285 (cit. on pp. 40, 49).

Davies, M. (2008). *Word frequency data from the Corpus of Contemporary American English (COCA)* (Version 1990–2019). Retrieved November 22, 2023, from https://www.wordfrequency.info (cit. on p. 72).

de Finetti, B. (1972). *Probability, induction and statistics: The art of guessing*. Wiley. https://lccn.loc.gov/73165954 (cit. on p. 11).

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540. https://doi.org/10.1146/annurev-linguistics-031220-010811 (cit. on p. 11).

de Marneffe, M.-C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. *Proceedings of the Fifth International Conference on*

Language Resources and Evaluation (LREC'06). http://www.lrec-conf.org/proceedings/lrec2006/pdf/440%E2%82%9Adf.pdf (cit. on p. 89).

de Marneffe, M.-C., & Manning, C. (2008). *Stanford typed dependencies manual*. manual. Version Stanford Parser v.3.7.0. Stanford NLP. https://nlp.stanford.edu/software/dependencies%E2%82%98anual.pdf (cit. on p. 94).

de Marneffe, M.-C., & Manning, C. D. (2008). The Stanford typed dependencies representation. *Coling 2008: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8. https://www.aclweb.org/anthology/W08-1301 (cit. on p. 89).

de Marneffe, M.-C., & Nivre, J. (2019). Dependency grammar. *Annual Review of Linguistics*, *5*(1), 197–218. https://doi.org/10.1146/annurev-linguistics-011718-011842 (cit. on p. 85).

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 (cit. on pp. 14, 37, 44).

de Paiva Alves, E. (1996). The selection of the most probable dependency structure in Japanese using mutual information. *34th Annual Meeting of the Association for Computational Linguistics*, 372–374. https://doi.org/10.3115/981863.981919 (cit. on pp. 86, 87).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423 (cit. on pp. 3, 83, 87, 89).

Dotlačil, J. (2021). Parsing as a cue-based retrieval model. *Cognitive science*, *45*(8), e13020. https://doi.org/10.1111/cogs.13020 (cit. on pp. 37, 48).

Doucet, A., Freitas, N., & Gordon, N. (Eds.). (2001). *Sequential Monte Carlo methods in practice*. Springer. https://doi.org/10.1007/978-1-4757-3437-9 (cit. on pp. 25, 60).

Doucet, A., & Johansen, A. M. (2008, December). A tutorial on particle filtering and smoothing: Fifteen years later. In D. Crisan & B. Rozovskiĭ (Eds.), *The Oxford Handbook of Nonlinear Filtering* (pp. 656–704). Oxford University Press. http://www.stats.ox.ac.uk/~doucet/doucet_johansen_tutorialPF2011.pdf (cit. on p. 60).

Note: Version 1.1 – December 2008 with typographical corrections March 2012.

Du, W., Lin, Z., Shen, Y., O'Donnell, T. J., Bengio, Y., & Zhang, Y. (2020). Exploiting syntactic structure for better language modeling: A syntactic distance approach. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6611–6628. https://doi.org/10.18653/v1/2020.acl-main.591 (cit. on p. 90).

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, *13*(2), 94–102. https://doi.org/10.1145/362007.362035 (cit. on p. 39).

Earman, J. (1992). *Bayes or bust? a critical examination of Bayesian confirmation theory*. MIT Press. (Cit. on p. 11).

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436. https://doi.org/10.1007/BF02143160 (cit. on p. 37).

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 (cit. on pp. 37, 38).

Eisner, J. (1997). *An empirical comparison of probability models for dependency grammar* (Technical report No. IRCS-96-11). Institute for Research in Cognitive Science, University of Pennsylvania. (Cit. on pp. 88, 191).

Eisner, J. M. (1996). Three new probabilistic models for dependency parsing: An exploration. *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C96-1058 (cit. on pp. 88, 191).

Elvira, V., Míguez, J., & Djurić, P. M. (2017). Adapting the number of particles in sequential Monte Carlo methods through an online scheme for convergence assessment. *IEEE Transactions on Signal Processing*, *65*(7), 1781–1794. https://doi.org/10.1109/TSP.2016.2637324 (cit. on p. 61).

Engelmann, F. (2016). *Toward an integrated model of sentence processing in reading* [Doctoral dissertation, Universität Potsdam]. Retrieved October 12, 2022, from https://publishup.uni-potsdam.de/frontdoor/index/index/docId/10086 (cit. on p. 48).

Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association on retrieval processes: A computational account. *Cognitive Science*, *43*(12), e12800. https://doi.org/10.1111/cogs.12800 (cit. on p. 48).

Fano, R. M. (1961). *Transmission of information: A statistical theory of communications* (1st ed.). MIT Press. https://b-ok.cc/book/5577269/50d40b (cit. on pp. 5, 86).

Farmer, T. A., Misyak, J. B., & Christiansen, M. H. (2012). Individual differences in sentence processing. In K. McRae, M. Joanisse, & M. Spivey (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 353–364). Cambridge University Press. https://doi.org/10.1017/CBO9781139029377.018 (cit. on p. 45).

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 (cit. on pp. 37, 44, 56).

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1 (cit. on p. 45).

Fossum, V., & Levy, R. (2012). Sequential vs. Hierarchical syntactic models of human incremental sentence processing. *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, 61–69. https://aclanthology.org/W12-1706 (cit. on p. 45).

Fox, D. (2003). Adapting the sample size in particle filters through KLD-Sampling. *The International Journal of Robotics Research*, *22*(12), 985–1003. https://doi.org/10.1177/0278364903022012001 (cit. on p. 61).

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 (cit. on pp. 44, 45).

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 (cit. on pp. 37, 44).

Frazier, L. (1987). Syntactic processing: Evidence from Dutch. *Natural Language & Linguistic Theory*, *5*(4), 519–559. https://doi.org/10.1007/BF00138988 (cit. on p. 37).

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325. https://doi.org/10.1016/0010-0277(78)90002-1 (cit. on p. 36).

Freer, C., Mansinghka, V. K., & Roy, D. (2010). When are probabilistic programs probably computationally tractable? *NIPS Workshop on Monte Carlo Methods for Modern Applications*. http://montecarlo.wdfiles.com/local--files/contributed-abstracts/nipsmc2010_freer_etal.pdf (cit. on pp. 41, 46).

Futrell, R. (2017). *Memory and locality in natural language* [Doctoral dissertation, Massachusetts Institute of Technology / Massachusetts Institute of Technology. Department of Brain and Cognitive Sciences]. http://hdl.handle.net/1721.1/114075 (cit. on p. 44).

Futrell, R., Gibson, E., & Levy, R. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814. https://doi.org/10.1111/cogs.12814 (cit. on pp. 12, 44, 107).

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 (cit. on pp. 49, 50, 73, 155).

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 688–698. https://www.aclweb.org/anthology/E17-1065 (cit. on pp. 9, 107).

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 (cit. on pp. 6, 86, 103).

Gale, W., & Church, K. (1994, January 1). What is wrong with adding one? In *Corpus-Based Research into Language* (pp. 189–198, Vol. 12). Brill. https://doi.org/10.1163/9789004653566_015 (cit. on p. 3).

Gao, L. (2021, May). *On the sizes of OpenAI API models*. EleutherAI Blog. Retrieved December 13, 2021, from https://blog.eleuther.ai/gpt3-model-sizes/ (cit. on p. 50).

Garner, W. R. (1962). *Uncertainty and structure as psychological concepts*. Wiley. https://lccn.loc.gov/62010919 (cit. on p. 2).

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2014). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCamDat). *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*, 240–254. https://www.lingref.com/cpp/slrf/2012/abstract3100.html (cit. on p. 81).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (Third edition). CRC Press, Taylor; Francis Group. http://www.stat.columbia.edu/~gelman/book/ (cit. on p. 77).

Gibbs, A. L., & Su, F. E. (2002). On choosing and bounding probability metrics. *International Statistical Review*, *70*(3), 419–435. https://doi.org/10.1111/j.1751-5823.2002.tb00178.x (cit. on pp. 27, 109).

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. https://doi.org/10.1016/S0010-0277(98)00034-1 (cit. on p. 39).

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium* (pp. 94–126). The MIT Press. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.592.5833 (cit. on pp. 39, 107).

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, *24*(7), 1079–1088. https://doi.org/10.1177/0956797612463705 (cit. on p. 1).

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 (cit. on pp. 14, 37, 38, 44, 45, 51, 52, 56, 59, 149, 150, 158).

Goodkind, A., & Bicknell, K. (2021). Local word statistics affect reading times independently of surprisal. *arXiv*. https://doi.org/10.48550/ARXIV.2103.04469 (cit. on pp. 37, 44, 51, 56, 71, 158).

Graf, T., Monette, J., & Zhang, C. (2017). Relative clauses as a benchmark for minimalist parsing. *Journal of Language Modelling*, *5*(1). https://doi.org/10.15398/jlm.v5i1.157 (cit. on p. 39).

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A. H., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., ... Hajishirzi, H. (2024, February 27). *OLMo: Accelerating the science of language models*. arXiv: 2402.00838 [cs]. (Cit. on pp. 74, 165).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1195–1205. https://doi.org/10.18653/v1/N18-1108 (cit. on p. 49).

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119. https://doi.org/10.1073/pnas.2122602119 (cit. on pp. 9, 12).

Hahn, M., Keller, F., Bisk, Y., & Belinkov, Y. (2019). Character-based Surprisal as a Model of Reading Difficulty in the Presence of Errors. *Proceedings of the 41th Annual Meeting of the Cognitive Science Society, CogSci 2019: Creativity + Cognition + Computation, Montreal, Canada, July 24-27, 2019*, 401–407. https://doi.org/10.48550/arXiv.1902.00595 (cit. on p. 81).

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 (cit. on pp. 2, 4, 9, 12, 13, 36, 37, 44, 46, 47, 66, 68).

Hale, J. T. (2003a). *Grammar, uncertainty and sentence processing* [Doctoral dissertation, Johns Hopkins University]. https://www.proquest.com/docview/288510490 (cit. on pp. 13, 32, 33).

Hale, J. T. (2003b). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123. https://doi.org/10.1023/A:1022492123056 (cit. on pp. 9, 12, 18, 32, 34).

Hale, J. T. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*(4), 643–672. https://doi.org/10.1207/s15516709cog0000_64 (cit. on p. 32).

Hale, J. T. (2014, September). *Automaton theories of human sentence comprehension*. CSLI Publications, Center for the Study of Language; Information. Retrieved July 1, 2022, from https://csli.sites.stanford.edu/publications/csli-studies-computational-linguistics/automaton-theories-human-sentence-comprehension (cit. on pp. 12, 39).

Hale, J. T. (2016). Information-theoretical Complexity Metrics. *Language and Linguistics Compass*, *10*(9), 397–412. https://doi.org/10.1111/lnc3.12196 (cit. on p. 9).

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 (cit. on pp. 44, 45, 59).

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877. https://doi.org/10.3758/s13421-018-0808-6 (cit. on p. 73).

Heiss, A. (2021, November 10). *A guide to correctly calculating posterior predictions and average marginal effects with multilievel Bayesian models*. Andrew Heiss's blog. https://doi.org/10.59350/wbn93-edb02 (cit. on p. 171).

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2733–2743. https://doi.org/10.18653/v1/D19-1275 (cit. on p. 188).

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4129–4138. https://doi.org/10.18653/v1/N19-1419 (cit. on pp. 92, 97, 191).

Hick, W. E. (1952). On the Rate of Gain of Information. *Quarterly Journal of Experimental Psychology*, *4*(1), 11–26. https://doi.org/10.1080/17470215208416600 (cit. on p. 2).

Hinton, G. E. (1986). Learning distributed representations of concepts. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, *1*, 12. https://www.cs.toronto.edu/~hinton/absps/families.pdf (cit. on p. 3).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on pp. 3, 40).

Hockett, C. F. (1953). Review of the mathematical theory of communication. *Language*, *29*(1), 69–93. https://doi.org/10.2307/410457 (cit. on p. 2).

Hofmann, M. J., Biemann, C., & Remus, S. (2017). Benchmarking n-grams, topic models and recurrent neural networks by cloze completions, EEGs and eye movements. In *Cognitive Approach to Natural Language Processing* (pp. 197–215). Elsevier. https://doi.org/10.1016/B978-1-78548-253-3.50010-X (cit. on p. 44).

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 (cit. on pp. 37, 38, 44, 45, 47, 56, 59, 153, 158).

Hoover, J. L., Du, W., Sordoni, A., & O'Donnell, T. J. (2021). Linguistic dependencies and statistical dependence. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2941–2963. https://doi.org/10.18653/v1/2021.emnlp-main.234 (cit. on pp. 7, 84).

Hoover, J. L., Sonderegger, M., & O'Donnell, T. J. (2022, September 6). *With better language models, processing time is superlinear in surprisal* (Poster). York, England. https://virtual.oxfordabstracts.com/#/event/3067/submission/297 (cit. on p. 7).

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind: Discoveries in Cognitive Science*, *7*, 350–391. https://doi.org/10.1162/opmi_a_00086 (cit. on pp. 7, 36).

Howson, C., & Urbach, P. (2006). *Scientific reasoning: The Bayesian approach* (3rd ed.). Open Court Publishing. (Cit. on p. 11).

Htut, P. M., Phang, J., Bordia, S., & Bowman, S. R. (2019, November 27). *Do attention heads in BERT track syntactic dependencies?* arXiv: 1911.12246 [cs]. (Cit. on p. 97).

Hu, J. (2023, May). *Neural language models and human linguistic knowledge* [Doctoral dissertation, Massachusetts Institute of Technology]. Retrieved May 15, 2024, from https://dspace.mit.edu/handle/1721.1/152578

Hu, X., Mi, H., Li, L., & de Melo, G. (2022, November 2). *Fast-R2D2: A pretrained recursive neural network based on pruned CKY for grammar induction and text representation*. arXiv: 2203.00281 [cs]. (Cit. on p. 39).

Hu, X., Mi, H., Wen, Z., Wang, Y., Su, Y., Zheng, J., & de Melo, G. (2021). R2D2: Recursive transformer based on differentiable tree for interpretable hierarchical language modeling. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4897–4908. https://doi.org/10.18653/v1/2021.acl-long.379 (cit. on p. 39).

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 (cit. on pp. 14, 74, 75).

Huang, K.-J., & Staub, A. (2021). Using eye tracking to investigate failure to notice word transpositions in reading. *Cognition*, *216*, 104846. https://doi.org/10.1016/j.cognition.2021.104846 (cit. on p. 71).

Icard, T. (2023, September 12). *Resource rationality* [Book draft]. https://philpapers.org/archive/ICARRT.pdf

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. https://doi.org/10.1016/j.visres.2008.09.007 (cit. on p. 19).

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech* [Doctoral dissertation, Stanford University]. https://searchworks.stanford.edu/view/6686010 (cit. on p. 47).

Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*. Retrieved October 12, 2022, from https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00617 (cit. on p. 48).

Jegerski, J. (2013). Self-paced reading. In *Research Methods in Second Language Psycholinguistics* (pp. 20–49). Routledge. (Cit. on p. 73).

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023, October 10). *Mistral 7B*. arXiv: 2310.06825 [cs]. (Cit. on pp. 74, 165).

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., ... Sayed, W. E. (2024, January 8). *Mixtral of experts*. arXiv: 2401.04088 [cs]. (Cit. on pp. 74, 165).

Jin, L., & Schuler, W. (2020). Memory-bounded neural incremental parsing for psycholinguistic prediction. *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, 48–61. https://doi.org/10.18653/v1/2020.iwpt-1.6 (cit. on pp. 39, 44).

Johnson, R. L. (2009). The quiet clam is quite calm: Transposed-letter neighborhood effects on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 943–969. https://doi.org/10.1037/a0015572 (cit. on p. 71).

Johnson, R. L., Perea, M., & Rayner, K. (2007). Transposed-letter effects in reading: Evidence from eye movements and parafoveal preview. *Journal of Experimental Psychology: Human*

Perception and Performance, *33*(1), 209–229. https://doi.org/10.1037/0096-1523.33.1.209 (cit. on p. 71).

Jurafsky, D. (2003, April 8). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic Linguistics* (pp. 39–96). The MIT Press. https://doi.org/10.7551/mitpress/5582.003.0006 (cit. on p. 11).

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194. https://doi.org/10.1207/s15516709cog2002_1 (cit. on pp. 44, 62).

Jurafsky, D., & Martin, J. H. (2024, February 3). N-gram Language Models. In *Speech and Language Processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/3.pdf (cit. on p. 3).

Kelley, P. (2018). *More people understand Eschers than the linguist does: The causes and effects of grammatical illusions* [Doctoral dissertation, Michigan State University]. Retrieved February 22, 2023, from https://www.proquest.com/docview/2041968142/abstract/42208A941C8E47AFPQ/1 (cit. on p. 105).

Keynes, J. M. (1921). *A treatise on probability*. Macmillan; Co., retrieved May 14, 2024, from https://lccn.loc.gov/21020432 (cit. on p. 11).

Kim, T., Choi, J., Edmiston, D., & Lee, S.-g. (2020). Are pre-trained language models aware of phrases? Simple but strong baselines for grammar induction. *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. https://openreview.net/forum?id=H1xPR3NtPB (cit. on p. 97).

Kim, T., Li, B., & Lee, S.-g. (2020). Chart-based zero-shot constituency parsing on multiple languages. (Cit. on p. 97).

Klein, D., & Manning, C. (2004). Corpus-based induction of syntactic structure: Models of dependency and constituency. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 478–485. https://doi.org/10.3115/1218955.1219016 (cit. on pp. 87, 90–92, 192, 195).

Kneser, R., & Ney, H. (1995). Improved backing-off for M-gram language modeling. *1995 International Conference on Acoustics, Speech, and Signal Processing*, *1*, 181–184 vol.1. https://doi.org/10.1109/ICASSP.1995.479394 (cit. on p. 3).

Kuhlmann, M. (2010). *Dependency structures and lexicalized grammars: An algebraic approach* (Vol. 6270). Springer. https://www.ida.liu.se/~marku61/pdf/kuhlmann2010dependency.pdf (cit. on p. 88).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694 (cit. on p. 17).

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32–59. https://doi.org/10.1080/23273798.2015.1102299 (cit. on p. 11).

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 (cit. on pp. 44, 45, 59, 60).

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13 (cit. on p. 173).

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241. https://doi.org/10.1111/cogs.12414 (cit. on p. 1).

Laverghetta, A., Nighojkar, A., Mirzakhalov, J., & Licato, J. (2022). Predicting human psychometric properties using computational language models. In M. Wiberg, D. Molenaar, J. González, J.-S. Kim, & H. Hwang (Eds.), *Quantitative Psychology* (pp. 151–169). Springer International Publishing. https://doi.org/10.1007/978-3-031-04572-1_12 (cit. on pp. 45, 59).

Leivada, E. (2020). Language Processing at Its Trickiest: Grammatical Illusions and Heuristics of Judgment. *Languages*, *5*(3), 29. https://doi.org/10.3390/languages5030029 (cit. on p. 31).

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912–928. https://doi.org/10.1111/oik.05985 (cit. on pp. 76, 77).

Lenth, R. V. (2024). ***emmeans****: Estimated marginal means, aka least-squares means* (Version 1.10.0). https://CRAN.R-project.org/package=emmeans (cit. on pp. 77, 78, 171, 173).

Levy, O., & Goldberg, Y. (2014, December). Neural word embedding as implicit matrix factorization. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27 (NIPS 2014)* (pp. 2177–2185). https://proceedings.neurips.cc/paper/2014/hash/feab05aa91085b7a8012516bc3533958-Abstract.html (cit. on pp. 90, 184).

Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity* [Doctoral dissertation, Stanford University]. https://www.proquest.com/dissertations-theses/probabilistic-models-word-order-syntactic/docview/305432573/se-2?accountid=12339 (cit. on pp. 2, 12, 13, 19, 20, 44, 46, 47, 61).

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006 (cit. on pp. 1, 2, 4, 9, 12–14, 20, 36, 37, 39, 44, 47, 66–68).

Levy, R. (2008b). A noisy-channel model of human sentence comprehension under uncertain input. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243. https://aclanthology.org/D08-1025 (cit. on p. 44).

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1055–1065. https://aclanthology.org/P11-1106 (cit. on p. 44).

Levy, R. (2013). Memory and surprisal in human sentence comprehension. In R. P. G. van Gompel (Ed.), *Sentence processing* (pp. 78–114). Psychology Press. https://www.mit.edu/%20rplevy/papers/levy-2013-memory-and-surprisal-corrected.pdf (cit. on pp. 1, 9, 12, 13, 44, 47, 67).

Levy, R. (2018, July). Communicative efficiency, uniform information density, and the rational speech act theory. In C. Kalish, J. Z. Martina Rau, & T. Rogers (Eds.), *Proceedings of the 40th annual meeting of the cognitive science society* (pp. 684–689). https://cogsci.mindmodeling.org/2018/papers/0146/ (cit. on p. 44).

Levy, R., Fedorenko, E., & Gibson, E. (2013). The syntactic complexity of Russian relative clauses. *Journal of Memory and Language*, *69*(4), 461–495. https://doi.org/10.1016/j.jml.2012.10.005 (cit. on p. 156).

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 (cit. on pp. 12, 47).

Levy, R., Reali, F., & Griffiths, T. L. (2008). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21, proceedings of the twenty-second annual conference on neural information processing systems, vancouver, british columbia, canada, december 8-11, 2008* (pp. 937–944). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2008/hash/a02ffd91ece5e7efeb46db8f10a74059-Abstract.html (cit. on pp. 40, 60, 61).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703 (cit. on p. 89).

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. https://doi.org/10.1207/s15516709cog0000_25 (cit. on pp. 37, 48, 107).

Li, X. L., & Eisner, J. (2019). Specializing word embeddings (for parsing) by information bottleneck. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2744–2754. https://doi.org/10.18653/v1/D19-1276 (cit. on pp. 93, 188).

Liang, D., Krishnan, R. G., Hoffman, M. D., & Jebara, T. (2018). Variational autoencoders for collaborative filtering. *Proceedings of the 2018 World Wide Web Conference*, 689–698. https://doi.org/10.1145/3178876.3186150 (cit. on p. 19).

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005. https://doi.org/10.1214/aoms/1177728069 (cit. on p. 33).

Linzen, T., & Jaeger, F. (2014, June). Investigating the role of entropy in sentence processing. In V. Demberg & T. O'Donnell (Eds.), *Proceedings of the Fifth Workshop on Cognitive Modeling and Computational Linguistics* (pp. 10–18). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-2002 (cit. on p. 34).

Llama team. (2024, July 23). *The Llama 3 herd of models*. AI@Meta. Retrieved July 25, 2024, from https://ai.meta.com/research/publications/the-llama-3-herd-of-models/ (cit. on pp. 74, 165).

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171 (cit. on p. 153).

Lounsbury, F. G. (1954). Transitional probability, linguistic structure, and systems of habit-family hierarchies. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguistics* (pp. 93–101, Vol. Psycholinguistics: A survey of theory and research problems). Waverly Press Baltimore. https://publish.iupress.indiana.edu/read/db9a6002-fd15-44c0-a60e-65b6af037d2b/section/cebab226-583c-4176-a54c-0c8461c45bbc#fn47 (cit. on p. 33).

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 (cit. on p. 44).

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology*, *7*(2), 183–188. https://doi.org/10.1037/1089-2680.7.2.183 (cit. on p. 2).

Luong, T., O'Donnell, T., & Goodman, N. (2015). Evaluating models of computation and storage in human sentence processing. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 14–21. https://doi.org/10.18653/v1/W15-2403 (cit. on p. 39).

Lupker, S. J., Perea, M., & Davis, C. J. (2008). Transposed-letter effects: Consonants, vowels and letter frequency. *Language and Cognitive Processes*, *23*(1), 93–116. https://doi.org/10.1080/01690960701579714 (cit. on p. 71).

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*(4), 590–604. https://doi.org/10.1162/neco.1992.4.4.590 (cit. on p. 33).

Magerman, D. M., & Marcus, M. P. (1990). Parsing a natural language using mutual information statistics. *AAAI*, *90*, 984–989. https://www.aaai.org/Library/AAAI/1990/aaai90-147.php (cit. on pp. 86, 87).

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117 (cit. on p. 97).

Marcus, M. P. (1978). *A theory of syntactic recognition for natural language*. http://hdl.handle.net/1721.1/16176 (cit. on pp. 36, 39).

Marcus, M. P., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., & Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 8-11, 1994*. https://www.aclweb.org/anthology/H94-1020 (cit. on p. 89).

Mareček, D. (2012). *Unsupervised dependency parsing*. http://ufal.mff.cuni.cz/biblio/attachments/2012-marecek-m1481417340536440366.pdf (cit. on pp. 190, 191).

Mareček, D., & Rosa, R. (2018). Extracting syntactic trees from transformer encoder self-attentions. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 347–349. https://doi.org/10.18653/v1/W18-5444 (cit. on p. 97).

Mareček, D., & Rosa, R. (2019). From balustrades to Pierre Vinken: Looking for syntax in transformer self-attentions. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 263–275. https://doi.org/10.18653/v1/W19-4827 (cit. on p. 97).

Markov, A. A. (1913). Essai d'une recherche statistique sur le texte du roman "Eugène Oněgin", illustrant la liaison des épreuves en chaîne. *Bulletin de l'Académie Impériale des Sciences de St.-Pétersbourg. VI série*, *7*(3), 153–162. https://www.mathnet.ru/eng/im6612 (cit. on p. 3).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information* [Reprinted by MIT Press, 2010]. W.H. Freeman. (Cit. on pp. 10, 47).

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861–904. https://doi.org/10.1017/S0142716418000036 (cit. on p. 73).

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522–523. https://doi.org/10.1038/244522a0 (cit. on p. 37).

Marslen-Wilson, W. D. (1975). Sentence perception as an interactive parallel process. *Science (New York, N.Y.)*, *189*(4198), 226–228. https://doi.org/10.1126/science.189.4198.226 (cit. on p. 37).

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 (cit. on pp. 37, 38, 44).

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 (cit. on pp. 37, 38, 44).

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Second edition). CRC Press. https://xcelab.net/rm/ (cit. on pp. 76, 77).

Meister, C., Cotterell, R., & Vieira, T. (2020). If beam search is the answer, what was the question? *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2173–2185. https://doi.org/10.18653/v1/2020.emnlp-main.170 (cit. on p. 62).

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 (cit. on pp. 12, 14, 37, 38, 47, 51, 56, 158).

Meister, C., Vieira, T., & Cotterell, R. (2020). Best-first beam search. *Transactions of the Association for Computational Linguistics*, *8*, 795–809. https://doi.org/10.1162/tacl_a_00346 (cit. on p. 62).

Merkx, D., & Frank, S. L. (2021). Human sentence processing: Recurrence or attention? *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. https://doi.org/10.18653/v1/2021.cmcl-1.2 (cit. on pp. 30, 44, 45, 47, 59).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26:*

27th annual conference on neural information processing systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States (pp. 3111–3119). https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html (cit. on pp. 90, 184).

Mikolov, T., Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. https://www.aclweb.org/anthology/N13-1090 (cit. on p. 3).

Miller, G. A. (1957). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81. https://doi.org/10.1037/h0043158 (cit. on p. 2).

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In D. Luce (Ed.), *Handbook of mathematical psychology* (pp. 2–419). John Wiley & Sons. https://www.semanticscholar.org/paper/Finitary-models-of-language-users-Miller-Chomsky/4f3695d5dd36bb0abd91c02d2725463fca556f46 (cit. on p. 36).

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference* [Doctoral dissertation, Massachusetts Institute of Technology]. Retrieved March 30, 2023, from https://dspace.mit.edu/handle/1721.1/86583 (cit. on p. 22).

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. (Cit. on pp. 44, 158).

Mitchell, J., Lapata, M., Demberg, V., & Keller, F. (2010). Syntactic and semantic factors in processing difficulty: An integrated measure. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 196–206. https://www.aclweb.org/anthology/P10-1021 (cit. on pp. 30, 44, 47).

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 (cit. on pp. 31, 32, 105).

Nádas, A. (1984). Estimation of probabilities in the language model of the IBM speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(4), 859–861. https://doi.org/10.1109/TASSP.1984.1164378 (cit. on p. 3).

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using **brms**: A case study of gender effects on vowel variability in standard indonesian. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1225–1242. https://doi.org/10.1044/2018_JSLHR-S-18-0006 (cit. on p. 77).

Narayanan, S., & Jurafsky, D. (2001). A Bayesian model predicts human parse preference and reading times in sentence processing. *Advances in Neural Information Processing Systems*, *14*. Retrieved June 28, 2022, from https://proceedings.neurips.cc/paper/2001/hash/f15d337c70078947cfe1b5d6f0ed3f13-Abstract.html (cit. on pp. 12, 44).

Narayanan, S., & Jurafsky, D. (2004, November 29). *A Bayesian model of human sentence processing* [Unpublished manuscript]. https://web.stanford.edu/~jurafsky/narayananjurafsky04.pdf (cit. on pp. 12, 44, 47).

Newell, A. (1981). The knowledge level: Presidential address. *AI Magazine*, *2*(2), 1. https://doi.org/10.1609/aimag.v2i2.99 (cit. on p. 10)
Also published as: Arificial intelligence, 18(1), 87–127, 1982.

Newell, A., & Paul, R. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive Skills and Their Acquisition*. Psychology Press. https://doi.org/10.4324/9780203728178-6 (cit. on p. 39).

Nicenboim, B., Schad, D., & Vasishth, S. (2024, March 11). *An introduction to Bayesian data analysis for cognitive science*. Bookdown. Retrieved May 21, 2024, from https://vasishth.github.io/bayescogsci/book/ (cit. on pp. 75, 77).

Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34. https://doi.org/10.1016/j.jml.2017.08.004 (cit. on p. 48).

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, *40*, 26–55. https://doi.org/10.1017/S0267190520000057 (cit. on p. 73).

Nivre, J. (2008). Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, *34*(4), 513–553. https://doi.org/10.1162/coli.07-056-R1-07-027 (cit. on p. 39).

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of the 12th Language Resources and Evaluation Conference*, 4034–4043. https://www.aclweb.org/anthology/2020.lrec-1.497 (cit. on p. 89).

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357. https://doi.org/10.1037/0033-295X.113.2.327 (cit. on p. 14).

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, *116*(1), 207–219. https://doi.org/10.1037/a0014259 (cit. on p. 14).

Oaksford, M., & Chater, N. (1996). Rational explanation of the selection task. *Psychological Review*, *103*(2), 381–391. https://doi.org/10.1037/0033-295X.103.2.381 (cit. on p. 34).

O'Connor, E. (2015). *Comparative illusions at the syntax-semantics interface* [Doctoral dissertation, University of Southern California]. Retrieved August 1, 2023, from https://www.proquest.com/docview/2158857705/abstract/281BBE0BF7B04B6DPQ/1 (cit. on p. 105).

O'Donnell, T. J. (2015, August 28). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. The MIT Press. https://doi.org/10.7551/mitpress/9780262028844.001.0001

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*. Retrieved May 2, 2023, from https://www.frontiersin.org/articles/10.3389/frai.2022.777963 (cit. on pp. 30, 44, 47, 59).

Oh, B.-D., & Schuler, W. (2023a). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350. https://doi.org/10.1162/tacl_a_00548 (cit. on pp. 30, 44, 45, 47, 59, 60).

Oh, B.-D., & Schuler, W. (2023b, December). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 1915–1921). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.128 (cit. on pp. 30, 44, 45, 47).

Oh, B.-D., Yue, S., & Schuler, W. (2024, March). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2644–2663). Association for Computational Linguistics. Retrieved May 16, 2024, from https://aclanthology.org/2024.eacl-long.162 (cit. on p. 30).

Opper, M. (2015, December 1). Expectation propagation. In F. Krzakala, F. Ricci-Tersenghi, L. Zdeborova, R. Zecchina, E. W. Tramel, & L. F. Cugliandolo (Eds.), *Statistical Physics, Optimization, Inference, and Message-Passing Algorithms: Lecture Notes of the Les Houches School of Physics: Special Issue, October 2013* (pp. 263–292). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198743736.003.0009 (cit. on p. 22).

Paape, D., Vasishth, S., & von der Malsburg, T. (2020). Quadruplex Negatio Invertit? The On-Line Processing of Depth Charge Sentences. *Journal of Semantics*, *37*(4), 509–555. https://doi.org/10.1093/jos/ffaa009 (cit. on pp. 31, 32, 105, 106).

Paskin, M. A. (2001, December). Grammatical bigrams. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14 (NIPS 2001)* (pp. 91–97). MIT Press. https://proceedings.neurips.cc/paper/2001/hash/89885ff2c83a10305ee08bd507c1049c-Abstract.html (cit. on p. 87).

Perea, M., & Lupker, S. J. (2003). Does jugde activate COURT? Transposed-letter similarity effects in masked associative priming. *Memory & Cognition*, *31*(6), 829–841. https://doi.org/10.3758/BF03196438 (cit. on p. 71).

Phillips, C., Wagers, M. W., & Lau, E. F. (2011, January 1). Grammatical illusions and selective fallibility in real-time language comprehension. In *Experiments at the Interfaces* (pp. 147–180, Vol. 37). Brill. https://doi.org/10.1163/9781780523750_006 (cit. on p. 31).

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6 (cit. on p. 43).

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529. https://doi.org/10.1073/pnas.1012551108 (cit. on p. 1).

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R. P., & Cotterell, R. (2023). On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*, *11*, 1624–1642. https://doi.org/10.1162/tacl_a_00603 (cit. on p. 33).

Polyanskiy, Y., & Wu, Y. (2024, November 30). *Information theory: From coding to learning* (1st ed.). https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf (cit. on p. 27).

Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, *36*(6), 1389–1401. https://doi.org/10.1002/j.1538-7305.1957.tb01515.x (cit. on pp. 88, 191).

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. MIT Press. https://lccn.loc.gov/84002913 (cit. on p. 10).

R Core Team. (2021). *R: A language and environment for statistical computing*. Manual. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/ (cit. on p. 149).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf (cit. on p. 40).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (cit. on pp. 3, 40, 46, 49, 50, 74, 83, 165).

Rasmussen, N. E., & Schuler, W. (2018). Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*, *42*(S4), 1009–1042. https://doi.org/10.1111/cogs.12511 (cit. on pp. 12, 36, 44).

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. http://is.muni.cz/publication/884893/en (cit. on p. 184).

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 (cit. on p. 44).

Rényi, A. (1961). On measures of entropy and information. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (pp. 547–561, Vol. 1). https://projecteuclid.org/proceedings/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fourth-Berkeley-Symposium-on-Mathematical-Statistics-and/Chapter/On-Measures-of-Entropy-and-Information/bsmsp/1200512181 (cit. on p. 26).

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x (cit. on pp. 51, 149).

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *27*(2), 249–276. https://doi.org/10.1162/089120101750300526 (cit. on pp. 37, 39).

Roark, B., Bachrach, A., Cardenas, C., & Pallier, C. (2009). Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 324–333. https://www.aclweb.org/anthology/D09-1034 (cit. on pp. 30, 44, 47, 62).

Rosenkrantz, D. J., & Lewis, P. M. (1970). Deterministic left corner parsing. *11th Annual Symposium on Switching and Automata Theory (Swat 1970)*, 139–152. https://doi.org/10.1109/SWAT.1970.5 (cit. on p. 39).

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, *181*, 141–150. https://doi.org/10.1016/j.cognition.2018.08.018 (cit. on p. 1).

Salle, A., & Villavicencio, A. (2019). Why so down? The role of negative (and positive) pointwise mutual information in distributional semantics. (Cit. on p. 186).

Samson, E. W. (1953). Fundamental natural concepts of information theory. *ETC: A Review of General Semantics*, *10*, 283–297. Retrieved March 25, 2024, from https://www.jstor.org/stable/42581366 (cit. on p. 2).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. (Cit. on p. 89).

Sanz-Alonso, D. (2018). Importance sampling and necessary sample size: An information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, *6*(2), 867–879. https://doi.org/10.1137/16M1093549 (cit. on pp. 26, 46, 61).

Savage, L. J. (1972). *The foundations of statistics* (2d rev. ed.). Dover Publications. https://lccn.loc.gov/79188245 (cit. on p. 28).

Schotter, E. R., Bicknell, K., Howard, I., Levy, R., & Rayner, K. (2014). Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition*, *131*(1), 1–27. https://doi.org/10.1016/j.cognition.2013.11.018 (cit. on p. 32).

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population marginal means in the linear model: An alternative to least squares means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031 (cit. on p. 171).

Shain, C. (2021). CDRNN: Discovering complex dynamics in human language processing. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3718–3734. https://doi.org/10.18653/v1/2021.acl-long.288 (cit. on p. 76).

Shain, C. (2024). Word Frequency and Predictability Dissociate in Naturalistic Reading. *Open Mind*, *8*, 177–201. https://doi.org/10.1162/opmi_a_00119 (cit. on p. 71).

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 (cit. on pp. 14, 47).

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022, November 25). *Large-scale evidence for logarithmic effects of word predictability on reading time*. https://doi.org/10.31234/osf.io/4hyna (cit. on pp. 46, 60, 153, 154).

Shain, C., & Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d18-1288 (cit. on p. 76).

Shain, C., & Schuler, W. (2021). Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, *215*, 104735. https://doi.org/10.1016/j.cognition.2021.104735 (cit. on p. 154).

Shain, C., & Schuler, W. (2022, September 24). *A deep learning approach to analyzing continuous-time systems*. arXiv: 2209.12128 [cs, stat]. (Cit. on pp. 46, 154).

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x (cit. on pp. 2, 3).

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50–64. https://doi.org/10.1002/j.1538-7305.1951.tb01366.x (cit. on p. 3).

Shen, Y., Tan, S., Sordoni, A., & Courville, A. C. (2019). Ordered neurons: Integrating tree structures into recurrent neural networks. *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. https://openreview.net/forum?id=B1l6qiR5F7 (cit. on p. 90).

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243 (cit. on p. 10).

Smith, N. J., & Levy, R. (2008a). 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 (cit. on pp. 14, 37, 44, 45, 47, 57, 149, 158).

Smith, N. J., & Levy, R. (2008b). Probabilistic prediction and the continuity of language comprehension. *9th Conference on Conceptual Structure, Discourse, and Language (CSDL9)*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1295346 (cit. on pp. 14, 44, 47).

Smith, N. J., & Levy, R. (2011). Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/69s3541f (cit. on p. 47).

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 (cit. on pp. 12, 14, 37, 38, 44, 45, 47, 51, 56, 57, 59, 149–151, 153, 158).

Sonderegger, M. (2023, June 6). *Regression modeling for linguistic data*. The MIT Press. https://lccn.loc.gov/2022026920 (cit. on pp. 78, 171).

Sóskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, *84*, 101017. https://doi.org/10.1016/j.wocn.2020.101017 (cit. on pp. 150, 151).

Stabler, E. P. (2013). Two models of minimalist, incremental syntactic analysis. *Topics in Cognitive Science*, *5*(3), 611–633. https://doi.org/10.1111/tops.12031 (cit. on p. 39).

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86. https://doi.org/10.1016/j.cognition.2010.04.002 (cit. on p. 156).

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics*, *21*(2), 165–201. https://www.aclweb.org/anthology/J95-2002 (cit. on pp. 37, 39).

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. *Interspeech 2012*, 194–197. https://doi.org/10.21437/Interspeech.2012-65 (cit. on p. 3).

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science (New York, N.Y.)*, *268*(5217), 1632–1634. https://doi.org/10.1126/science.7777863 (cit. on p. 37).

Taylor, W. L. (1953). 'Cloze procedure': A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401 (cit. on p. 47).

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. (Cit. on pp. 93, 188).

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023, July 19). *Llama 2: Open foundation and fine-tuned chat models*. arXiv: 2307.09288 [cs]. (Cit. on pp. 3, 74, 165).

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, *47*(1), 69–90. https://doi.org/10.1006/jmla.2001.2836 (cit. on p. 156).

Van der Mude, A., & Walker, A. (1978). On the inference of stochastic regular grammars. *Information and Control*, *38*(3), 310–329. https://doi.org/10.1016/S0019-9958(78)90106-7 (cit. on p. 87).

Vani, P., Wilcox, E. G., & Levy, R. (2021). Using the interpolated maze task to assess incremental processing in English relative clauses. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Retrieved March 9, 2023, from https://escholarship.org/uc/item/3x34x7dz (cit. on p. 156).

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 (cit. on pp. 14, 44, 48, 59).

Vasishth, S. (2006). On the proper treatment of spillover in real-time reading studies: Consequences for psycholinguistic theories. *Proceedings of the International Conference on Linguistic Evidence*, 96–100 (cit. on p. 158).

Vasishth, S., & Engelmann, F. (2021, October 31). *Sentence comprehension as a cognitive process: A computational approach* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781316459560 (cit. on pp. 37, 48).

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, *23*(11), 968–982. https://doi.org/10.1016/j.tics.2019.09.003 (cit. on p. 48).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 5998–6008). https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (cit. on pp. 3, 45).

Vieira, T., & Eisner, J. (2017). Learning to prune: Exploring the frontier of fast and accurate parsing. *Transactions of the Association for Computational Linguistics*, *5*, 263–278. https://doi.org/10.1162/tacl_a_00060 (cit. on p. 62).

Wainwright, M. J., & Jordan, M. I. (2007). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, *1*(1–2), 1–305. https://doi.org/10.1561/2200000001 (cit. on p. 22).

Wald, A. (1947). Foundations of a general theory of sequential decision functions. *Econometrica*, *15*(4), 279–313. https://doi.org/10.2307/1905331 (cit. on p. 14).

Wang, B., & Komatsuzaki, A. (2021, May). GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax (cit. on p. 49).

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 (cit. on pp. 32, 105, 106).

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 (cit. on p. 105).

Wieling, M., Tomaschek, F., Arnold, D., Tiede, M., Bröker, F., Thiele, S., Wood, S. N., & Baayen, R. H. (2016). Investigating dialectal differences using articulography. *Journal of Phonetics*, *59*, 122–143. https://doi.org/10.1016/j.wocn.2016.09.004 (cit. on pp. 150, 151).

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 (cit. on p. 14).

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/ (cit. on pp. 37, 38, 44–46, 51, 56, 57, 59, 149–151, 158, 160).

Wilcox, E. G., Pimentel, T., Meister, C., & Cotterell, R. (2024). An information-theoretic analysis of targeted regressions during reading. *Cognition*, *249*, 105765. https://doi.org/10.1016/j.cognition.2024.105765 (cit. on p. 107).

Wilson, K., & Carroll, J. B. (1954). Applications of entropy measures to problems of sequential structure. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguistics* (pp. 103–110, Vol. Psycholinguistics: A survey of theory and research problems). Indiana University Press. https://doi.org/10.1037/h0063655 (cit. on p. 33).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6 (cit. on pp. 74, 90, 146, 165, 192).

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 95–114. https://doi.org/10.1111/1467-9868.00374 (cit. on p. 150).

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686. https://doi.org/10.1198/016214504000000980 (cit. on pp. 44, 148).

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x (cit. on p. 56).

Wood, S. N. (2017, May). *Generalized additive models*. Chapman; Hall/CRC. https://doi.org/10.1201/9781315370279 (cit. on pp. 44, 148–152).

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, *111*(516), 1548–1563. https://doi.org/10.1080/01621459.2016.1180986 (cit. on pp. 37, 51, 149, 153).

Wu, S., Bachrach, A., Cardenas, C., & Schuler, W. (2010). Complexity Metrics in an Incremental Right-Corner Parser. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1189–1198. Retrieved May 31, 2022, from https://aclanthology.org/P10-1121 (cit. on p. 34).

Wu, Z., Chen, Y., Kao, B., & Liu, Q. (2020). Perturbed masking: Parameter-free probing for analyzing and interpreting BERT. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4166–4176. https://doi.org/10.18653/v1/2020.acl-main.383 (cit. on pp. viii, 97–99).

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, *108*(1), 40–55. https://doi.org/10.1016/j.bandl.2008.10.002 (cit. on p. 32).

Xu, W., Chon, J., Liu, T., & Futrell, R. (2023, December). The linearity of the effect of surprisal on reading times across languages. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 15711–15721). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1052 (cit. on pp. 12, 14).

Yang, K., & Deng, J. (2020). Strongly incremental constituency parsing with graph neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 21687–21698, Vol. 33). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/f7177163c833dff4b38fc8d2872f1ec6-Paper.pdf (cit. on p. 39).

Yang, S., Jiang, Y., Han, W., & Tu, K. (2020). Second-order unsupervised neural dependency parsing. *Proceedings of the 28th International Conference on Computational Linguistics*, 3911–3924. https://doi.org/10.18653/v1/2020.coling-main.347 (cit. on p. 89).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019, December). XLNet: Generalized autoregressive pretraining for language understanding. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 5754–5764, Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html (cit. on pp. 87, 89).

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*(5), 444–466. Retrieved March 10, 2023, from https://www.jstor.org/stable/985230 (cit. on p. 36).

Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction* [Doctoral dissertation, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science]. (Cit. on p. 87).

Zehr, J., & Schwarz, F. (2018, March 15). *PennController for Internet Based Experiments (IBEX)*. https://doi.org/10.17605/OSF.IO/MD832 (cit. on p. 72).

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022, June 21). *OPT: Open Pre-trained Transformer language models*. arXiv: 2205.01068 [cs]. (Cit. on pp. 59, 74, 165).

Zhang, T., & Hashimoto, T. (2021). On the inductive bias of masked language modeling: From statistical to syntactic dependencies. *Proceedings of the 2021 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies. https://doi.org/10.18653/v1/2021.naacl-main.404 (cit. on p. 97).

Zhang, Y., & Clark, S. (2008). A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 562–571. https://aclanthology.org/D08-1059 (cit. on p. 62).

Zhang, Y., Gibson, E., & Davis, F. (2023, December). Can language models be tricked by language illusions? Easier with syntax, harder with semantics. In J. Jiang, D. Reitter, & S. Deng (Eds.), *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 1–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-1.1 (cit. on p. 106).

Zhang, Y., Kauf, C., Levy, R. P., & Gibson, E. (2024, May 23). *Comparative illusions are evidence of rational inference in language comprehension*. https://doi.org/10.31234/osf.io/efr3q (cit. on p. 105).

Zhang, Y., Ryskin, R., & Gibson, E. (2023). A noisy-channel approach to depth-charge illusions. *Cognition*, *232*, 105346. https://doi.org/10.1016/j.cognition.2022.105346 (cit. on p. 105).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2023, November 24). *A survey of large language models*. arXiv: 2303.18223 [cs]. (Cit. on p. 3).

Appendices

A

Supplemental material for chapter 2

A.1 Runtime variance of guessing without replacement

In §2.2.3.2 in the main text we gave an expression for $\mathrm{Var}[N]$, the variance in the number of draws needed in guessing without replacement (eq. 2.7). Here we give the derivation of that expression. From general identities about covariance, we have the following.

$$\begin{aligned}\mathrm{Var}[N] = \mathrm{Var}[N-1] = \mathrm{Var}[\textstyle\sum_i X_i] &= \sum_{i,j} \mathrm{Cov}[X_i, X_j] \\ &= \sum_{i,j} \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]\end{aligned}$$

In each element of this sum, the first expectation term $\mathbb{E}[X_i X_j]$ is simply the probability that items i and j both come before the target, 0. There are two cases to consider. If $i = j$ this simplifies to $\mathbb{E}[X_i^2] = \mathbb{E}[X_i] = \Pr(i \prec 0) = \frac{u_i}{u_i + u_0}$. Otherwise $i \neq j$, and we have

$$\begin{aligned}\mathbb{E}[X_i X_j] &= \Pr(i \prec 0, j \prec 0) \\ &= \Pr(i \prec j \prec 0) + \Pr(j \prec i \prec 0)\end{aligned} \tag{A.1}$$

where, by an argument similar to that given in the proof of proposition 2.1,

$$\begin{aligned}\Pr(i \prec j \prec 0) = \Pr(i \prec j \wedge j \prec 0) &= \Pr(i \prec j \mid j \prec 0)\Pr(j \prec 0) \\ &= \Pr(i \prec (j \vee 0))\Pr(j \prec 0) \\ &= \frac{u_i}{u_i + u_j + u_0}\frac{u_j}{u_j + u_0}\end{aligned} \tag{A.2}$$

and likewise for $\Pr(j \prec i \prec 0)$.

So,

$$\begin{aligned}
\mathrm{Var}[N] &= \sum_{i,j} \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\
&= \sum_{i} \mathbb{E}[X_i] - (\mathbb{E}[X_i])^2 + \sum_{i \neq j} \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\
&= \sum_{i} \left[\frac{u_i}{u_i + u_0} - \left(\frac{u_i}{u_i + u_0} \right)^2 \right] + \sum_{i \neq j} \left[\left(\frac{u_i}{u_i + u_j + u_0} \frac{u_j}{u_j + u_0} + \frac{u_j}{u_i + u_j + u_0} \frac{u_i}{u_i + u_0} \right) - \frac{u_i}{u_i + u_0} \frac{u_j}{u_j + u_0} \right]
\end{aligned} \tag{A.3}$$

This is the expression for variance given in eq. 2.7, and plotted in fig. 2.1 for Pareto-distributed weights.

A.2 Language model surprisals

For our surprisal estimates, we used the pretrained models from Huggingface Transformers (Wolf et al., 2020) identified by the following model IDs: `transfo-xl-wt103`, `gpt2`, `gpt2-large`, `gpt2-xl`, `EleutherAI/gpt-neo-2.7B`, and `EleutherAI/gpt-j-6B`. For the proprietary GPT-3 models, we used log probabilities provided via the OpenAI API for the original GPT-3 base models with model IDs `davinci`, `curie`, and `ada` the (accessed with free trial account, March, 2022). For the n-gram and LSTM models, as well as unigram frequency predictors, we use data made available in Boyce (2020, July 28/2022). Code we used for retrieving all surprisal estimates we use will be released with supplemental material.

For each of the Transformer-based LMs, we obtain surprisal estimates with different amounts of context: In addition to the *maximum* context and *within-sentence* context amounts described in the main text, we also computed surprisals using *80 words* of context for the Huggingface models. These surprisals were estimated for each token using a sliding window of at most 80 tokens immediately preceding it within the story.

Figure A.1 is the full version of fig. 2.2, giving the GAM fits for the overall effect of surprisal on reading time, for surprisals estimated by each of language models we use and each of the context amounts.

Tokenization Because of tokenization differences between the reading time corpus and the language models, some words seen by participants as single units correspond to multiple tokens according to the tokenizers used by the language models. In order to avoid making unnecessary assumptions, we discard words where the tokenization is different (excluding punctuation and whitespace differences). Because the different language models use different tokenization schemes,

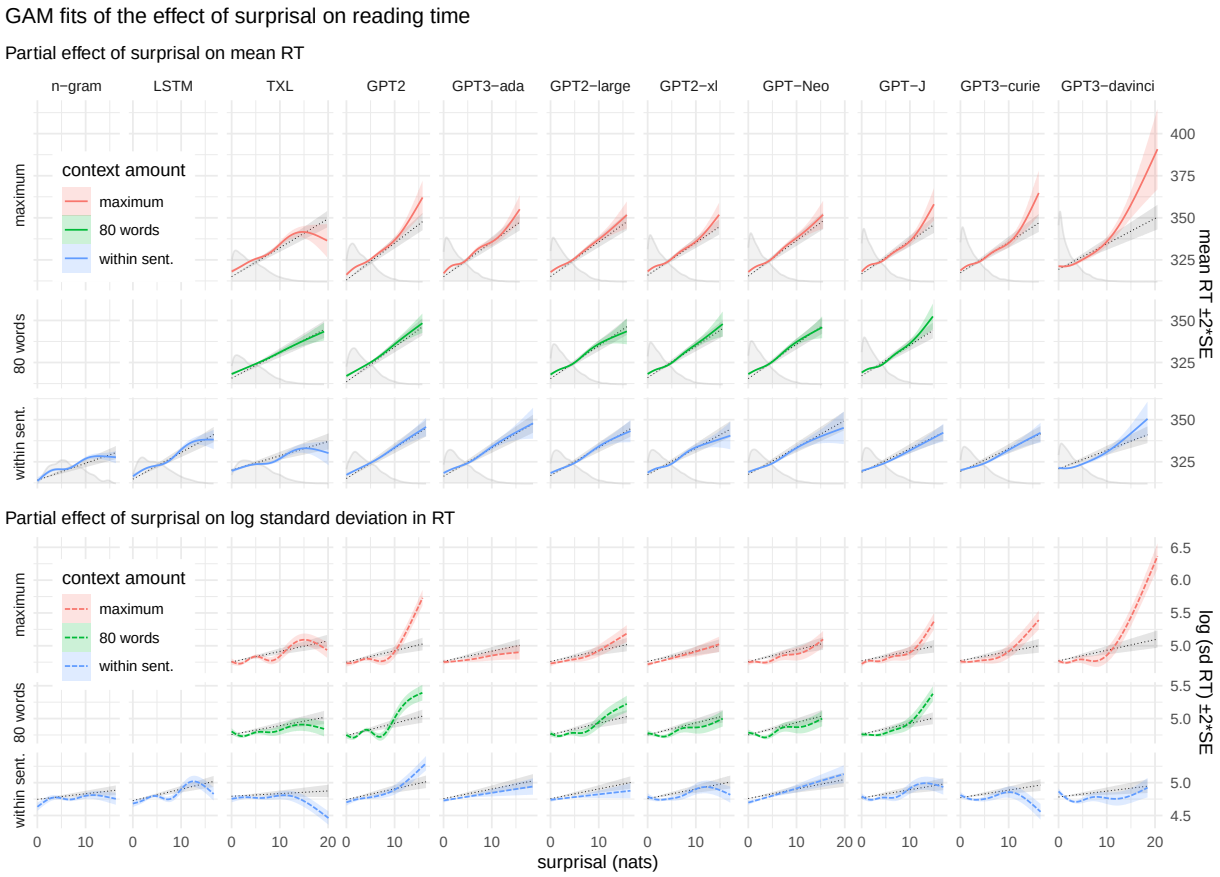


Figure A.1: Plots of all GAM models repeated from fig. 2.2, with the addition of select LM models with 80 previous words of context (middle row, green) as a middle-ground between maximum context and within sentence context.

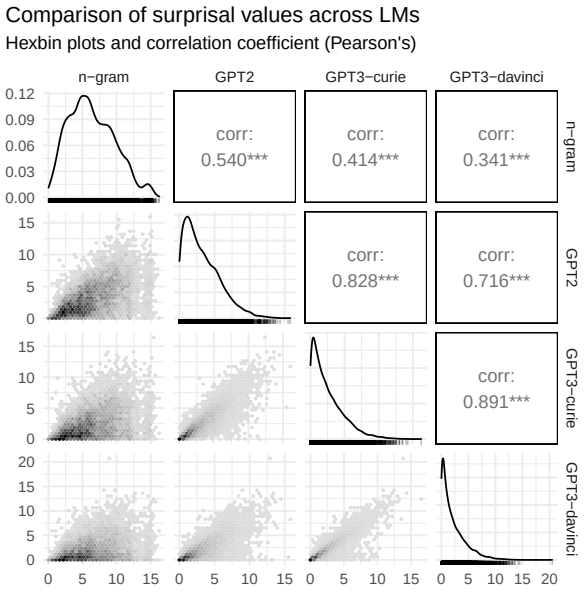


Figure A.2: Comparison of surprisal estimates from a selection of language models, by item in the corpus, with density and rug plots for each LM on the diagonal. Pearson's correlation coefficients for each of the pairs are given in the upper right.

the set of corpus tokens we use differs across language models, though not substantially.[1]

We do not estimate surprisal for the first word in each text (or sentence, for LMs using only within-sentence context), and so these words are removed before fitting the models. Similarly, words immediately following an excluded word are also excluded since their previous-word surprisal predictor (included to control for spillover) is undefined.

Comparison of LM surprisals Figure A.2 gives comparison of selected language models' surprisals against each other, by item in the corpus. We can see that as the language models get lower mean surprisal, not all words' surprisal is lowered proportionally. Also, as is clear from the density plots, at the higher end of surprisal, there is very little data, especially for the better language models. Given that it is in the high surprisal region that the predicted reading times according to the nonlinear GAMs we fit differ most from the predictions of the linear control, it is crucial to have data with constructions with high surprisal, something which is increasingly difficult with lower-perplexity language models.

A.3 Generalized additive models

Generalized additive models (GAMs; Wood, 2004, 2017) are a family of statistical models which allow nonlinear functions to be captured as linear combinations of basis functions. GAMs are

[1] After removing words with different tokenizations, 91% of tokens remain for the n-gram and LSTM, 80% for Transformer-XL, and 78% for the GPT models.

a nonlinear generalization of generalized linear models, and as such similarly allow the use of different response distributions and linking functions. For our purposes, a GAM allows us to fit a regression of the form

$$\text{Time}(w_n) = f_\theta(s(w_n)) \tag{A.4}$$

where the function f_θ is the linking function (as in previous literature since Smith & Levy, 2008a). GAMs are fit using penalized regression, of the form,

$$\arg\max_\theta \{\text{likelihood}(f_\theta) - \lambda J(f_\theta)\} \tag{A.5}$$

where the 'wiggliness' penalty functional J is specified so that $J(f_\theta) = 0$ if f_θ is linear, and, crucially, wiggliness is controlled by a parameter λ, which controls the trade-off between smoothness and fit to the data. This parameter itself may fit by cross validation, so the resulting regression model will be only be as nonlinear as necessary.

For our GAMs, we use the implementation provided by `gam` in `mgcv` 1.8-40 using R 4.2.1 (Wood, 2017; R Core Team, 2021). All GAMs we report in the main text were fit using with the default restricted maximum likelihood (REML) method for smoothing parameter estimation. Additional models given in this appendix that have a constant-variance assumption were fit using the more efficient `mgcv::bam` routine, and fast REML (fREML) for smoothing parameter estimation for computational efficiency.

A.3.1 Nonlinear GAM details

Formula A.1 gives the `mgcv` formulæ we use for the GAM fits of the nonlinear effect of surprisal on reading time. We fit Gaussian location-scale models (Rigby & Stasinopoulos, 2005; Wood et al., 2016), which lets us specify smooth predictors for the mean and standard deviation separately (with `family=gaulss()`). The LHS of the first formula specifies the response, while the RHS specifies the structure of the linear predictor for mean RT. The second formula is one-sided, and specifies the structure of the linear predictor for standard deviation. In all our models, we use the default links: identity link for the mean, and a log-shift link for the standard deviation (so the relationship between the linear predictor and the standard deviation is $\eta = \log(\sigma + b)$, with parameter $b = 0.01$).

For the predictors of mean, following Smith and Levy (2013), Goodkind and Bicknell (2018), and Wilcox et al. (2020), we include a nonlinear term for the main effect of surprisal, and also include a tensor product term for the interaction between log-frequency and word length (orthographic) of the current word. Also following this previous literature, we include predictors likewise for the effect of the previous word on current reading time, to help control for spillover effect (see discussion in, e.g., Smith & Levy, 2013). We additionally include subject-specific terms (using `bs='fs'` in `mgcv` to use the factor-smooth interaction basis) to allow for by-subject nonlinear effects

```
list(# mean
     RT ~ s(surp, bs='tp', k=6) + s(subj, surp, bs='fs', m=1, k=6) +
          te(freq, len) +
          s(prev_surp, bs='tp', k=6) + s(subj, prev_surp, bs='fs', m=1, k=6) +
          te(prev_freq, prev_len),
     # standard deviation
     ~ s(surp, bs='tp', k=6) + s(subj, surp, bs='fs', m=1, k=6) +
       te(freq, len))
```

Formula A.1: The `mgcv` formulæ used for the **nonlinear GAM** fits. RT is predicted as a nonlinear function of surprisal, controlling for nonlinear by-subject effects, and interactions between frequency and word length. The mean formula also includes similar predictors previous word as well as current, to control for spillover effects.

on reading time, to avoid the assumption of linearity, rather than just random slopes and or intercepts as in Goodkind and Bicknell (2018). Unlike by-subject random smooths, which fit potentially nonlinear effects independently for each subject (or separate by-subject models, as used by Smith & Levy, 2013, for their experiment with eye-tracking data only), including the subject predictor as a factor-smooth interaction allows us to control for potentially different nonlinear effects of each participant (and random intercept) while sharing the same smoothing parameter, as is appropriate for by-subject random smooths (Wood, 2017, §7.7.4).[2]

Basis and Order of Penalty Term Since we are particularly interested in the shape of this curve in the high-surprisal region, where there is the least data, we choose not to use cubic regression splines (unlike Goodkind & Bicknell, 2018; Wilcox et al., 2020), for which knot locations are by default chosen by quantile. Instead we use thin-plate regression splines (TPRS; Wood, 2003), avoiding the problem of knot placement. Using TPRS results in evenly distributed knots.

We set the order of the penalty functional to 1 (`m=1`) in the factor smooth, which penalizes towards a slope of zero (flat line). This results in penalizing deviation from the global effect, limiting the wiggliness per speaker, suitable for these by-speaker nonlinear effects (cf. Wieling et al., 2016). While this choice is principled, changing it does not affect our qualitative results. Our choice to set the penalty term `m=1` in the factor smooth interaction terms is motivated by the fact that the default `m=2` would allow more wiggliness per speaker smooths, and could lower our power to detect the population-level positive effect. In preliminary testing with `m=2`, the qualitative results were unchanged. We note however that the confidence intervals on the resulting main smooths were somewhat wider than the results using `m=1` which we report, and to this extent, the choice may be somewhat anticonservative. Since we are interested in the overall effect, the choice to set a stronger

[2]The factor smooth interaction basis we use fits a nonlinear random effect for each subject (with a TPRS basis and basis dimension $k = 10$, by default). The key point is that using factor smooth rather than random slopes, not which exact factor smooths used, which matters less (as explored in Sóskuthy, 2021).

penalty on the factor smooths is warranted, and follows previous literature on using similar GAMs (e.g. Wieling et al., 2016; Sóskuthy, 2021), though we are the first to introduce it to this application.

Restricting maximum wiggliness We must choose a value for basis dimension for the main smooth term, k. This parameter effectively controls the maximum degrees of freedom of the curve, with a higher values allowing a potentially very wiggly curve to be fit, while at the lower value, the curve would be forced toward the null space (linear). The arbitrary default in `mgcv` is $k = 10$. Some previous work chooses a large number for the basis dimension (such as $k = 20$, in e.g., Smith & Levy, 2013; Wilcox et al., 2020) and allows the smoothing parameter to be fit according to the data, resulting in only as smooth a curve as is necessary. Instead, we set $k = 6$, effectively allowing a maximum of 5 degrees of freedom ($k - 1$, because one degree is lost to the identifiability constraint). The result is nonlinear effects which are restricted to simpler curves. We limit the basis dimension since we are in particular interested in the rather simple question: given a few degrees of freedom, whether the GAM will use them to bend the curve, or not. In preliminary experimentation, increasing the basis dimension leads to local nonlinearities which obscure the global pattern somewhat, but don't change the qualitative interpretation.

A.3.2 Linear control GAM details

As described in the main text (§2.4.3), for each language model and context amount, in addition to the GAM fit using formula A.1 (the nonlinear GAM), we also fit a GAM using formula A.2 (the linear control GAM), where the effects of surprisal on reading time mean and likewise on variance are assumed to be linear, but otherwise the model is the same. For this linear control, the global nonlinear terms of surprisal and previous word surprisal are replaced with linear parametric terms, and the factor-smooth subject terms are replaced with linear random effects (via the basis `bs='re'`). One caveat is that this model specification includes the additional assumption that the random slopes and intercepts are independent, which is not assumed in the case of the nonlinear model.[3] We leave the tensor product terms for the interactions between frequency and length the same for maximum similarity between the two. The interpretation of the linear control models is as a baseline to which the nonlinear models would converge if the true effect of surprisal on reading time were perfectly linear.

A.3.3 Significance of superlinearity

We are interested in whether an assumption of linearity is justified to model the effect of surprisal on processing difficulty, or if a nonlinear fit is necessary. One way to specifically test whether a smooth term may be replaced with a linear parametric term in a GAM is to explicitly separate the

[3] A smooth term `s(x, g, bs='re')` for the random effect of variable `x` with grouping factor `g` encode a random effect of `x` for each level of `g`, but not by-group means. Adding random intercepts in separately, with an additional term `s(g, bs='re')` will encode an assumption that all slopes and intercepts are independent (see Wood, 2017, §3.5.2)

```
list(# mean
     RT ~ surp + s(subj, bs='re') + s(surp, subj, bs='re') + te(freq, len) +
          prev_surp + s(prev_surp, subj, bs='re') + te(prev_freq, prev_len),
     # standard deviation
     ~ surp + s(subj, bs='re') + s(surp, subj, bs='re') + te(freq, len))
```

Formula A.2: The formulæ used for **linear control GAM** fits. The interpretation is effectively the same as that of formula A.1, except that the fit effect of surprisal on mean/variance in reading time is forced to be linear.

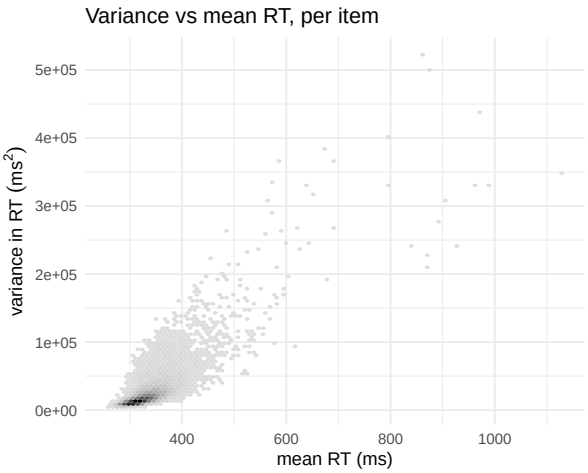


Figure A.3: Variance in self-paced reading time versus mean, by item in the Natural Stories corpus. Variance increases with mean.

basis for the penalty range space from the basis for the null space when parametrizing the smooth, effectively allowing one to ask the question "is this curve significantly nonlinear?" This technique can be accomplished in `mgcv` with thin-plate regression splines by setting the smooth up without a null space basis, and including a parametric term (as described in Wood, 2017, §6.12.3).[4] For our purposes, we are interested in the shape of the nonlinear fit (namely, whether it is superlinear), not simply whether it is significant. Nonetheless, we experimented with using this technique to get a p-value testing whether the nonlinear components were required. Unsurprisingly, we found across models that the nonlinear components were significant, though not in an illuminating way: even for the worst LMs and the most qualitatively linear fits, there are small but statistically significant nonlinearities. For this reason this technique is not a useful way to quantify nonlinearity.

A.3.4 Nonconstant variance of data

Most modelling of the relationship between surprisal and reading time, both using generalized linear mixed models and GAMs, has used the default Gaussian distribution for the dependent

[4] However, as Wood notes, this technique is generally unnecessary when the smoothing parameter is efficient to fit, as a smooth would be automatically shrunk to linear if the data merit it.

variable, with identity linking function. The primay exceptions are Hofmann et al. (2022), who use a gamma family with the default logarithmic linking function, and Smith and Levy (2013), who also mention that their results were robust to switching from Gaussian to a heavy-tailed (gamma) family. The choice of dependent variable distribution and linking function for models of reading time data in general is explored in detail in Lo and Andrews (2015), who point out that RTs are better modelled by waiting time distributions such as the gamma or inverse-Gaussian.

Looking at the reading time data we use empirically, before fitting any models, it is clear that the variance in reading time is not constant across mean reading time values, as illustrated in fig. A.3. This already suggests that the assumption of constant error variance implicit in using least squares estimation (constant Gaussian distributed error) is not warranted. This lack of constant variance is a known feature of reaction time data, and motivates the use of a response distribution that is better matched to these data (see Lo & Andrews, 2015, for detailed discussion). In fitting Gaussian scale-location models (Wood et al., 2016) where variance is allowed to vary as a smooth function of the predictors, we can effectively probe the correspondance between mean and variance. In our results, the similarity between the fitted curves for mean and variance (figs. 2.2 and A.1) suggest that use of a member of the exponential family for which variance increases smoothly with the predictor value is indeed justified (for example, gamma or inverse-Gaussian distributions). An expansion of the current study using such models is material for future work.

A.3.5 Relationship between mean and variance

The GAMs we fit did not assume any particular relationship between RT and variance in RT. Yet, comparing the nonlinear GAM's mean and variance fits for a given LM in fig. 2.2, it is clear that these two curves are generally similar to each other in shape. The similarity between these fitted curves may justify the use of statistical models where variance is *assumed* to be a fixed increasing function of the predicted mean.[5] Making this assumption *a priori*, rather than fitting that relationship simultaneously for mean and variance, as we did, would have the benefit of making the models much less computationally costly to fit. We leave to future work the exploration of models with variance that scales parametrically with the mean.

A.4 Comparison with Shain et al. (2022)

Shain et al. (2022) present a meticulous and large-scale study of the relationship between surprisal and processing difficulty, using multiple datasets (including Natural Stories) and reading modalities

[5] A model with a gamma-distributed response (as used by, e.g. Hofmann et al., 2022) has this property. This is likewise true for inverse Gaussian, or even log-normal models, though the specific assumption is different in each case (see Lo and Andrews, 2015 for a discussion of these choices for modelling reading-time data with generalized linear models). An assumption of a inverse Gaussian or gamma distribution would also potentially be a principled choice for an underlying process involving sampling, given these distributions model waiting time.

(eye-tracking and Maze task data, in addition to self-paced reading) and using surprisal estimates from multiple language models (including a 5-gram model, and GPT-2, GPT-J, and GPT-3 Davinci as well as a PCFG model and cloze probabilities). Unlike the current study and much previous literature, Shain et al. (2022) do not use GAMs, but instead make use of continuous-time deconvolutional regressive neural networks (CDRNNs; Shain & Schuler, 2021, 2022), a new modelling technique which describes the influence of predictors in terms of overlapping additive impulse response functions in continuous time. This technique also allows modelling of the effect of predictors on all parameters of the response distribution (not just the mean), with full nonlinear random effects.

While their study and the empirical component of our study both target the shape of the linking function, and use surprisal estimates from some of the same pre-trained language models, the differing analytical models make it difficult to compare results directly. Still, for the Natural Stories dataset (which, of the datasets they include, has the largest number of observations, and is also the dataset we use), they report qualitative confirmation of the superlinear relationship we observe between surprisal and self-paced reading time. Namely, their results for this data show curves that increase superlinearly with surprisal for the larger LMs, with superlinear models tending also to show stronger performance (larger psychometric predictive power). However, they do not find such a trend in the other datasets and modalities, and find that *overall* (when aggregating across all and datasets and modalities) the larger models GPT-3 and GPT-J perform worse as psychometric models than GPT-2, especially if the linking function is constrained to be linear[6]. Their overall conclusion is that empirical evidence favors a linear relationship.

As discussed in the main text, we believe our choice of the Natural Stories dataset is well-motivated, given the design of the corpus, a well as the large number of participants, which allows us to better control for a large amount of potential variation between individuals. However, the difference between the results on this dataset, which do show superlinearity (in both our study and theirs), and those on the other datasets and modalities in their study, which do not, complicates the picture. It is also worth noting (as Shain et al., 2022 do) that if the uncertainty interval covers an a superlinear function, it is not possible to falsify the hypothesis of superlinearity in favor of a linear linking function. This observation leads back to our fundamental motivating question: What predictions do algorithmic theories of processing make about the relationship between surprisal and processing difficulty? In this work we have argued that the only algorithms we know of which naturally scale in surprisal predict a superlinear linking function. The tensions between this prediction and the results of Shain et al. (2022) motivate further study from both empirical

[6]With an unconstrained (nonlinear) linking function this is less clear: GPT-J does not underperform GPT-2, but GPT-3 does. However, we note this trend reverses in their results when considering just the self-paced reading datasets in their study. In fact fully nonlinear GPT-3 and GPT-J perform better than GPT-2 for self-paced reading data from both available corpora (Natural Stories and Brown).

and theoretical directions.

A.5 Surprisal explorer

To facilitate exploration of the words of the corpus in full context, with language model surprisal estimates and reading time annotations, we provide an interactive utility in the repository for this paper: github.com/mcqll/plausibility-sampling-processing/.

A.6 Effect of highest surprisal words

The difference between a linear and superlinear linking function is naturally most appreciable in the high end of the surprisal range. However, for low-perplexity LMs, the vast majority of words in the corpus are relatively low surprisal, as can be seen in the highly skewed density plots of surprisal values (plotted above in figs. 2.2 and A.1, and compared across LMs in fig. A.2). This is to be expected for any corpus of fluent text, and remains true of the Natural Stories corpus, despite its being designed to contain rare and marked constructions. Since this skew is especially pronounced for the lowest-perplexity LMs, the models for which we see the most superlinearity are also the models for which we have the smallest amount of data in the high end of the surprisal range. To understand how the particular words in this region of the surprisal range affect our results, in this appendix we take a detailed look at the highest-surprisal words according to GPT-3 Davinci—the lowest-perplexity of the LMs we use, and the one for which the relationship with reading time is the most superlinear. Then we assess their contribution to this superlinearity, by re-fitting the GAM without these words.

A.6.1 Highest surprisal words

For GPT-3 Davinci, the top 40% of the surprisal range (surprisal $>$ 12 nats) contains only about 0.3% of the words in the corpus. Table A.1 gives each of these words, in order of decreasing surprisal, with part-of-speech tag and dependency label (provided with the Natural Stories corpus; see Futrell et al., 2021, §2.3).

Inspecting each of the words on this list in context, it is possible to identify intuitive reasons why it is plausible that they would be high-surprisal for humans, yet it is not possible to put them into one common category. Most are examples of unusual grammatical constructions. The notable exceptions are items 1, 2, and 4: The highest surprisal word (item 1) seems to be the result of a typo or at least unconventional usage ("**US**" rather than "the US"). Also high on the list are two numbers which are dates written out longform (items 2 and 4 in the table), where presumably numerals would be more expected. Of the remaining items on the list, many are examples of the kinds of marked syntactic constructions that Natural Stories is designed to contain. For example, four are words at critical regions in object-extracted relative clauses (ORCs). Two are on the verb

	word in context	GPT-3 D. surprisal	POS	dep. label	story #	word #	mean RT
1	...in military programs **US** conducted in the...	20.51	NNP	nsubj	8	836	413.83
2	...mania in February **sixteen** thirty-seven, tulip...	18.11	NN	compound	9	38	862.11
3	...His brother had **blatantly** peeked and even...	15.50	RB	advmod	2	748	391.25
4	...movie brought the **nineteen** forty-seven incident...	14.73	CD	nummod	8	404	361.44
5	...names, such as **even** 'Admiral of Admirals' and...	14.49	RB	advmod	9	343	460.22
6	...classic that many **publishing** houses continue...	14.22	NN	compound	9	884	317.26
7	...well which seems **puzzling** at first, but the reason...	13.84	JJ	xcomp	1	137	375.21
8	...the little bird **guarded** by the owl peeped out,...	13.62	VBN	acl	4	904	326.08
9	...who Abby still **strained** to remain upset with, ...	13.23	VBN	dep	6	772	374.57
10	...sight, and then **folding** his wings together, he...	13.12	VBG	dep	4	479	359.07
11	...was called and **though** they understood the birds...	13.10	IN	mark	4	37	366.11
12	...were supposed to **slowly** wait to be called, I...	12.72	RB	advmod	5	448	346.46
13	...little girl no one **sheltered** from the gelid air...	12.59	VBD	acl:relcl	3	28	439.62
14	...markets, which **merchants** used to sell and buy...	12.38	NNS	nsubj	9	544	330.08
15	...vocalizations, which **motor** tics typically precede,...	12.27	NN	compound	10	315	389.39
16	...September, and thus **actual** purchases occurred...	12.13	JJ	amod	9	488	388.97
17	...the boar? By the **handsome** reward many felt...	12.10	JJ	amod	1	346	355.48
18	...who they knew looked **dirt** poor and helpless....	12.04	RB	advmod	3	978	368.29
19	...The Dutch Golden Age **growers** named their...	12.02	NNS	nsubj	9	297	387.70

Table A.1: All 19 words in the corpus with GPT-3 Davinci surprisal > 12, with surrounding context, mean RT, part of speech tag, and dependency label annotations from the parses provided with the corpus.

(item 13: "little girl [CP ∅ no one **sheltered** ...]" and 9: " Mom, [CP who Abby still **strained** to ...]"), and the other two the onset of the subject NP (item 14: "markets, [CP which **merchants** used ...]", and 15: "vocalizations [CP which **motor** tics ...]").[7] Item 7 is at the critical region of a garden path sentence: "It shows a sinister looking boar's head sitting on top of a well [which seems **puzzling** at first]"—the word "**puzzling**" disambiguates attachment ambiguity for the relative clause, in favor of the matrix CP as the subject, rather than the local NP "well". Item 19 is another where temporary ambiguity is resolved in favor of the less-likely alternative "The Dutch Golden Age **growers**...", a noun following an NP modifier, where presumably a verb would be more expected. Item 8, "Then the little bird **guarded** by the owl peeped out, ..." is in an example of main verb / reduced-relative (MV/RR) garden-path, however the surprising word comes *before* the disambiguating word in the noun phrase (where surprisal-based processing difficulty is theoretically predicted). Item 11 begins a CP subordinating conjunction "A meeting of all the birds was called and [**though** they understood the birds ... would be unable to come], many birds came from faraway meadows and woods." Item 10 is a gerund modifier "...and then

[7] Difficulty in ORCs has been explored in a number of previous studies focusing on predictions about where the locus of difficulty is—the subject or the verb, with the former traditionally being the prediction of surprisal-based theories (see e.g., Traxler et al., 2002; Staub, 2010; R. Levy et al., 2013; Vani et al., 2021). It is perhaps interesting to note that words from *both* critical places in ORCs are represented in the list of highest-surprisal items—not just at the subject, but also at the verb.

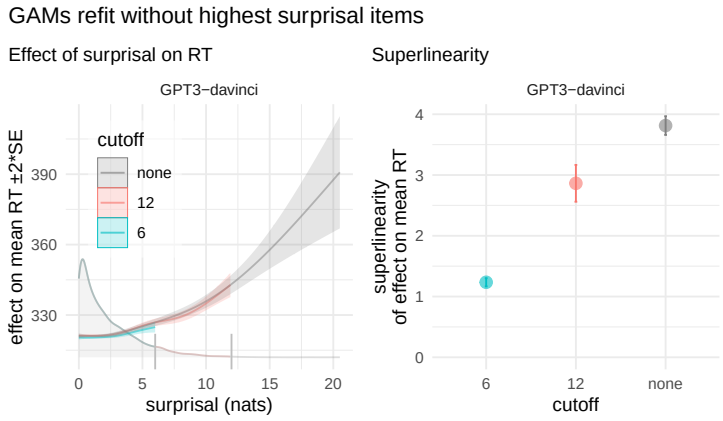


Figure A.4: GAMs fit on GPT-3 Davinci surprisals with highest surprisal items removed. **Left**: The effect of surprisal on mean RT, fit on data subset with surprisal ≤ 6 (blue) and ≤ 12 (red). For comparison we also plot the fit on all data (grey; repeated from fig. 2.2). **Right**: Superlinearity of these curves (grey point repeated from fig. 2.4).

folding his wings together, he sank to earth…". The remaining handful of words are other somewhat rare modifiers (items 3, 5, 6, 12, 17, 18), which are plausibly hard to predict especially given they come before their heads. Note that for the purpose of understanding the empirical relationship between surprisal and processing time, what matters about these words is simply that they are surprising. It is reassuring to see that for the most part they seem like items which would be intuitively hard for humans to predict.

A.6.2 Models without highest surprisal words

To determine the extent to which our conclusions about superlinearity rely on the relatively few highest-surprisal items, we re-fit nonlinear GAMs (formula A.1) including only those items in the corpus with surprisal below a cutoff value: $\{w \in \text{Corpus} \mid s(w) \leq s_{\text{cutoff}}\}$.

We fit two versions of this control: one with $s_{\text{cutoff}} = 12$, and and one with $s_{\text{cutoff}} = 6$. Cutting off above surprisal threshold $s_{\text{cutoff}} = 12$ removed the 19 words discussed above in table A.1 (which comprise 1557 RT observations, roughly 0.3% of total observations in the data). Cutting off above $s_{\text{cutoff}} = 6$ removed an additional 470 words (489 words total, comprising 41261 RT observations, roughly 7.6% of total observations in the data).

Figure A.4 (left) shows the fitted effect of surprisal on mean RT from these GAMs ($s_{\text{cutoff}} = 12$ in red, $s_{\text{cutoff}} = 6$ in blue), compared to the model fit on all words (grey, repeated from fig. 2.2). Figure A.4 (right) shows the superlinearity of these curves. We observe that the exclusion of these high-surprisal items leaves the shape of the curve basically unchanged in the remaining lower-surprisal region. Truncating the curve like this naturally reduces the amount of superlinearity we see, but the curve remains superlinear, even with the more drastic cutoff.

```
RT ~ s(surp, bs='tp', k=6) + s(subj, surp, bs='fs', m=1) + te(freq, len) +
     s(prev_surp, bs='tp') + s(subj, prev_surp, bs='fs', m=1) + te(prev_freq, prev_len)
```

Formula A.3: The `mgcv` formula for the nonlinear GAM with constant variance.

```
RT ~ surp + s(subj, bs='re') + s(surp, subj, bs='re') + te(freq, len) +
     prev_surp + s(prev_surp, subj, bs='re') + te(prev_freq, prev_len)
```

Formula A.4: The `mgcv` formula for the linear control GAM with constant variance. The interpretation of this formula is essentially the same as that of formula A.3, except that the effect of surprisal on reading time is assumed to be linear.

A.7 Additional controls

A.7.1 Gaussian GAMs with constant variance assumption

For comparison with the GAMs discussed in the main text, which fit the effect of surprisal on variance in reading time as well as mean, we also fit versions of these models with a constant variance assumption (formulæ A.3 and A.4). In addition to allowing a more direct comparison with previous work, which has largely used Gaussian constant-variance GAMs (Smith & Levy, 2008a, 2013; Goodkind & Bicknell, 2018; Wilcox et al., 2020; Hofmann et al., 2022), these models also function as a control for the effect that fitting variance might have had on the shape of the relationship with mean RT. They also have the benefit of being much less costly to compute than the models which must fit the effect on variance as well as mean of the response.

Figure A.5 shows the relationship between surprisal and RT according to these models (compare with the mean effect in fig. A.1). As with the results presented in the main text, these results show increasing superlinearity with LM quality.

A.7.2 Spillover and autocorrelation

When fitting a mixed-effects model or GAM to predict reaction time data, it is common practice to include additional predictors for the previous word—or, more generally all words within a M-word window including the current word to control for spillover effects (D. C. Mitchell, 1984; Vasishth, 2006). For our models, we follow previous literature in this area (e.g., Goodkind & Bicknell, 2018, 2021; Meister et al., 2021) in including predictors for one previous word for spillover control ($M = 2$). However, some other studies (e.g., Wilcox et al., 2020) have used $M = 4$, following Smith and Levy (2013) who noted that a window size of $M = 4$ was empirically best to capture the effect of surprisal on self-paced reading time in their study. For our models, we found that including more than one previous word was computationally intractable, since predictors for each

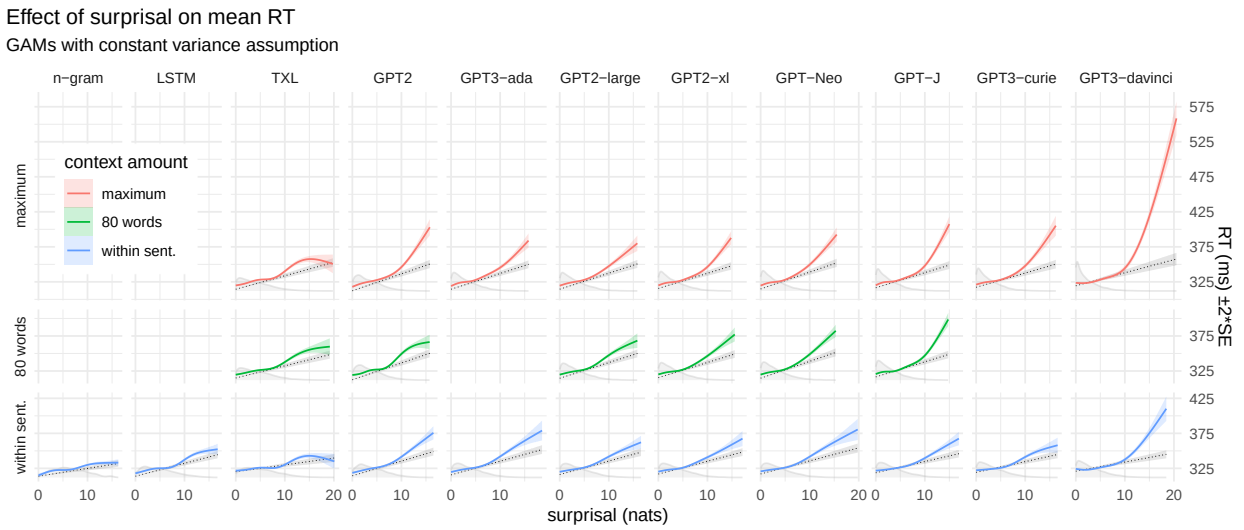


Figure A.5: The effect of surprisal on self-paced reading time from GAM models which assume constant variance (formulæ A.3 and A.4). Solid lines are the fitted effects from the nonlinear GAMs, dashed lines beneath are from the corresponding linear control GAMs. Shaded regions represent 95% CIs. Cf. fig. A.1, top panel (effect on mean RT).

additional spillover word adds a full set of by-subject nonlinear effects for both location and scale.[8] In this section we investigate the degree to which this choice could have affected our results.

Autocorrelation plots One way to assess whether a larger M would have likely affected our results is to look for residual autocorrelation in our models. Intuitively, spillover effects cause time-dependence in the response, since higher surprisal on a word will result not just in higher reading time on the current word, but this effect will also "spill over" to the subsequent word (or words). Intuitively, if such time-dependence is not fully captured by our models, this will result in time-dependence in the residuals. We can look for evidence of such time-dependence by looking for autocorrelation in the residuals.

Figure A.6 shows the mean (complete) autocorrelation (left) and mean partial autocorrelation (right) for the nonlinear GAM fit on GPT-3 Davinci surprisals, averaged across stories and subjects.[9] 95% CIs are shaded red. Autocorrelation for GAMs fit on surprisals from other LMs are similar.

These plots indicate that there amount of residual autocorrelation is small for any lag. In the PACF plot, for all $k > 3$ partial autocorrelation is not significantly different from zero, and even for $k \leq 3$, partial correlation values are small. This suggests that optimally we should include predictors for three previous words ($M = 4$), but we may expect that doing so would not have a

[8] We attempted fitting models with more previous words ($M = 3$ and $M = 4$), but found that this resulted in models whose design matrices that were too big for `mgcv::gam`. Unfortunately the more efficient procedure `bam` is not currently implemented for location-scale GAMs.

[9] For lag k, the autocorrelation function $\mathrm{ACF}(k)$ gives the correlation between observations k words apart; partial autocorrelation $\mathrm{PACF}(k)$ is the amount of correlation that is not accounted for by lags 1 through $k - 1$.

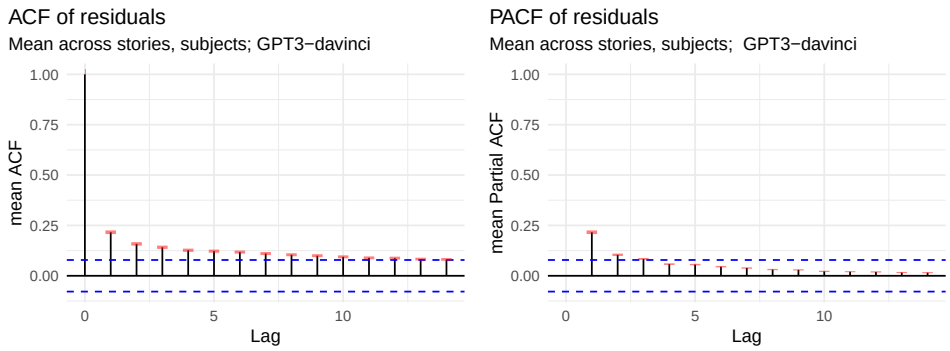


Figure A.6: Plots of mean autocorrelation function (ACF; **left**) and mean partial autocorrelation function (PACF; **right**) of residuals for the nonlinear GAM for GPT-3 Davinci. For a given lag value, bar height represents the mean (P)ACF across stories and subjects, with 95% CI in red. Dashed blue lines indicate significance thresholds (against white noise null hypothesis).

large effect on results.

Additional predictors for spillover We also experimented with fitting the simpler constant-variance models (described above in the first subsection of this appendix), but with predictors for the previous three words, to control for spillover.

These GAMs are plotted in fig. A.7 (solid lines), together with GAMs with only one previous word (dashed lines; repeated from fig. A.5), for comparison. Grey dotted lines are the linear control models (also repeated from fig. A.5). We observe that in most cases there is little difference between the curves with three spillover words compared to those with only one: Some fits become slightly more visually superlinear, and others slightly less. One large change is in GPT-3 Davinci, which does become much less steeply superlinear in the high end of the surprisal range, but remains superlinear overall.

A.7.3 Without by-subject effects

Unlike our study, Wilcox et al. (2020) use GAMs to model mean item reading time as the response, and do not control for by-subject random effects. For comparison with their results, we also fit models of mean RT without the by-subject effects (formula A.5). These models were fit with a constant-variance assumption, for computational efficiency, given that the superlinearity we observed was robust to this simplifying assumption, as discussed above. Figure A.8 (analogous to fig. A.1) provides plots of GAMs fit with this formula. The results show much larger confidence intervals, suggesting that properly modelling by-subject variation in this data gives us higher power to detect population-level nonlinear effects.

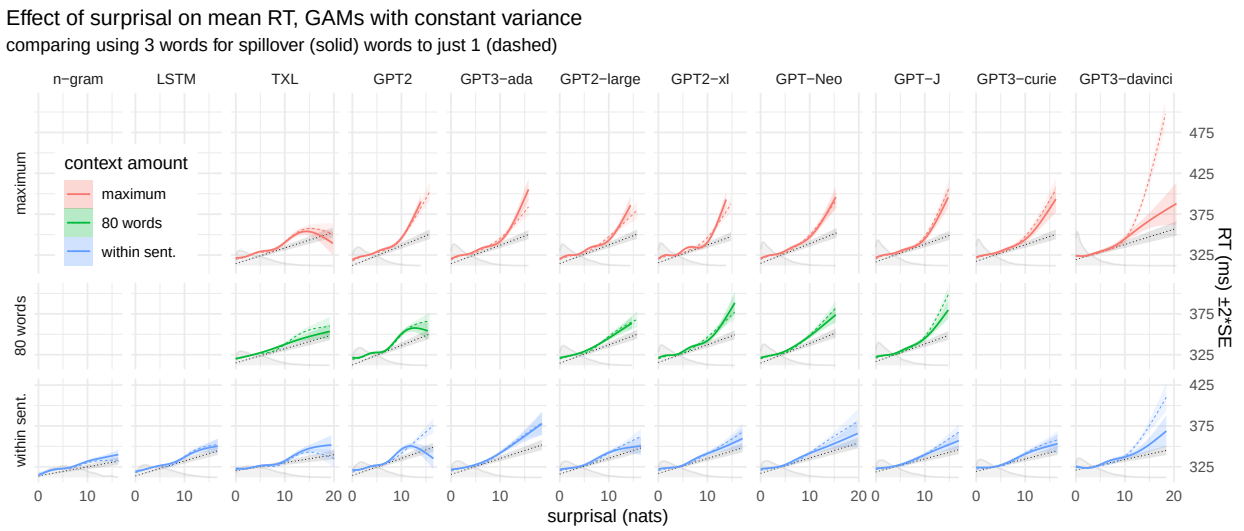


Figure A.7: Comparing the relationship between surprisal and RT using GAMs with spillover control predictors for three previous words (solid lines) to GAMs with only one word for spillover (dashed lines, repeated from fig. A.5). Linear control models plotted as dotted lines (also repeated from fig. A.5). All GAMs for this plot were fit with an assumption of constant variance.

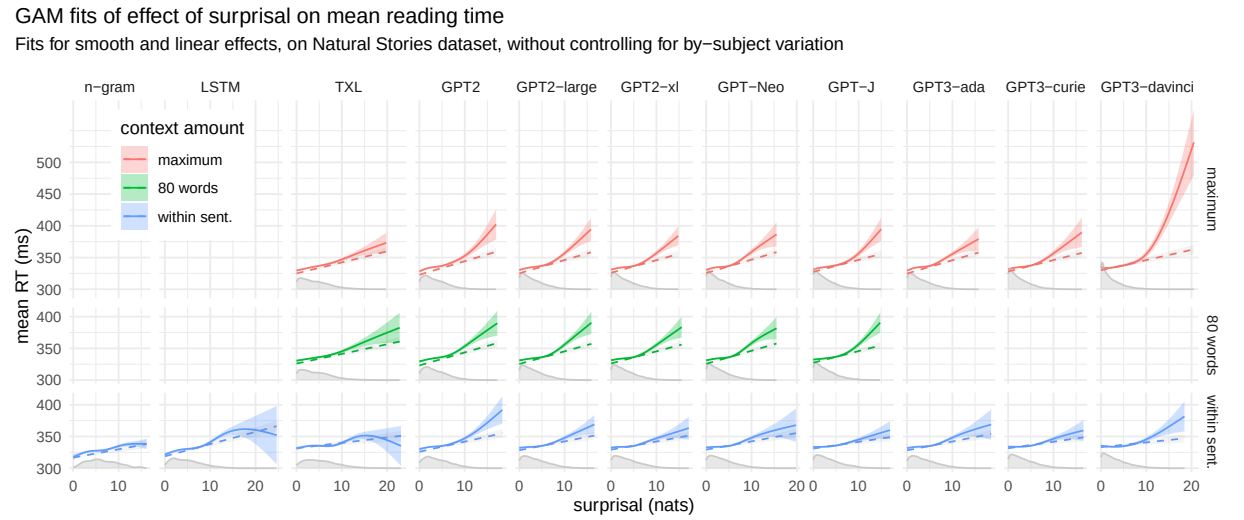


Figure A.8: Plots of effect of surprisal on mean RT for constant-variance GAMs which do not control for by-subject differences (formula A.5).

```
RT ~ s(surp, bs='tp', k=6) + te(freq, length) +
     s(prev_surp, bs='tp') + te(prev_freq, prev_length)
```

Formula A.5: The `mgcv` formula for nonlinear GAM fits without by-subject effects. Mean reading time is predicted as a nonlinear global effect of surprisal, controlling for interactions between log frequency and orthographic length, all for the current word as well as the previous. Compare to formula A.3, which also includes factor smooths by subject.

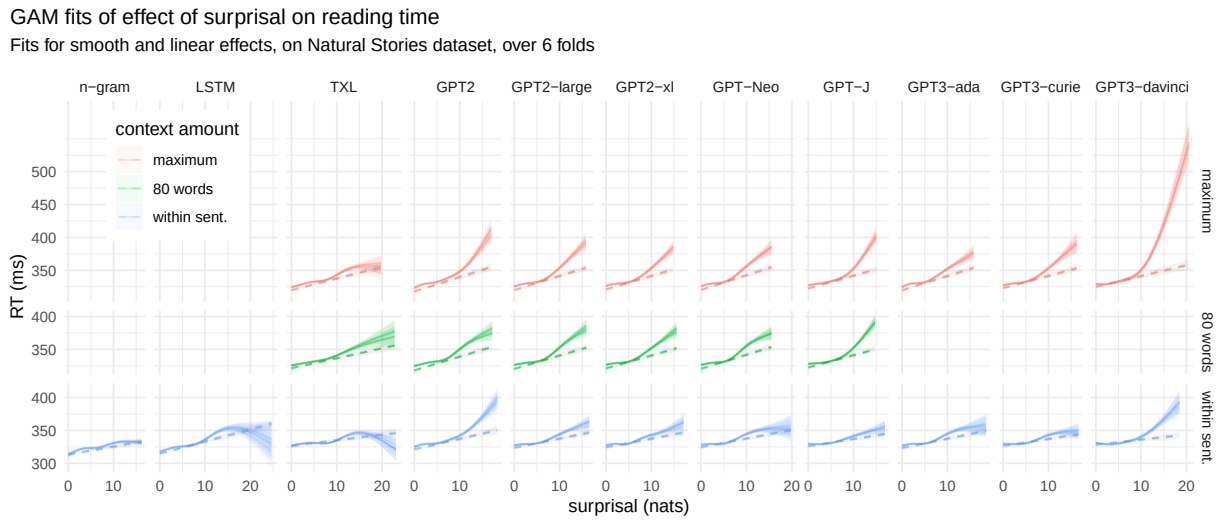


Figure A.9: Plots as in fig. A.5, except that here we plot fitted curves for each of 6 GAMs fit on randomized folds of 5/6ths of the dataset. Similarity across folds to suggests the models are not overfitting.

A.7.4 GAM plots from folds of data

To insure against potential high-leverage outliers, we carried out a cross-validation control. For this control, we partitioned the data into 6 folds, and refit the GAMs 6 times leaving out one fold each time. These models were fit with a constant-variance assumption, for computational efficiency (as with the previous control).

The fitted effect of surprisal on reading time for each of the 6 folds, with confidence intervals, are plotted superimposed in fig. A.9. Comparing these results with the plots for GAMs fit on all of the data in fig. A.1 we can visually confirm that the results are effectively identical, and conclude that the superlinearity we see is robust.

A.8 Probability-ordered search runtime

For deterministic ranked search, if items are sorted in order of decreasing probability, runtime is simply the number of items with higher probability (assuming no two items have the same probability, in which case the sorting is not well defined). Here I will derive that runtime of probability-ordered search increases exponentially in surprisal, for two cases: when the probabilities are Pareto-distributed, or when the odds are.

A.8.1 Assuming Pareto weights

Assume the item weights (that is, unnormalized probabilities) are distributed according to a Pareto distribution. The probability of an item with weight w is $\frac{w}{Z}$, for some global normalizing constant Z. Then item weights have density[10] $\text{pdf}(\mathsf{w}) = a\mathsf{w}^{-(\alpha+1)}$, and so then item surprisal s ($= -\log \frac{\mathsf{w}}{Z}$), has density:

$$\text{pdf}(\mathsf{s}) = \frac{a}{Z^\alpha} e^{\alpha \mathsf{s}} \tag{A.6}$$

Derivation (via transformation of a random variable):

Surprisal is a (monotonic smooth) deterministic function of item weight, so, starting with the pdf for weights $f(\mathsf{w})$, we can derive the expression for the pdf of surprisal $h(\mathsf{s})$ as follows. With the inverse transformation $\mathsf{w}(\mathsf{s}) = Ze^{-\mathsf{s}}$, we have

- $f(\mathsf{w}(\mathsf{s})) = a[Ze^{-\mathsf{s}}]^{-(\alpha+1)} = aZ^{-(\alpha+1)}e^{(\alpha+1)\mathsf{s}}$, and
- $\frac{\mathrm{d}}{\mathrm{d}\mathsf{s}}\mathsf{w}(\mathsf{s}) = -Ze^{-\mathsf{s}}$

so the pdf of surprisal is

$$\begin{aligned} h(\mathsf{s}) &= f(\mathsf{w}(\mathsf{s})) \cdot \left| \frac{\mathrm{d}}{\mathrm{d}\mathsf{s}}\mathsf{w}(\mathsf{s}) \right| \\ &= \frac{a}{Z^{\alpha+1}} e^{(\alpha+1)\mathsf{s}} \cdot Ze^{-\mathsf{s}} = \frac{a}{Z^\alpha} e^{\alpha \mathsf{s}} \end{aligned} \tag{A.7}$$

Then, for target item i, the proportion of items with surprisal lower than $s(i)$ is

$$\Pr(\mathsf{s} < s(i)) = \int_0^{s(i)} \text{pdf}(\mathsf{s})\, \mathrm{ds} = \frac{a}{\alpha Z^\alpha}(e^{\alpha s(i)} - 1) \tag{A.8}$$

so, (collecting constants for simplicity) since the runtime to find item i is proportional to the number of items of lower surprisal, we have

$$\text{Time}(i) = K(e^{\alpha\, s(i)} - 1) \tag{A.9}$$

[10] The density of $X \sim \text{Pareto}(\alpha; \theta)$, with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$, is $f(x) = \alpha\theta^\alpha x^{-(\alpha+1)}$ (for domain $x \geq \theta$). Here I abbreviate with constant $a := \alpha\theta^\alpha$.

Thus search runtime increases exponentially with surprisal.

A.8.2 Assuming Pareto odds

Assume instead that the odds are Pareto distributed, rather than the weights, then we likewise still get that search runtime increases exponentially with (negative) log-odds (as derived in Anderson & Lebiere, 1998): Odds have density $\text{pdf}(\mathsf{o}) = a\mathsf{o}^{-(\alpha+1)}$ so log-odds, r $(= \log \mathsf{o})$, have density

$$\text{pdf}(\mathsf{r}) = ae^{-\alpha \mathsf{r}} \tag{A.10}$$

by transformation, similar to above. Then, as above, for target item i one can derive relationship between log-odds and search runtime as the proportion of items with *higher* log odds:

$$\Pr(\mathsf{r} > \text{logodds}(i)) = \int_{\text{logodds}(i)}^{\infty} \text{pdf}(\mathsf{r})\, d\mathsf{r} = \frac{a}{\alpha} e^{-\alpha \text{logodds}(i)} \tag{A.11}$$

Letting constant $K = a/\alpha$, this is the form of the equation called the 'latency formula' in ACT-R:

$$\text{Time}(i) = Ke^{-\alpha \text{logodds}(i)} = K(e^{s(i)} - 1)^{\alpha} \tag{A.12}$$

by the identity $\text{logodds}(\cdot) = -\log(e^{s(\cdot)} - 1)$. So, search runtime increases exponentially with surprisal. Additionally assuming $\alpha = 1$, as is common in the ACT-R literature, this simplifies to $\text{Time}(i) \propto e^{-s(i)} - 1$.

B

Supplemental material for chapter 3

B.1 Language model details

With the exception of OpenAI's GPT-3 models, we obtained all surprisal estimates from pretrained models using the implementations available through Huggingface Transformers (Wolf et al., 2020), version 4.35.2.

The models we used were the following (model names followed by corresponding Transformers model IDs):

- GPT-2 (Radford et al., 2019): `gpt2`, `gpt2-xl`;
- GPT-Neo and NeoX (Black et al., 2021; Black et al., 2022): `EleutherAI/gpt-neo{-2.7B,x-20b}`;
- OPT (S. Zhang et al., 2022): `facebook/opt-{350m,2.7b,6.7b,13b,30b,66b}`
- OLMo (Groeneveld et al., 2024): `allenai/OLMo-{1B,7B}`;
- Llama-2 (Touvron et al., 2023): `meta-llama/Llama-2-{7b-hf,13b-hf,70b-hf}`
- Llama-3 (AI@Meta, 2024; Llama team, 2024): `meta-llama/Meta-Llama-3-{8B,70B}`
- Mistral and Mixtral (Jiang et al., 2023, 2024): `mistralai/{Mistral-7B-v0.1,Mixtral-8x7B-v0.1}`.

For surprisal estimates from GPT-3, we used log probabilities provided via the OpenAI's "Completions" API, for `babbage-002`, and `davinci-002` models. Note this API endpoint was retired on January 4 2024,[1] and is labelled as "legacy" as of time of access (mid-January 2024), but is to our

[1] See migration announcement here https://openai.com/index/gpt-4-api-general-availability/.

knowledge the only interface currently made available by OpenAI which allows access to models' estimated log-probability of input tokens. For all more recent models from OpenAI, using the "Chat Completions" API, this functionality is no longer supported: Token log probabilities are available only for the model's generated completions, not the user-provided prefix, ruling out the use of GPT-3.5 and later for surprisal estimates.

B.2 Additional empirical plots

B.2.1 Spillover vs reading speed

To the extent to which there is a difference in processing cost between the experimental conditions, observable as a slowdown in reading time, this slowdown may occur on the target word or the next few words, due to spillover. For this reason, as a measure of processing cost, we use the average reading time in a three-word region of interest, starting at the target word.

Faster readers may be expected to show the effect at later lags. Figure B.1 shows the mean slowdown on different lags, for participants binned by quantiles of average reading speed. It is apparent that for most readers, the effect is largest at the word after the target (lag=1), while for the slowest readers, the effect shows up most clearly on target word itself, and for the fastest readers it shows up most clearly at lag=2 (as highlighted in the plot by facets with light-brown shaded background in the first three rows).

This observation provides a justification for averaging across a 3-word window as a rough way to capture the slowdown effect of interest despite the variation between individuals. The systematic relationship between average reading speed and lag at which the effect is most pronounced suggests the effects sizes reported in this study would likely be even larger if we controlled for this relationship between reading speed and lag directly in more sophisticated regressions.

B.2.2 Surprisal means by language model family

Figure B.2 provides an alternative view of the data presented in fig. 3.3 (right), with the language models separated into four groups, to make the patterns across conditions for each language model more easily visible.

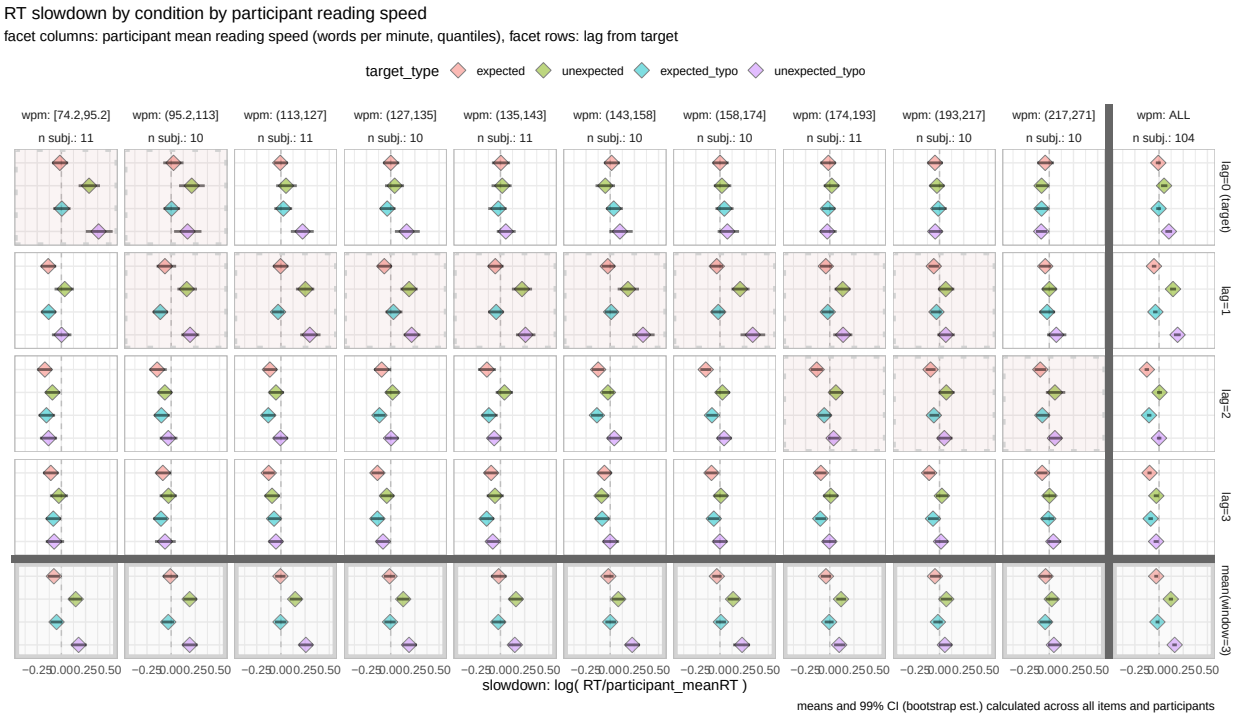


Figure B.1: Reading time response across the four experimental conditions, by participant mean reading speed and lag. In each subplot, horizontal axis is reading time slowdown (log RT relative to participant mean RT), vertical axis is experimental condition (target type). Diamonds mark mean values; horizontal lines indicate 99% CIs. **Columns:** Participants are partitioned into deciles by mean reading speed; slowest on the left, faster to the right. The mean for all participants together is on the right of the grey vertical line. **Rows:** First four rows show slowdown for different lags: Lag=0 refers to slowdown on target word, lag=1 is slowdown on the subsequent word, etc. Bottom row (below horizontal line) is RT averaged on three-word window starting at target word (the response used in our analyses). Highlighted cells in first three rows indicate the trend in lag values for which the difference between `expected` and `unexpected` is largest.

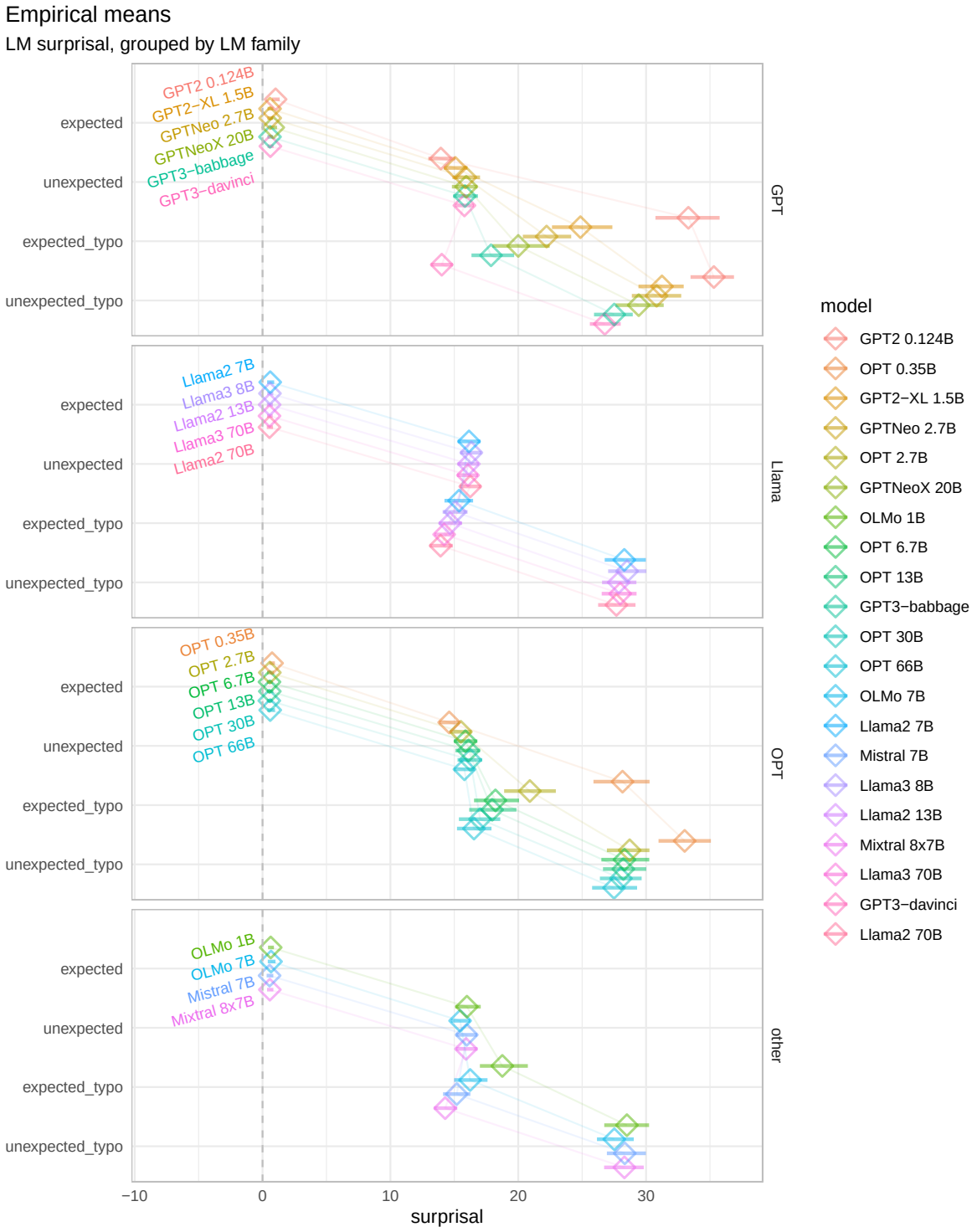


Figure B.2: This figure repeats the data in fig. 3.3 (right), showing empirical means of LM surprisal across the four experimental conditions. LMs are grouped into four subplots to make within-LM patterns more easily visible: GPT models, Llama models, OPT models, and other. Horizontal axis is surprisal. Diamonds mark empirical mean values, with horizontal lines indicating 99% CIs. Within each group, LMs are ordered by their mean surprisal on the `expected_typo` condition.

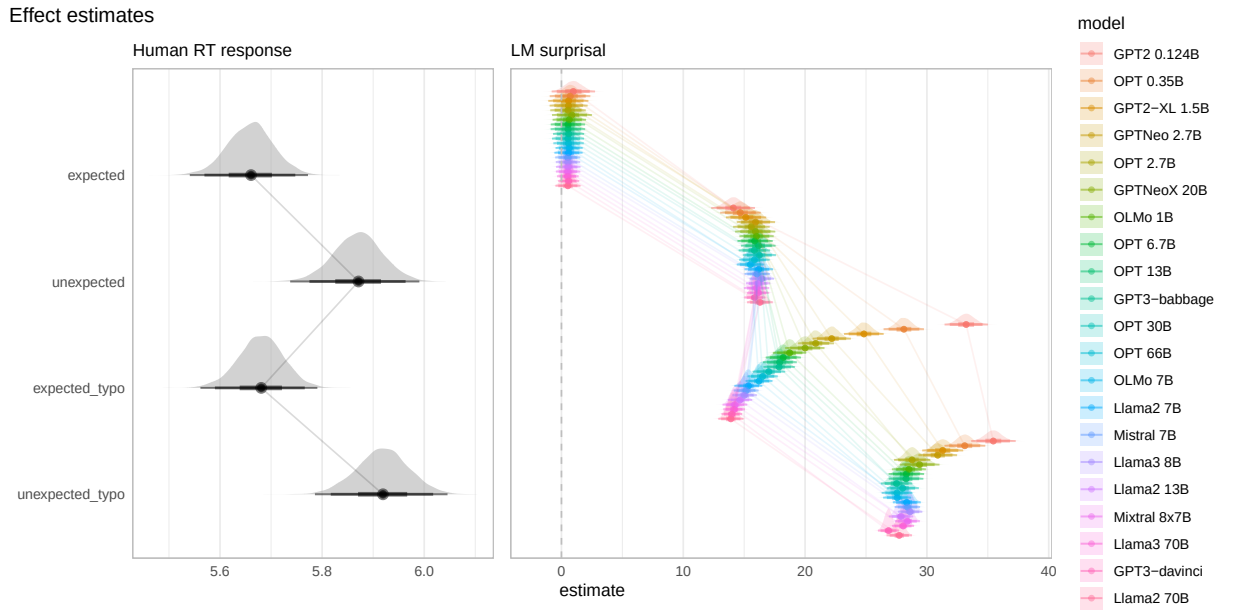


Figure B.3: Effect estimates for regressions of human reading time response and LM surprisal, across the four experimental conditions. Dots mark regression models' estimates; uncertainty is represented as density plot, with horizontal bars for credible intervals (66%, 95%, 99%). These estimates closely matches the patterns displayed in plots of empirical means (fig. 3.3).

B.3 Bayesian linear regressions

As described in the main text, we fit Bayesian mixed-effects linear regression models to predict reading time and, separately, surprisal from each LM, using `brms` (Bürkner, 2017). Figure B.3 displays the estimated marginal effect on response (RT or surprisal) for each of the four experimental conditions, with median effect estimate marked as a dot, and uncertainty indicated in half-eye density plots with horizontal bars at 66%, 95%, and 99% CIs. Contrasts between these conditions, according to these regression models, selected to address our research questions, were displayed in fig. 3.4 (in the main text).

B.3.1 Regression fit diagnostics

All models fit without any divergent transitions, and with $\hat{R}$ values below 1.01.

Graphical posterior predictive checks Posterior predictive density plots from the Bayesian models are displayed in fig. B.4, for comparison with empirical data density plots. The thick dark line in each plot is the empirical data density (labeled $\mathbf{y}$), and the narrow light orange curves give 50 density plots (labeled $\mathbf{y}_{\text{rep}}$) each representing a simulation from the posterior predictive distribution defined by the model. The posterior predictive distribution for hypothetical data $\mathbf{y}_{\text{rep}}$ can be written as follows, marginalizing over the posterior beliefs about the regression model's

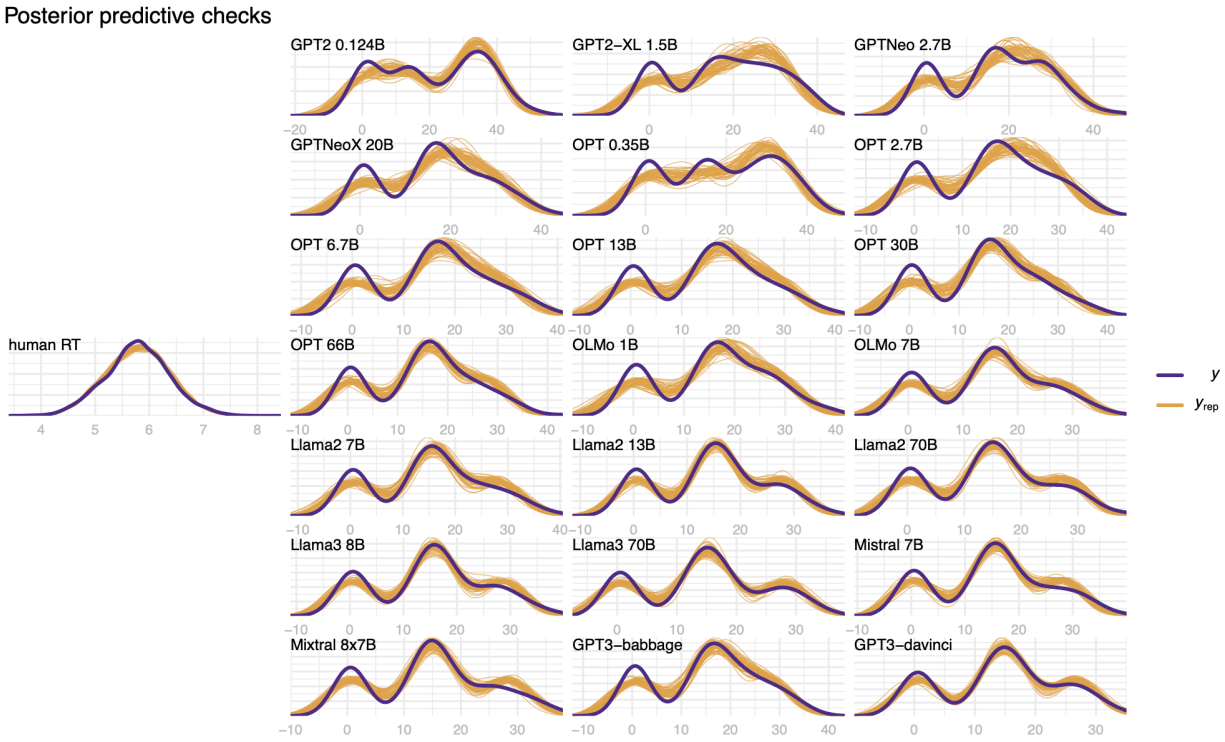


Figure B.4: Overlaid density plots for 50 simulations representing the posterior predictive distributions ($\mathbf{y}_{\text{rep}}$, narrow orange curves), overlaid with the density plot for the empirical data ($\mathbf{y}$, thick dark curve).

parameters: $p(\mathbf{y}_{\text{rep}} \mid \breve{\mathbf{y}}) = \mathbb{E}_{\theta \sim p_{\Theta|\breve{\mathbf{y}}}}[p(\mathbf{y}_{\text{rep}} \mid \theta)]$. Each $\mathbf{y}_{\text{rep}}$ sampled from this distribution replicates the data generation process, according to the model, with variation between samples representing the uncertainty in the distribution about θ given the data.

These graphical posterior predictive comparisons provide a simple check of model adequacy: If a regression model is fit successfully, the data it generates (the posterior predictive distribution) should look similar to the empirical data. In our case we see that across the models, the posterior predictive distribution is generally similar to the empirical distribution for the regression of human RT, and for each of the LM surprisals.

B.3.2 Group-level consistency in results

Figure B.5 shows by-item and by-subject conditional effects. Each subplot represents a single item (left group) or participant (right group) in the data. In each subplot, as in the overall plot in fig. 3.4, the vertical axis presents the three contrasts of interest, with the estimated contrast (units of log ms), with uncertainty represented as a density plot, a dot at the median, and horizontal bars for 0.66, 0.95, and 0.99 CIs. These plots represent the average marginal effect for a single typical item (cf. participant), when conditioned on a given participant (cf. item). For further description of the

computation and interpretation of estimated conditional and marginal mean (a.k.a. least-squares mean; Searle et al., 1980) effects in the analysis multilevel regression models, see, e.g., Cai (2014) and Sonderegger (2023, chs. 7–9), and less formal discussion with examples in, e.g., Heiss (2021) and documentation and vignettes for the `emmeans` package (Lenth, 2024).

Taken as a group, despite there being substantially more uncertainty in each individual plot than the overall estimated marginal mean effects for these contrasts, these plots confirm that the same pattern holds within grouping variables: The typo effect is small or nonexistent, and the unexpectedness effect is generally larger.

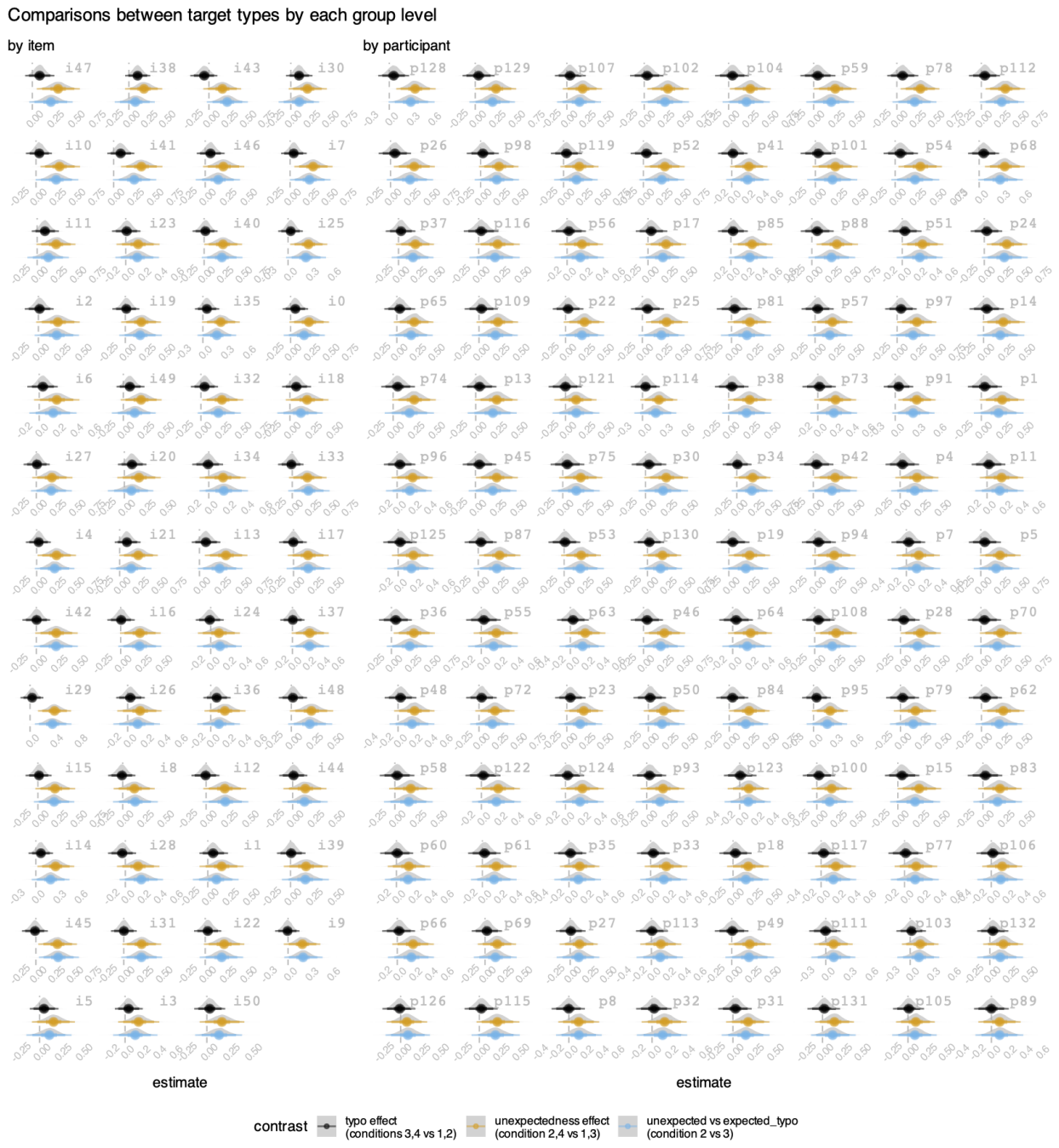


Figure B.5: Average marginal contrasts conditioning on each item (left subplots), or participant (right subplots). Each item-subplot represents the effect on RT for that item, for a single typical participant, based on the distribution of existing participants in the data. Likewise, each participant-subplot represents the effect on RT for that participant, for a single typical item.

B.4 Frequentist linear regressions

Using same structure as the Bayesian regressions (formulæ 3.1 and 3.2), we fit frequentist linear mixed-effects regressions to predict reading time and, separately, surprisal from each LM, using `lme4` (Bates et al., 2015; Kuznetsova et al., 2017). These regressions results conform with the interpretation from the Bayesian regressions reported in the main text.

Figure B.6 displays the post-hoc contrasts (equivalent to the comparisons in fig. 3.4), according to the regression models of human RT (left), and LM surprisals (right), computed with `emmeans` in R (Lenth, 2024). On the vertical axis is the comparison in question, and on the horizontal axis is the estimated contrast (in units of log ms for the RT regression, and nats for the surprisal regressions). Estimated 99% confidence intervals are displayed as shaded regions. These results mirror the interpretation of the Baysian regressions (§3.4.1): The typo effect on human RT is small if it exists at all, whereas the effect of unexpectedness on RT is large, and of similar magnitude to the `unexpected` vs `expected_typo` effect. By contrast, for LM surprisal, the typo effect is larger or similar size to the unexpectedness effect (see below). Correspondingly, the `unexpected` vs `expected_typo` effect is negative for the smallest LMs and near zero for the better LMs.

We can't directly compare the relative effect of these contrasts for a given LM by just checking if CIs overlap in fig. B.6 to determine significance of differences. However, significance checks for these comparisons (fig. B.7) confirm that the typo effect is significantly smaller than the unexpectedness effect only for the largest/most recent few LMs, and even for those few the effect sizes for the unexpectedness and typo effects are similar, and both are significantly and substantially larger than the `unexpected` vs `expected_typo` effect. The typo effect is larger than the unexpectedness effect for the smaller/older LMs, and it is only for the very best few LMs that the unexpectedness effect is in fact larger than the typo effect (slightly, but significantly, at the 0.01 level). For many LMs in the middle, they are not significantly different sizes. None of the models have typo effect *smaller* than the `unexpected` vs `expected_typo` effect, which is the pattern seen clearly in the human reading times.

B.5 Experimental materials

Table B.1 contains all experimental items and comprehension questions for the self-paced reading time experiment. Table B.2 contains the practice sentences and comprehension questions.

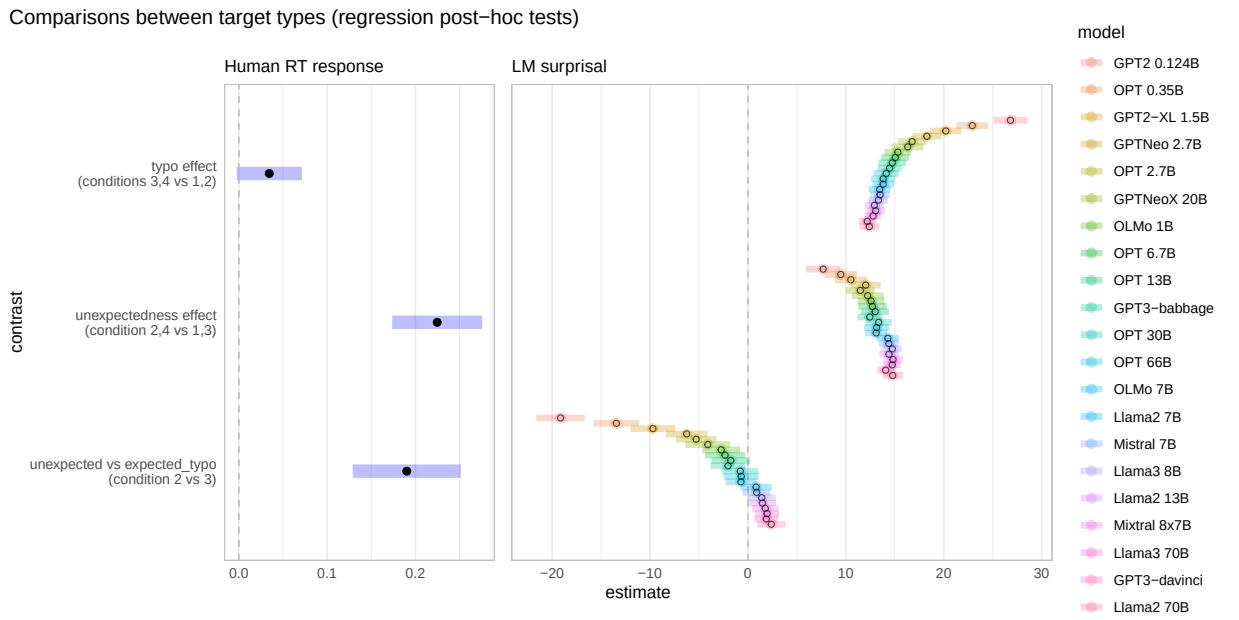


Figure B.6: Results of post-hoc comparisons between conditions' effect on RT (left) vs surprisal (right), from frequentist regressions. Dots mark estimated marginal means, horizontal bars give 99% CIs. Results are equivalent in interpretation to those from Bayesian regressions (fig. 3.4).

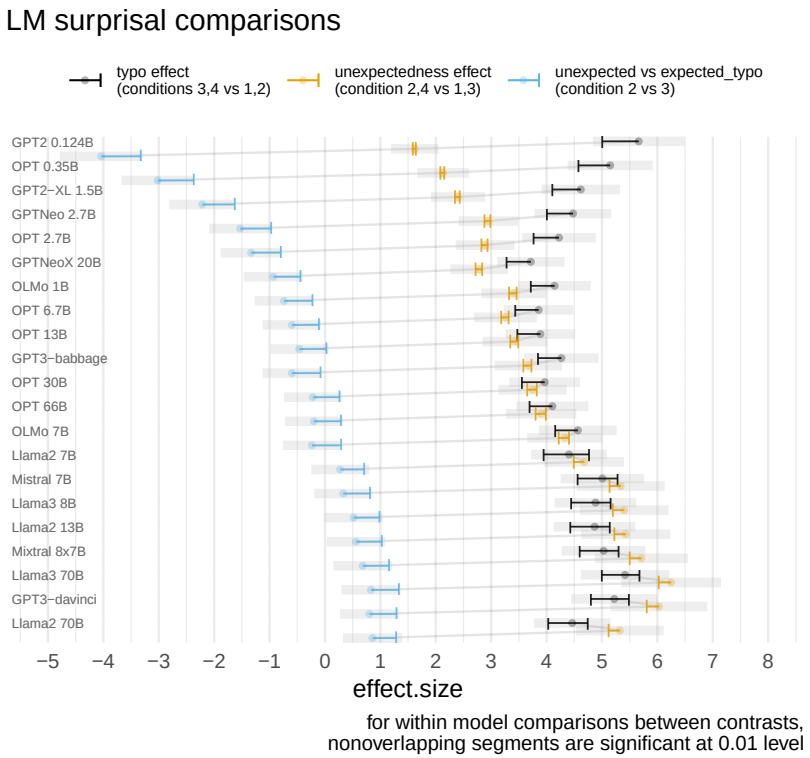


Figure B.7: Surprisal post-hoc contrasts for each LM, from frequentist regressions. For contrast of interest (typo effect, unexpectedness effect, and `unexpected` vs `expected_typo`), dots indicate estimated standardized effect size, with grey horizontal bars indicating 99% CIs. Segments indicate significance of comparisons between effect for a given LM: When segments do not overlap, a directed comparison is significant at the 0.01 level.

Table B.1: Stimuli and comprehension questions for self-paced reading experiment.

	stimulus					answer choices			
item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
0	It is unfortunately hard to drive through the city in the summertime, because so many of the streets are closed due to	expected unexpected expected_typo unexpected_typo	**construction** **democracy** **construtcion** **democarcy**	causing inevitable delays.	*What time of year was referred to?*	Summer.	Spring.	Autumn.	Winter.
1	Upon examining the evidence closely, the detective turned to his assistant and said, 'Indeed! It's just as I	expected unexpected expected_typo unexpected_typo	**suspected** **rejected** **suspetced** **rejetced**	it could be. The clues all point to a single suspect.'	*Who examined the evidence?*	The detective.	The assistant.	The judge.	The scientist.
2	Following a series of setbacks and performance metrics that were lower than expected, it was obvious that the team leader was having second	expected unexpected expected_typo unexpected_typo	**thoughts** **analyses** **thougths** **anaylses**	regarding the project.	*How was the project performing?*	Worse than expected.	Better than expected.	As well as expected.	Not enough information to say.
3	With its compelling storyline and groundbreaking special effects, the new film impressed the critics in every	expected unexpected expected_typo unexpected_typo	**category** **village** **catgeory** **vilalge**	and was nominated for a number of awards.	*What was nominated for awards?*	The film.	The play.	The book.	The TV show.
4	Due to her business acumen and strong leadership skills, she was quickly promoted to become the vice	expected unexpected expected_typo unexpected_typo	**president** **required** **presidnet** **requried**	of the company.	*Which of the following positive attributes did she have?*	Business acumen.	Academic achievement.	Job experience.	Foreign language proficiency.
5	Among the artifacts found in the archaeological site, several have unique markings, indicating they might be of historical	expected unexpected expected_typo unexpected_typo	**significance** **celebrations** **significnace** **celebratoins**	and are worth more detailed investigation.	*Where were the artifacts found?*	An archaeological site.	A historical museum.	A private collection.	An inventory list.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
6	In the business startup community, Lisa is admired for her ability to start and grow new ventures from the ground up. She's often invited to speak about her experiences as a successful	expected unexpected expected_typo unexpected_typo	**entrepreneur** **adolescent** **entreperneur** **adolecsent**	and enjoys sharing her insights with the community.	*In what community is Lisa admired?*	The business startup community.	The academic philosophy community.	The local political community.	The performing arts community.
7	Today marks an important milestone for the couple. They're planning a large celebration for their thirtieth wedding	expected unexpected expected_typo unexpected_typo	**anniversary** **frustration** **anniverasry** **frustartion**	to reflect on their decades together.	*Who is reflecting on their time spent together?*	A couple.	A sports team.	A comedy duo.	A parent and child.
8	As a foreign exchange trader, he constantly monitors the exchange rates between dollars and pounds, or other different national	expected unexpected expected_typo unexpected_typo	**currencies** **weaknesses** **currecnies** **weankesses**	to capitalize on market fluctuations.	*What activity did the trader constantly monitor?*	Exchange rates.	Stock market prices.	Political news.	Commodity prices.
9	Ever since the haunted house visit, the children couldn't sleep peacefully, and were troubled by recurring	expected unexpected expected_typo unexpected_typo	**nightmares** **advertising** **nighmtares** **advetrising**	which left them feeling uneasy.	*What location did they visit?*	A haunted house.	A technology fair.	An art gallery.	A nature reserve.
10	Studies have shown a strong correlation between student engagement and academic	expected unexpected expected_typo unexpected_typo	**achievement** **revolutions** **acheivement** **revolutoins**	within the context of overall educational progress.	*Which aspect of the educational environment was examined?*	Student engagement.	Student-to-teacher ratio.	Graduation rates.	Assignment workload.
11	In a sudden expression of surprise, her eyes widened and she involuntarily wrinkled her forehead by raising her	expected unexpected expected_typo unexpected_typo	**eyebrows** **buckets** **eyeborws** **bukcets**	as high as she could.	*Which emotion did she express?*	Surprise.	Fear.	Disgust.	Joy.
12	In the language learning class, the teacher emphasized the importance of expanding one's	expected unexpected expected_typo unexpected_typo	**vocabulary** **processors** **vocabluary** **procesosrs**	to become more fluent.	*What was the subject of the class?*	Language.	Mathematics.	History	Geography.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
13	After tripping over the rug in front of everyone at the party, she quickly got up, but her cheeks turned red and she felt deeply	expected unexpected expected_typo unexpected_typo	**embarrassed** **innovative** **embarrsased** **innovaitve**	as she walked carefully back to her chair.	*What did she trip on?*	A rug.	A banana peel.	A toy.	A cat.
14	With the recent rapid advancements in technology, many industries are now integrating artificial	expected unexpected expected_typo unexpected_typo	**intelligence** **opposition** **intellignece** **oppositoin**	to their workflows, to improve efficiency and reduce costs.	*What reason was given for updating workflows?*	Better efficiency.	Increased reliability.	Higher quality.	Better image.
15	In the geography class, the teacher explained the importance of conserving our planet's natural	expected unexpected expected_typo unexpected_typo	**resources** **movements** **resuorces** **movemnets**	to ensure sustainability for future generations.	*What was the subject of the class?*	Geography.	Mathematics.	History.	Language.
16	The environmental science lecture focused on the impact of greenhouse gas emissions on global	expected unexpected expected_typo unexpected_typo	**warming** **hunting** **wamring** **hutning**	and its effects on climate patterns.	*What action's impact did the lecture focus on?*	Gas emissions.	Deforestation.	Resource extraction.	Long-term climatic cycles.
17	In order to reduce traffic congestion and air pollution, the city council allocated funding to improve public	expected unexpected expected_typo unexpected_typo	**transportation** **architecture** **transpotration** **architetcure**	throughout the city, after a heated debate.	*Who allocated funds?*	The city council.	The national government.	The foundation.	The activists.
18	In the recent tech conference, the primary focus was on how the latest algorithms for machine	expected unexpected expected_typo unexpected_typo	**learning** **breaking** **leanring** **braeking**	can enhance data analysis and predictive modeling in various industries.	*What event was focused on these algorithms?*	The technology conference.	The trade show.	The governmental hearing.	The public lecture.
19	The United Nations summit this year is centered around the theme of sustainable	expected unexpected expected_typo unexpected_typo	**development** **newspapers** **develompent** **newspapres**	to balance economic growth with environmental protection.	*What organization held the summit?*	The United Nations.	The World Economic Forum.	The European Union.	The International Monetary Fund.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
20	With the increasing threat to endangered species, the report stressed the importance of wildlife	expected unexpected expected_typo unexpected_typo	**conservation** **tournaments** **consevration** **touranments**	and outlined policies meant to protect natural habitats.	*What did the report intend to protect?*	Endan-gered species.	Mineral resources.	National security.	Cultural diversity.
21	To reduce plastic waste, many companies are now shifting towards the use of biodegradable	expected unexpected expected_typo unexpected_typo	**materials** **sciences** **materails** **sciecnes**	in packaging to lessen environmental impact.	*Who is changing behavior?*	Companies.	Individuals.	Govern-ments.	Schools.
22	The research institute received a substantial grant for studying Alzheimer's, Parkinson's, and other neurological	expected unexpected expected_typo unexpected_typo	**disorders** **resources** **disodrers** **resuorces**	with a focus on developing new treatments.	*Who received the grant?*	The research institute.	The company.	The doctor.	The professor.
23	Before the judge makes the final decision on sentencing, it is important to consider any extenuating	expected unexpected expected_typo unexpected_typo	**circumstances** **associations** **circumtsances** **associaitons**	which may give justification for the defendant's alleged actions.	*Who is described as making the final sentencing decision?*	The judge.	The jury.	The prosecutor.	The defense attorney.
24	The team of astronomers looked at the distant nebula using an array of powerful	expected unexpected expected_typo unexpected_typo	**telescopes** **crystals** **telecsopes** **crytsals**	to uncover the mysteries of the early universe.	*What did the astronomers investigate?*	A distant nebula.	A nearby galaxy.	The solar system.	Exoplanets.
25	After four decades, Mary finally reconnected with and then eventually married her high-school	expected unexpected expected_typo unexpected_typo	**sweetheart** **physician** **sweethaert** **physicain**	who had remained a trusted friend and confidant over the intervening years.	*Roughly how long had it been since Mary was in high school?*	Forty years.	Three years.	One year.	Ten years.
26	As a new member of the United States Congress, he understood his main responsibility was to introduce new	expected unexpected expected_typo unexpected_typo	**legislation** **curricula** **legisaltion** **curriucla**	to promote the interests of his constituents.	*What body was he a new member of?*	The US Congress.	The Board of Governors.	The Parliament.	The Ad-ministrative Council.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
27	In the physics seminar, the professor explained how different particles vibrate at different	expected unexpected expected_typo unexpected_typo	**frequencies** **hurricanes** **frequecnies** **hurricnaes**	in accordance with their wave-like nature.	*What was the subject of the seminar?*	Physics.	Mathematics.	History.	Geography.
28	Most families are unable to afford to send their children to private schools without financial aid or prestigious	expected unexpected expected_typo unexpected_typo	**scholarships** **negotiations** **scholasrhips** **negotiaitons**	that are difficult to obtain.	*According to the passage, what are most families unable to afford?*	Private school.	Private academic tutors.	Educational materials.	Transportation costs.
29	At most dentists' offices you cannot simply walk in and get your teeth cleaned; you have to have made an	expected unexpected expected_typo unexpected_typo	**appointment** **exhibition** **appoinmtent** **exhibtiion**	there previously, often weeks or months earlier.	*What type of dentistry service was described?*	Teeth cleaning.	Cosmetic dental procedures.	Dental surgery.	Orthodontic braces fitting.
30	Due to a complicated medical condition, he had to use multiple prescription	expected unexpected expected_typo unexpected_typo	**medications** **cigarettes** **medicatoins** **cigartetes**	daily to manage his symptoms effectively.	*Why did he have the described daily regimen?*	As a result of a medical condition.	Because he was participating in a trial.	Because of the high cost of medical care.	As a personal choice.
31	When it was discovered that the athletes were taking illegal drugs to enhance their	expected unexpected expected_typo unexpected_typo	**performance** **photographs** **perfomrance** **photorgaphs**	in elite competitions, it caused a major scandal.	*What caused the scandal?*	Elite athletes' use of illegal drugs.	A disagreement about competition results.	Mismanagement of the competition.	The compensation levels for the elite athletes.
32	These days many people use online banking to carry out simple financial	expected unexpected expected_typo unexpected_typo	**transactions** **testimonies** **transatcions** **testimnoies**	rather than going to a brick and mortar bank.	*According to the passage, what has replaced traditional visits to the bank for many people?*	Online banking.	Automated teller machines.	Financial advisors.	Postal check deposits.
33	In recent years, nonfiction films such as biopics or nature	expected unexpected expected_typo unexpected_typo	**documentaries** **championships** **documetnaries** **championhsips**	have become increasingly popular on streaming services.	*Where have these films recently become more popular?*	On streaming services.	In movie theaters.	On network television.	In physical media sales.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
34	The lower court was unable to rule on the case involving a dispute between international corporations, because international disputes were not within its	expected unexpected expected_typo unexpected_typo	**jurisdiction** **headquarters** **jurisditcion** **headquatrers**	requiring the matter to be taken to a higher judicial body.	*What was the subject of the case that the court was unable to rule on?*	A dispute between international corporations.	A dispute between local businesses.	An internal governmental policy issue.	A disagreement over environmental regulations.
35	When I was a child, I always liked to visit my grandfather in the tiny apartment he shared with my	expected unexpected expected_typo unexpected_typo	**grandmother** **photographer** **grandmohter** **photograhper**	in the outskirts of the capital.	*When did the visits to take place?*	When the narrator was a child.	During the narrator's adulthood.	In the grandfather's early years.	Recently.
36	In a business negotiation, you are unlikely to get exactly what you want, so you will usually have to settle for some kind of	expected unexpected expected_typo unexpected_typo	**compromise** **suspicion** **comrpomise** **susipcion**	and remain flexible in your discussions.	*What is the context of negotiation described?*	Business negotiation.	Family dispute.	Political negotiation.	Academic debate.
37	In modern politics, there is a danger that military interests will use their power and	expected unexpected expected_typo unexpected_typo	**influence** **libraries** **infleunce** **libraires**	to shape government and public policy.	*What kind of interest does the passage warn about?*	Military.	Economic.	Technological.	Environmental.
38	Glancing in the shop window, the tourists saw the faces of movie stars printed on the covers of glossy	expected unexpected expected_typo unexpected_typo	**magazines** **vehicles** **magaznies** **vehilces**	prominently displayed facing the street.	*Whose faces did the tourists see?*	Film stars.	Famous authors.	Directors.	Historical figures.
39	The consulting company advertised that they would invest passion and talent to consistently exceed their clients'	expected unexpected expected_typo unexpected_typo	**expectations** **governments** **expetcations** **govenrments**	in delivering innovative solutions	*What type of company advertised?*	A consulting agency.	A publisher.	A tech startup.	A manufacturer.
40	In today's business forum, the main topic of discussion focused on the effects of increasing governmental	expected unexpected expected_typo unexpected_typo	**regulation** **surprises** **regluation** **surpirses**	and how to get around them without violating legal boundaries.	*Where did the discussion take place?*	A business forum.	A local community meeting.	A hearing.	A university lecture.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
41	The museum houses many famous pieces of art, including watercolor and oil	expected unexpected expected_typo unexpected_typo	**paintings** **chickens** **paitnings** **chikcens**	from renowned artists in a special collection.	*What types of art does the museum house, as mentioned in the passage?*	Watercolor and oil works.	Sculptures and ceramics.	Photographs and digital art.	Textile and woodcraft.
42	Following the election, there were major policy and staffing changes across multiple federal government	expected unexpected expected_typo unexpected_typo	**departments** **restaurants** **departemnts** **restuarants**	in order to project an image of reform.	*When were the major changes made?*	After the election.	Before the budget announcement.	At the start of the fiscal year.	During the summit.
43	Intervening in a complex ecological system without fully understanding its dynamics almost invariably leads to unintended unexpected	expected unexpected expected_typo unexpected_typo	**consequences** **achievements** **conseqeunces** **acheivements**	affecting both biodiversity and the environment.	*What type of system was the passage about?*	A complex ecological system.	A complicated economic system.	A computer operating system.	An educational system.
44	The common conception of political alignment makes a binary contrast between liberal and	expected unexpected expected_typo unexpected_typo	**conservative** **unidentified** **consrevative** **unidnetified**	positions on most prominent topics.	*What is being contrasted in the passage?*	Political alignments.	Scientific hypotheses.	Rhetorical techniques.	Personality types.
45	For their second date, the couple went to the city's historic district, where they got dinner in a fancy	expected unexpected expected_typo unexpected_typo	**restaurant** **classroom** **restuarant** **clasrsoom**	near the old town square.	*Where in the city did the couple go?*	The historic district.	A city park.	A suburban shopping mall.	The business district.
46	Cast iron cookware has generally been manufactured using the same techniques since the industrial	expected unexpected expected_typo unexpected_typo	**revolution** **gentlemen** **revolutoin** **gentelmen**	introduced methods for pouring liquid metal into sand molds.	*What item's production methods were discussed?*	Cookware.	Automobile parts.	Machinery components.	Building materials.
47	During the traditional music concert, the soloist played a rare, antique musical	expected unexpected expected_typo unexpected_typo	**instrument** **vegetable** **insrtument** **vegetbale**	that captivated the audience with its unique sound.	*What type of concert was it?*	Traditional music.	Pop music.	Classical orchestral music.	Choral music.

Table B.1: Stimuli and comprehension questions. *(continued)*

item	pretarget	target_type	target	posttarget	question	correct	incorrect0	incorrect1	incorrect2
48	At my annual review, the boss gave feedback that wasn't just negative commentary, instead it was more of a constructive	expected unexpected expected_typo unexpected_typo	**criticism** **disaster** **criticsim** **disatser**	aimed at changing my behavior on the project.	*What type of review did the person receive from the boss?*	Annual.	Quarterly.	Biannual.	Monthly.
49	In this exhibition, the artwork's abstract nature leaves it open to multiple different	expected unexpected expected_typo unexpected_typo	**interpretations** **disappoint-ments** **interpertations** **disappoitn-ments**	in a manner that will be uniquely personal to the viewer.	*What did the exhibition comprise of?*	Abstract art.	Classical sculpture.	Antique cars.	Modern technology.
50	The journalist found a job overseas, working as a foreign	expected unexpected expected_typo unexpected_typo	**correspondent** **revolutionary** **correpsondent** **revolutoinary**	for a national newspaper back home.	*Who found a job overseas?*	The journalist.	The banker.	The artist.	The publisher.

Table B.2: Practice stimuli and comprehension questions.

			answer choices			
item	sentence	question	correct	incorrect0	incorrect1	incorrect2
practice1	The driver waited until all the passengers were aboard before starting the engine.	*Who waited for the passengers?*	The driver.	The mechanic.	The attendant.	The conductor.
practice2	When the journlaists arrived, the editor removed the documents from the table.	*What was removed from the table?*	Some documents.	The plates.	Some cups.	The folders.
practice3	The story began on Halloween, all the street lights had gone out, and there was no moon, so it was dark and starry on the cloudless night.	*What did the passage say about the street lights?*	They had gone out.	The passage said nothing in particular about them.	They were flickering.	They were bright.
practice4	The spectators' cheering grew increasnigly loud as the swimmers approached the finish line.	*What were the swimmers approaching?*	The end of the race.	The halfway mark	The leader.	The spectators.

C

Supplemental material for chapter 4

C.1 CPMI-dependency implementation details

C.1.1 Word2Vec as noncontextual PMI control

We use Word2Vec (Mikolov, Sutskever, et al., 2013) to obtain a non-conditional PMI measure as a control/baseline. Additionally, in contrast with the CPMI values extracted from contextual language models, this estimate does not take into account the positions of the words in a particular sentence, but otherwise reflects global distributional information similarly to the contextualized models. Word2Vec should therefore function as a control with which to compare the PMI estimates derived from the contextualized models.

Word2Vec maps a given word w_i in the vocabulary it to a 'target' embedding vector $\mathbf{w}_i$, as well as an 'context' embedding vector $\mathbf{c}_i$ (used during training). As demonstrated by O. Levy and Goldberg (2014) and Allen and Hospedales (2019), Word2Vec's training objective is optimized when the inner product of the target and context embeddings equals the PMI, shifted by a global constant (determined by k, the number of negative samples): $\mathbf{w}_i^\top \mathbf{c}_j = \text{pmi}(w_i; w_j) - \log k$. This type of embedding model thus provides a non-contextual PMI estimator. A global shift will not change the resulting PMI-dependency trees, so we simply take $\text{pmi}_{\text{w2v}}(w_i; w_j) \coloneqq \mathbf{w}_i^\top \mathbf{c}_j$, with embeddings calculated using a Word2Vec model trained on the same data as BERT.[1] Note: since we are ignoring the global shift of k, an absolute valued version of PMI estimate will not be meaningful, and for this reason we only ever extract dependencies from the Word2Vec PMI estimate without taking

[1] We use the implementation in *Gensim* (Řehůřek & Sojka, 2010), trained on BookCorpus and English Wikipedia, and use a global average vector for out-of-vocabulary words.

the absolute value.

C.1.2 LtoR-CPMI for one-directional models

Our CPMI measure as defined above requires a bidirectional model (to calculate probabilities of words given their context, both preceding and following). The LSTM models we test in this study are left-to-right, so we define an slightly modified version of CPMI, to use with such unidirectional language models. That is, for a left to right model M_{LtoR}

$$\begin{aligned}&\text{CPMI}_{M_{\text{LtoR}}}(\mathbf{w}_I; \mathbf{w}_J) = \\ &\log \frac{p_{M_{\text{LtoR}}}(\mathbf{w}_I \mid \mathbf{w}_{0:I-1})}{p_{M_{\text{LtoR}}}(\mathbf{w}_I \mid \mathbf{w}_{0:J-1,J+1:I-1})},\end{aligned} \tag{C.1}$$

where $\mathbf{w}_{0:I-1}$ is the sentence up to before $\mathbf{w}_I$, and $\mathbf{w}_{0:J-1,J+1:I-1}$ is the sentence up to before$\mathbf{w}_I$, with $\mathbf{w}_J$ masked.

C.1.3 Calculating CPMI scores

C.1.3.1 Subtokenization

We must formulate the CPMI measure between sequences of subtokens, rather than tokens (words), because the large pretrained language models we use break down words into subtokens, for which gold dependencies and part of speech tags are not defined.

The calculation of CPMI between two lists of subtokens $\mathbf{w}_I$ and $\mathbf{w}_J$ in sentence $\mathbf{w}$ is

$$\begin{aligned}&\text{CPMI}_M(\mathbf{w}_I; \mathbf{w}_J) = \\ &\log \frac{p_M(\mathbf{w}_I \mid \mathbf{w}_{-I,J}, \mathbf{w}_J)}{p_M(\mathbf{w}_I \mid \mathbf{w}_{-I,J})} = \log \frac{p_M(\mathbf{w}_I \mid \mathbf{w}_{-I})}{p_M(\mathbf{w}_I \mid \mathbf{w}_{-I,J})}\end{aligned} \tag{C.2}$$

where I and J are spans of (sub)token indices, $\mathbf{w}_I$ is the set of subtokens with indices in I (likewise for $\mathbf{w}_J$), $\mathbf{w}_{-I}$ is the entire sentence without subtokens whose indices are in I, and $\mathbf{w}_{-I,J}$ is the sentence without subtokens whose indices are in I or J.

Likewise, POS-CPMI is defined in terms of subtokens. Note that gold POS tags are defined for PTB word tokens, which may correspond to multiple subtokens. POS-CPMI is calculated as:

$$\text{POS-CPMI}_M(\pi_I; \pi_J) = \log \frac{p_{M_{\text{POS}}}(\pi_I \mid \mathbf{w}_{-I})}{p_{M_{\text{POS}}}(\pi_I \mid \mathbf{w}_{-I,J})} \tag{C.3}$$

where M_{POS} is the contextual embedding model M with a POS embedding network on top, and π_I is the POS tag of $\mathbf{w}_I$ (the set of subtokens with indices in I, as in the definition of CPMI above).

To get the probability estimate for a multiple-subtoken word, we use a left-to-right chain

rule decomposition. To get an estimate for a probability $p(\mathbf{w})$ of a subtokenized word $\mathbf{w} = w_0, w_1, \ldots, w_n$ (that is, a joint probability, which we cannot get straight from a language model), we use a left-to-right chain rule decomposition of conditional probability estimates within the word:

$$p(\mathbf{w}) = p(w_0) \cdot p(w_1 \mid w_0) \cdots p(w_n \mid w_{0:n-1}) \tag{C.4}$$

This decomposition allows us to estimate conditional pointwise information between words made of multiple subtokens, at the expense of specifying a left-to-right order within those words.

C.1.3.2 Symmetrizing matrices

PMI is a symmetric function, but the estimated CPMI scores are not guaranteed to be symmetric, since nothing in the models' training explicitly forces their conditionaly probability estimates of words given context to respect the identity $p(x \mid y)p(y) = p(y \mid x)p(x)$. For this reason, we have a choice when assigning a score to a pair of words v, w, whether we use the model's estimate of $\text{CPMI}_M(v; w)$, which compares the probability of v with conditioner w masked and unmasked, or of $\text{CPMI}_M(w; v)$. In our implementation of CPMI we calculate scores in both directions, and use their sum (as mentioned in the main text §4.3.1), though experiments using one or the other (using just the upper or lower triangular of the matrix), or the max (equivalent to extracting a tree from the unsymmetrized matrix) led to very similar overall results. Likewise for the Word2Vec PMI estimate, and the POS-CPMI estimates.

C.1.3.3 Negative PMI values

PMI may be positive or negative. Results in the main text are all computed for CPMI dependencies extracted from signed matrices (so arcs with large negative CPMI will be rarely included). However, there is some discussion of interpreting the magnitude of PMI as indicating dependency, independent of sign (see Salle & Villavicencio, 2019). The choice to use an absolute-valued version of CPMI might be justified by arguing that words which influence each other's distribution should be connected, whether this influence is positive or negative.

In §C.4.1 we include full results both with and without taking the absolute value of the CPMI matrices before extracting trees. The absolute-valued CPMI dependencies show a models increase in UUAS over the corresponding matrices without taking the absolute value in general. But, it is not clear whether the choice to use absolute-valued CPMI would be justified conceptually. Contrary to the conceptual motivation for CPMI dependencies, in which words which often occur together should be linked, an absolute-valued version links words which are highly informative of each others' *not* being present. For this reason we do not choose to use an absolute-valued version of CPMI by default, but report those results for comparison, note that the UUAS is in fact higher with the absolute value, and refrain from further speculation.

C.1.4 Additional analysis of CPMI dependencies

C.1.4.1 Similarity between models

Figure C.3 shows the similarity of the CPMI dependency structures extracted from the different contextual embedding models. We measure similarity of dependency structures with the Jaccard index for the sets of the predicted edges by two models. Jaccard index measures similarity of two sets A, B and is defined as $J(A, B) = |A \cap B| / |A \cup B|$. The contextualized models agree with each other on around 30–50% of the edges, and agree with the the noncontextual baseline W2V slightly less. In general, they agree with the linear baseline at somewhat higher rates.

C.1.4.2 Accuracy versus arc length

Breaking down the results by dependency length, Figure C.1 shows the recall accuracy of CPMI dependencies, grouped by length of gold arc. In general, length 1 arcs have the highest accuracy; longer dependencies have lower accuracy. CPMI dependencies from BERT (large) have 81% recall accuracy on length 1 arcs, with arcs longer than 1 having much lower recall (13% overall) near random (10%). In other models, XLNet in particular, this distinction is less of a binary distinction, but the trend is still for lower recall on longer arcs.

C.1.4.3 Accuracy versus perplexity

Here we investigate the correlation between language model performance and CPMI-dependency accuracy. If models' confidence in predicting were tied to accuracy, it would be hard to argue that the relatively low accuracy score we see was due to the lack of connection between syntactic dependency and statistical dependency, rather than to the models' struggling to recover such a structure. Here we measure model confidence by obtaining a perplexity score for each sentence, calculated as the negative mean of the pseudo log-likelihood, that is, for a sentence $\mathbf{w}$ of length n,

$$\text{pseudo PPL}(\mathbf{w}) = \exp\left[-\frac{1}{n}\sum_{I=1}^{n}\log p(\mathbf{w}_I \mid \mathbf{w}_{-I})\right] \tag{C.5}$$

Figure C.2 shows that accuracy is not correlated with sentence-level perplexity for any of the models (fitting a linear regression, $R^2 < 0.05$ for each model). That is, the accuracy of CPMI-dependency structures is roughly the same on the sentences which the model predicts confidently (lower perplexity) as on the sentences which it predicts less confidently (higher perplexity).

C.1.4.4 UUAS during training

We examined the accuracy of CPMI dependencies during training of BERT (base uncased) from scratch. Figure C.4 shows the average perplexity of this model, along with the sentence-wise average accuracy of CPMI structures at selected checkpoints during training. After about one

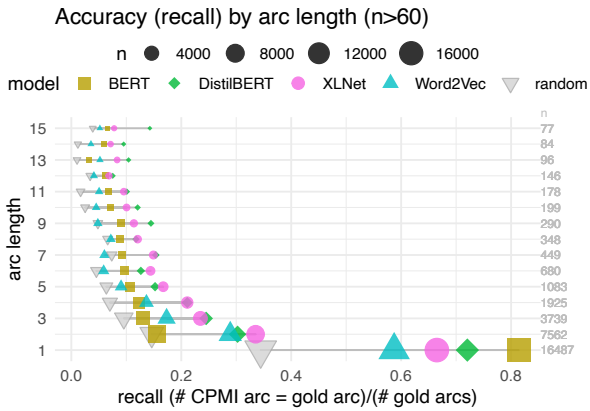


Figure C.1: Recall accuracy is higher for shorter arcs. The distinction is mostly between arcs of length 1 vs longer arcs. Note that the relatively higher accuracy of BERT (large)'s estimates overall are driven by its very large proportion of length 1 arcs.

million training steps the model has reached a plateau in terms of performance (perplexity stops decreasing), and we see that the peak UUAS has also plateaued at that point, but in fact reached its highest value after one hundred thousand training steps.

C.1.4.5 UUAS by dependency label

Table C.1 gives per-dependency label recall accuracy of CPMI-dependencies extracted from the subset of dependency labels for which XLNet (base) achieves UUAS higher than both the linear and a random (projective) baselines.

C.2 Information Bottleneck for POS probe

The simple POS probe is a d-by-h-matrix, where the input dimension h is the contextual embedding network's hidden layer dimension, and the output dimension d is the number of different POS tags in the tagset. Interpreting the output as an unnormalized probability distribution over POS tags, we train the layer to minimize the cross-entropy loss between the predicted and observed POS (using the labels from the Treebank). Training a simple linear probe is a rough way to get a compressed representations from contextual embeddings, but it has limitations (Hewitt & Liang, 2019).

A more correct way of extracting these representations is by a variational information bottleneck technique Tishby et al., 2000. We implement this technique (roughly following Li & Eisner, 2019), as follows. Optimization is to minimize $\mathcal{L}_{\text{IB}} = -\text{I}(Y;Z) + \beta\,\text{I}(H;Z)$, where H is the input embedding, Z the latent representation and Y the true label, and $\text{I}(\cdot;\cdot)$ is mutual information. This technique trains two sets of parameters: the decoder, a linear model just as in the simple linear POS probe, and the encoder, another linear model, whose output in our case is interpreted as means and log-variances of a multivariate Gaussian (a simplifying assumption). Minimizing this

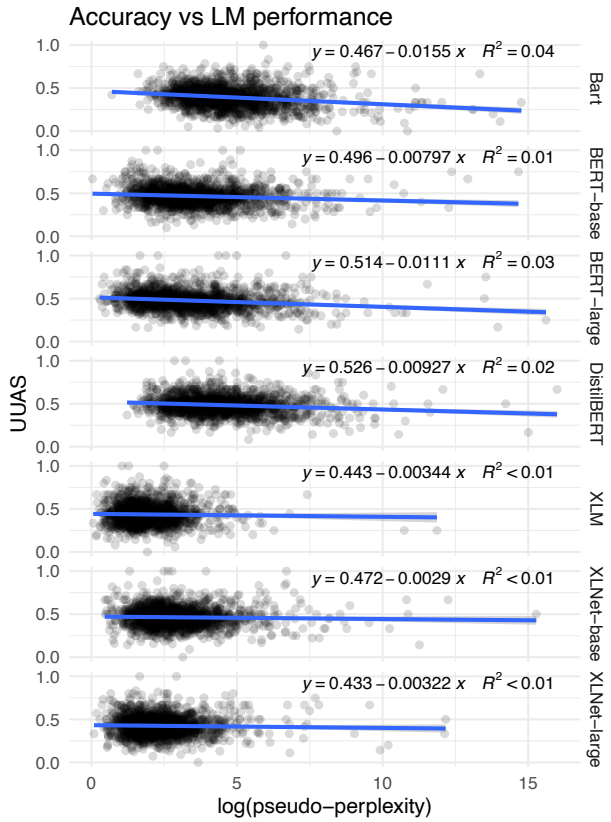


Figure C.2: Per-sentence accuracy (UUAS) against log psuedo-perplexity. Accuracy is not tied to the confidence of the language model on a given sentence, for any of the models (there is a slight tendency to have higher accuracy on sentences of lower perplexity, but the effect size is negligible, and correlation is very low).

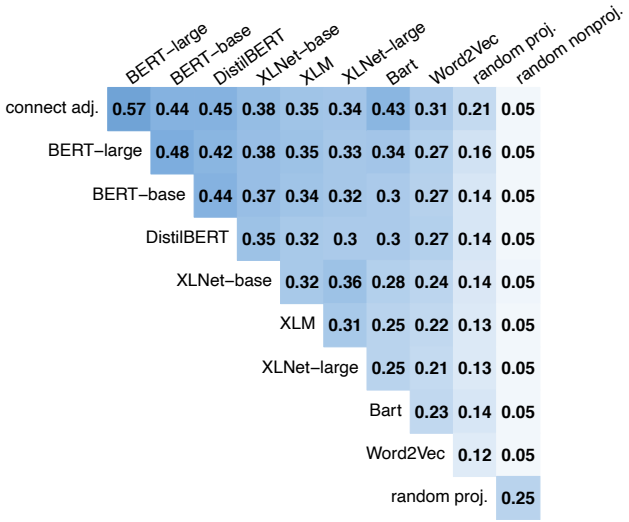


Figure C.3: Similarity of models' predictions, by wordpair, reported as Jaccard index, the intersection of the two models' sets of dependency edges divided their union.

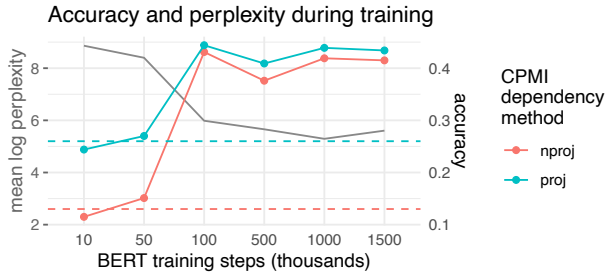


Figure C.4: Training checkpoints for BERT-base uncased. After about 1 million training steps, the perplexity (gray, axis left) has plateaued. The UUAS (axis right) of extracted CPMI structures does not increase past the level it reaches at 100k steps.

relation	mean length	n	BERT	Distil-BERT	Bart	XLNet	XLM	W2V	connect-adjacent	random projective
xcomp	3.1	398	0.24	0.23	0.18	**0.43**	0.40	0.26	0.07	0.13
mark	5.0	421	0.18	0.29	0.11	**0.30**	0.20	0.09	0.05	0.10
conj	6.1	1009	0.12	0.19	0.21	0.28	0.26	**0.29**	0.03	0.10
ccomp	6.9	550	0.11	0.15	0.07	**0.19**	0.14	0.06	0.03	0.08
dobj	2.4	1637	0.37	0.38	0.33	**0.47**	0.42	0.35	0.21	0.16
advcl	8.7	293	0.05	0.04	0.05	**0.11**	0.07	0.06	0.00	0.06
nsubjpass	4.3	253	0.13	0.15	0.12	0.21	**0.26**	0.19	0.00	0.13
rcmod	4.1	290	0.11	0.07	0.12	0.12	**0.14**	0.11	0.00	0.08
poss	2.4	709	0.30	0.28	0.21	**0.32**	0.31	0.30	0.24	0.17
pobj	2.3	3745	0.33	**0.39**	0.28	0.36	0.32	0.30	0.30	0.17
tmod	3.0	244	0.31	0.35	0.30	0.39	**0.40**	0.18	0.33	0.18
cop	2.1	330	0.39	**0.49**	0.39	0.42	0.33	0.33	0.39	0.22
det	1.7	3327	0.52	**0.64**	0.24	0.53	0.43	0.41	0.52	0.23

Table C.1: Recall accuracy by label for the labels which XLNet achieves above the baselines, for the models BERT large, Distilbert base, Bart large, XLNet base, XLM, as well as Word2Vec, and the connect adjacent and random baselines.

loss maximizes information in the compressed representations about the output labels given a constraint on the amount of information that the compressed representations carry about the original embeddings.

C.3 Equivalence of max pmi and max conditional probability objectives

Mareček (2012) describes the equivalence of optimizing for trees with maximum conditional probability of dependents given heads and optimizing for the maximum PMI between dependents and heads. This equivalence relies on an assumption that the marginal probability of words is independent of the parse tree.

For a corpus C, a dependency structure t can be described as a function which maps the index of a word to the index of its head. If net mutual information between dependents and heads according to dependency structure t is $\text{pmi}(t) \coloneqq \sum_i \text{pmi}(w_i; w_{t(i)})$, and the log conditional probability of dependents given heads is $\ell_{\text{cond}}(t) \coloneqq \prod_{w \in s} p(w_i \mid w_{t(i)})$, the optimum is the same:

$$\arg\max_t \text{pmi}(t) = \arg\max_t \log \prod_{i=1}^{|C|} \frac{p(w_i, w_{t(i)})}{p(w_i)p(w_{t(i)})} \tag{C.6}$$

$$= \arg\max_t \log \prod_{i=1}^{|C|} \frac{p(w_i, w_{t(i)})}{p(w_{t(i)})} \tag{C.7}$$

$$= \arg\max_t \ell_{\text{cond}}(t) \tag{C.8}$$

The step taken in (C.7) follows only under the assumption that the marginal probability of dependent words is independent of the structure t. That is, that "probabilities of the dependent words ... are the same for all possible trees corresponding to a given sentence" (Mareček, 2012, §5.1.2). This must be stipulated as an assumption in a probabilistic model for the above derivation to hold.

C.4 Augmented tables of results

We give results in further detail for the CPMI-dependencies on the English PTB Wall Street Journal (WSJ) and on the multilingual PUD treebanks. Tables described below follow this appendix.

C.4.1 Results on WSJ data

Results presented in this section repeat those given in the main text, with two independent additional parameters: projectivity and absolute value.

Projectivity As described in §4.3.1, in the main text we report results for projective CPMI dependency trees extracted from CPMI matrices using Eisner's algorithm J. M. Eisner (1996) and J. Eisner (1997). These results are also repeated below, but we additionally present UUAS results for maximum spanning trees (MSTs) extracted from CPMI matrices using Prim's algorithm (Prim, 1957), following Hewitt and Manning (2019).

Absolute value In the main text we consider dependencies extracted from signed CPMI matrices. As described in §C.1.3.3, we also compute UUAS from absolute-valued matrices, and report them here.

- Table C.2 is an augmented version of Table 4.1 from the main text, containing results for CPMI-dependencies both with and without the projectivity constraint.

- Table C.3 is as the previous, but using an absolute valued version of CPMI.
- Table C.6 is likewise an augmented version of Table 4.3 from the main text, containing results for POS-CPMI-dependencies both with and without the projectivity constraint.
- Table C.7 is as the previous but using an absolute valued version of POS-CPMI.

In these tables, we also include the UUAS of randomized 'lengthmatched' control. For each sentence, this control consists of a randomized tree whose distribution of arc lengths is identical to the gold tree (obtained by rejection sampling).

C.4.1.1 WSJ10

Tables C.4 and C.5 give augmented UUAS results as in to Tables C.2 and C.3, resp., but for only the sentences of length ≤ 10 from the test split (section 23) of the WSJ corpus (WSJ10). We include these results for comparison with much of the unsupervised dependency parsing literature following Klein and Manning (2004), which reports results on that subset. Note that the UUAS is naturally higher across the board on this corpus of shorter sentences.

C.4.2 Results on multilingual PUD data

Table C.5 gives results on the 20 languages of the Parallel Universal Dependencies (PUD) treebanks. These parallel treebanks were included in the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies. The PUD treebank for each language consists of 1000 sentences annotated for Universal Dependencies. The sentences are translated into each of the languages, with the majority (750) being originally in English.

We compute CPMI for these sentences using the multilingual pretrained BERT-base model made available by Hugging Face Transformers (Wolf et al., 2020).[2] This model was trained using masked language modelling and next sentence prediction on the 104 languages with the largest Wikipedias, including all 20 in the PUD. UUAS for CPMI dependency trees for all languages is plotted in Figure C.6.

[2] https://huggingface.co/bert-base-multilingual-cased

		UUAS						
			MSTs			Projective MSTs		
				CPMI			CPMI	
language	mean sent. len	connect-adjacent	rand	(signed)	(abs)	rand	(signed)	(abs)
Arabic	17.52	.58	.11	.43	.48	.27	.45	.51
Chinese	17.51	.45	.11	.38	.39	.23	.40	.42
Czech	14.99	.48	.12	.47	.48	.25	.48	.50
English	17.73	.42	.10	.41	.43	.22	.43	.45
Finnish	12.47	.52	.15	.45	.46	.28	.47	.48
French	21.18	.45	.08	.44	.46	.23	.47	.49
German	17.56	.42	.11	.44	.46	.22	.46	.48
Hindi	20.53	.51	.09	.38	.39	.24	.41	.42
Icelandic	15.88	.49	.12	.40	.41	.25	.42	.44
Indonesian	16.06	.56	.12	.44	.46	.27	.46	.49
Italian	20.43	.45	.09	.45	.46	.23	.47	.48
Japanese	24.73	.48	.08	.30	.39	.23	.34	.43
Korean	13.99	.58	.13	.46	.48	.28	.49	.50
Polish	14.73	.54	.12	.50	.51	.27	.52	.53
Portuguese	19.83	.45	.10	.44	.46	.23	.47	.48
Russian	15.38	.51	.12	.49	.50	.26	.51	.51
Spanish	20.00	.45	.09	.46	.47	.23	.48	.50
Swedish	16.14	.44	.11	.41	.43	.24	.43	.45
Thai	21.05	.56	.09	.39	.38	.25	.42	.42
Turkish	13.73	.55	.14	.46	.48	.27	.48	.50

Figure C.5: UUAS for multilingual Parallel UD dataset, for CPMI dependencies extracted from from BERT base multilingual. Note that while the dataset consists of the same 1000 sentences translated into the 20 languages, there is some variation across languages in mean sentence length. Projective (signed) UUAS are plotted in fig. C.6 with random and connect-adjacent baselines.

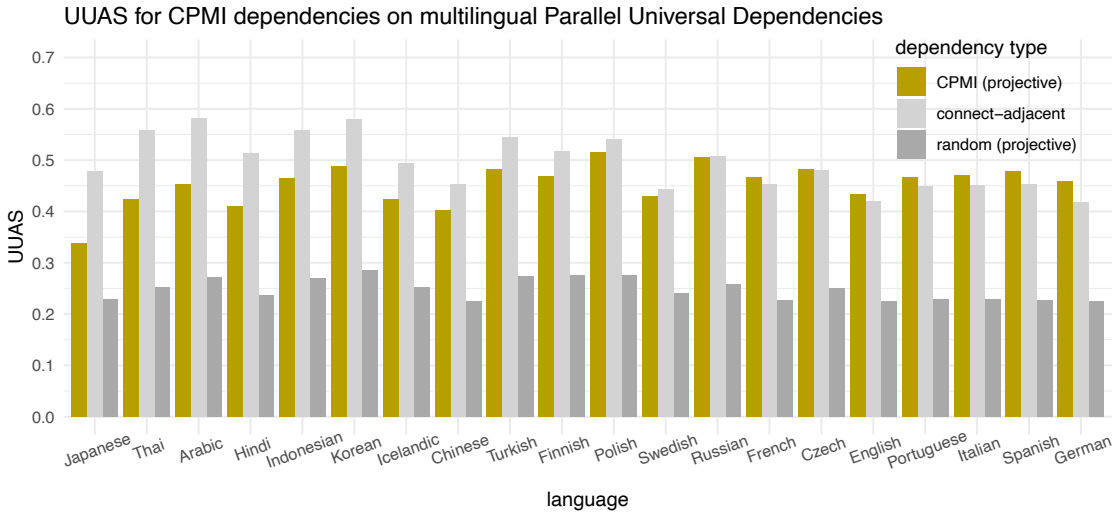


Figure C.6: CPMI UUAS (signed, projective) from BERT base multilingual, ordered by the difference between CPMI UUAS and the connect-adjacent baseline UUAS. For most languages the CPMI UUAS is below or comparable to the connect-adjacent baseline.

	MSTs			Projective MSTs		
	all	len = 1	len > 1	all	len = 1	len > 1
		prec. | recall	prec. | recall		prec. | recall	prec. | recall
random	.09	.49 | .10	.05 | .09	.22	.49 | .34	.08 | .10
connect-adjacent	**.49**	.49 | 1	– | 0	**.49**	.49 | 1	– | 0
lengthmatched	.37					
Word2Vec	.27	**.67** | .36	.13 | .19	.39	**.61** | .59	.19 | .19
BERT base	.44	.59 | .68	.26 | .22	.46	.57 | .72	.27 | .21
BERT large	.46	.56 | **.79**	.23 | .14	.47	.55 | **.81**	.24 | .13
DistilBERT	.46	.58 | .68	**.30** | .25	.48	.57 | .72	**.32** | .24
Bart large	.36	.53 | .60	.15 | .14	.38	.52 | .64	.16 | .13
XLM	.38	.64 | .55	.20 | .22	.42	.60 | .64	.23 | .22
XLNet base	.42	.61 | .59	.25 | **.26**	.45	.59 | .66	.29 | **.25**
XLNet large	.36	.63 | .51	.19 | .22	.41	.59 | .61	.23 | .22
vanilla LSTM	.40	.56 | .60	.23 | .22	.44	.54 | .70	.26 | .19
ONLSTM	.41	.57 | .61	.23 | .22	.44	.55 | .71	.27 | .19
ONLSTM-SYD	.41	.57 | .61	.23 | .22	.45	.55 | .71	.27 | .19

Table C.2: Total UUAS on the WSJ data, for CPMI dependencies extracted by both with a simple MST (Prim's algorithm; left) with a projectivity constraint (Eisner's algorithm; right, repeating Table 4.1). In each case, overall scores are in the first column, followed by precision and recall UUAS for the subset consisting only of adjacent words (len = 1), and likewise for nonadjacent words (len > 1).

	MSTs			Projective MSTs		
	all	len = 1	len > 1	all	len = 1	len > 1
		prec. | recall	prec. | recall		prec. | recall	prec. | recall
BERT base	.48	.60 | .75	.29 | .22	.49	.59 | .78	.31 | .21
BERT large	.48	.56 | **.84**	.25 | .13	.48	.56 | **.86**	.26 | .13
DistilBERT	.48	.58 | .73	**.32** | .25	.50	.58 | .77	**.35** | .24
Bart large	.38	.55 | .59	.19 | .17	.40	.54 | .64	.20 | .16
XLM	.41	.65 | .59	.22 | .24	.44	**.63** | .67	.25 | .23
XLNet base	.44	.61 | .62	.27 | **.26**	.47	.60 | .70	.30 | **.25**
XLNet large	.37	.63 | .53	.19 | .23	.42	.61 | .62	.22 | .22
vanilla LSTM	.42	.55 | .63	.25 | .22	.45	.54 | .73	.28 | .18
ONLSTM	.42	.56 | .63	.25 | .22	.45	.54 | .73	.29 | .19
ONLSTM-SYD	.42	.56 | .64	.25 | .22	.46	.54 | .74	.29 | .19

Table C.3: As above in Table C.2, but with dependencies extracted from absolute-valued matrices. As noted in §C.1.1, due to the fact that Word2Vec estimates PMI only up to a global shift, an absolute-valued version would be meaningless, so we do not include that model here.

	MSTs			Projective MSTs		
	all	len = 1	len > 1	all	len = 1	len > 1
		prec. | recall	prec. | recall		prec. | recall	prec. | recall
random	.29	.56 | .30	.18 | .28	.34	.54 | .45	.18 | .21
adjacent	**.53**	.53 | 1	– | 0	**.53**	.53 | 1	– | 0
lengthmatched	.51					
Word2Vec	.42	**.61** | .51	.28 | .32	.46	**.60** | .63	.29 | .27
BERT base	.51	.60 | .69	.36 | .29	.52	.59 | .72	.38 | .28
BERT large	.52	.59 | **.81**	.34 | .20	**.53**	.59 | **.82**	.36 | .20
DistilBERT	.51	.59 | .71	**.38** | .29	.52	.58 | .75	**.40** | .27
Bart large	.44	.54 | .63	.27 | .21	.45	.54 | .66	.28 | .21
XLM	.48	**.61** | .61	.32 | .32	.49	**.60** | .66	**.34** | **.31**
XLNet base	.51	**.61** | .64	**.38** | **.35**	**.53**	**.60** | .69	.42 | .35
XLNet large	.46	**.61** | .57	.32 | .34	.48	.59 | .64	.34 | .31

Table C.4: Total UUAS on WSJ10, for CPMI dependencies extracted both without the projectivity constraint (MSTs), and with it (Projective MSTs). Compare with an overall UUAS of **.637** reported in Klein and Manning (2004, Fig. 3) for the complete WSJ10.

	MSTs			Projective MSTs		
	all	len = 1	len > 1	all	len = 1	len > 1
		prec. | recall	prec. | recall		prec. | recall	prec. | recall
BERT base	.53	.60 | .75	.39 | .28	.54	.60 | .78	.41 | .27
BERT large	.54	.60 | **.85**	.37 | .19	.54	.59 | **.86**	.38 | .19
DistilBERT	.54	.60 | .77	**.41** | .28	.55	.60 | .79	**.43** | .27
Bart large	.47	.58 | .63	.31 | .28	.48	.58 | .67	.33 | .27
XLM	.50	**.64** | .65	**.33** | **.32**	.51	**.63** | .69	.35 | **.31**
XLNet base	.52	**.62** | .68	.39 | .34	.55	.62 | .73	.42 | .34
XLNet large	.48	**.62** | .61	.33 | .34	.51	.61 | .66	.37 | .33

Table C.5: Total UUAS on WSJ10, MST and Projective MST, as above, but extracted from absolute-valued CPMI matrices.

		MSTs			Projective MSTs		
		all	len = 1	len > 1	all	len = 1	len > 1
			prec. | recall	prec. | recall		prec. | recall	prec. | recall
simple-POS	**BERT** base	.47	.57 | .77	.29 | .20	.48	.56 | .79	.32 | .19
	BERT large	.44	.54 | .73	.25 | .17	.45	.53 | .75	.27 | .16
	XLNet base	.29	.56 | .41	.14 | .17	.36	.55 | .56	.17 | .17
	XLNet large	.26	.59 | .38	.11 | .15	.32	.56 | .51	.14 | .15
IB-POS	**BERT** base	.38	.60 | .58	.18 | .18	.41	.58 | .65	.20 | .18
	BERT large	.39	.56 | .64	.17 | .14	.41	.55 | .69	.18 | .14
	XLNet base	.36	.57 | .52	.19 | .20	.40	.55 | .60	.22 | .20
	XLNet large	.30	.60 | .44	.13 | .17	.36	.56 | .56	.16 | .16

Table C.6: Total UUAS for POS-CPMI, both MST (left) and projective MST (right, a repeat of Table 4.3), using the simple POS probe and IB POS probe, from BERT and XLNet models.

		MSTs			Projective MSTs		
		all	len = 1	len > 1	all	len = 1	len > 1
			prec. | recall	prec. | recall		prec. | recall	prec. | recall
simple-POS	**BERT** base	.49	.57 | .78	.32 | .21	.50	.57 | .80	.34 | .21
	BERT large	.47	.56 | .79	.28 | .17	.48	.55 | .81	.30 | .16
	XLNet base	.31	.57 | .44	.15 | .18	.36	.56 | .56	.17 | .17
	XLNet large	.27	.59 | .40	.12 | .15	.31	.57 | .49	.13 | .14
IB-POS	**BERT** base	.35	.60 | .52	.16 | .18	.39	.59 | .61	.19 | .18
	BERT large	.40	.58 | .67	.17 | .15	.43	.57 | .72	.19 | .14
	XLNet base	.38	.58 | .56	.20 | .21	.42	.57 | .63	.23 | .21
	XLNet large	.30	.59 | .44	.13 | .16	.35	.57 | .55	.16 | .16

Table C.7: As above in Table C.6, but with dependencies extracted from absolute-valued matrices.

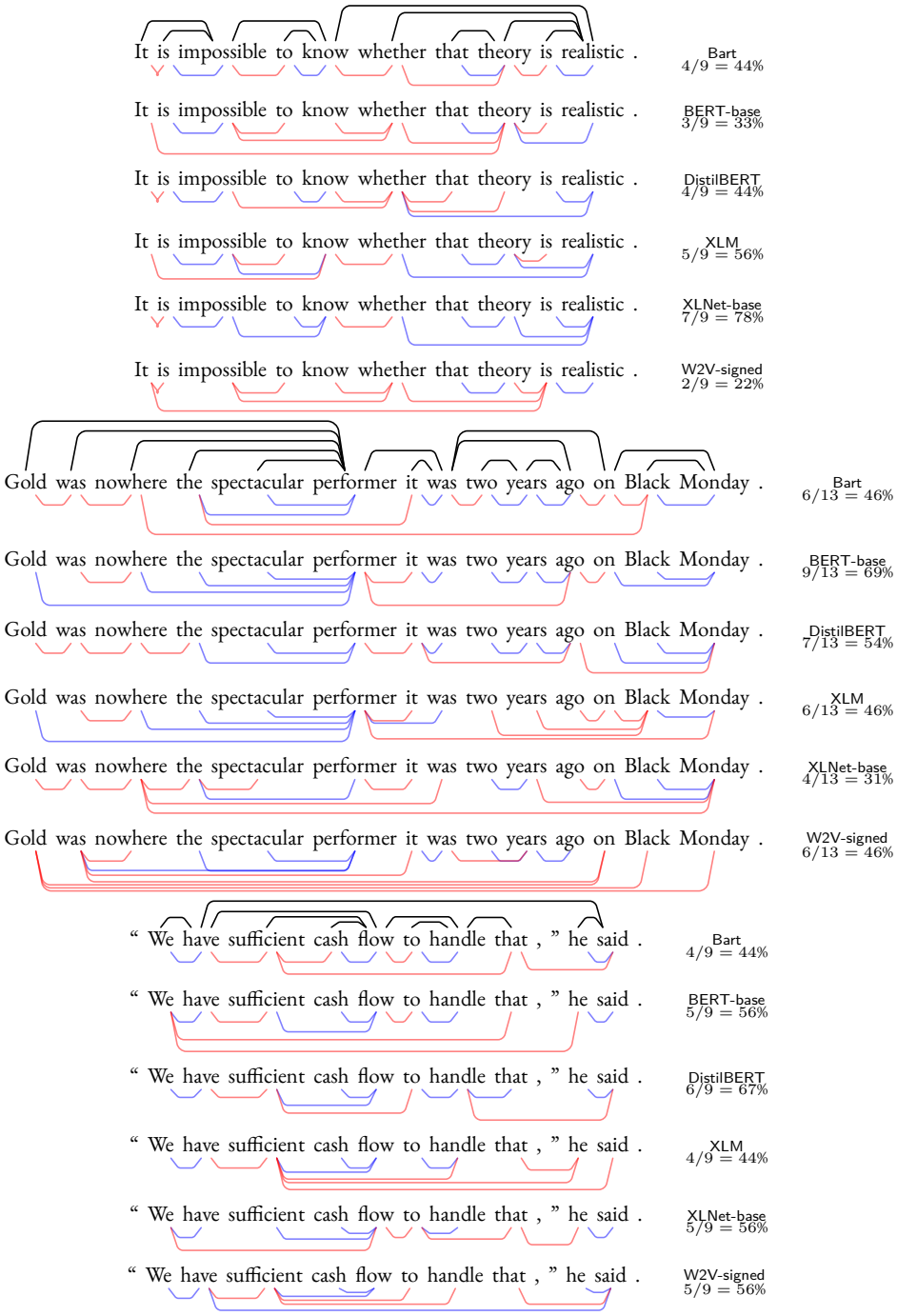


Figure C.7: Additional examples of projective parses from Bart, BERT, DistilBERT, XLM, XLNet, and the noncontextual baseline Word2Vec. Gold standard dependency parse above in black, CPMI-dependencies below, blue where they agree with gold dependencies, and red when they do not. Accuracy scores (UUAS) are given for each sentence.

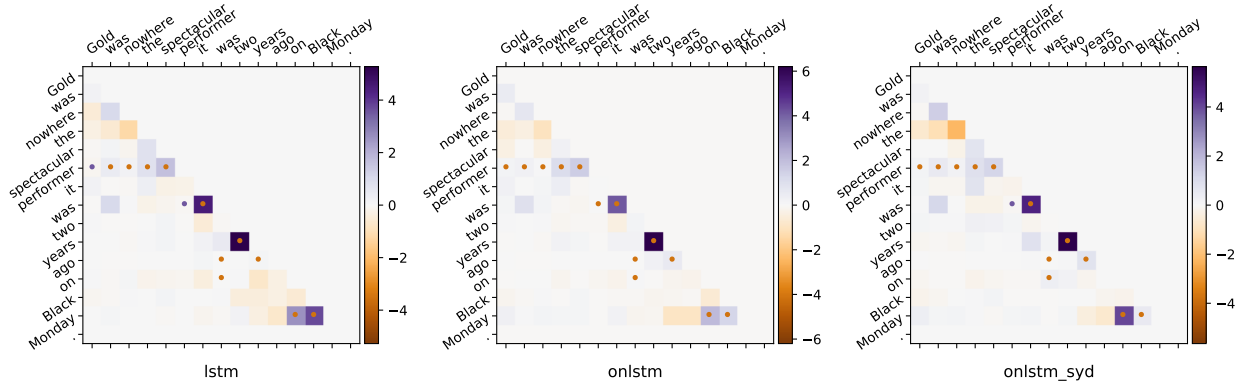


Figure C.8: CPMI matrices for ONLSTM and ONLSTM-SYD, with vanilla LSTM baseline. Gold edges are marked with a dot. Compare with dependency structures in Figure C.9

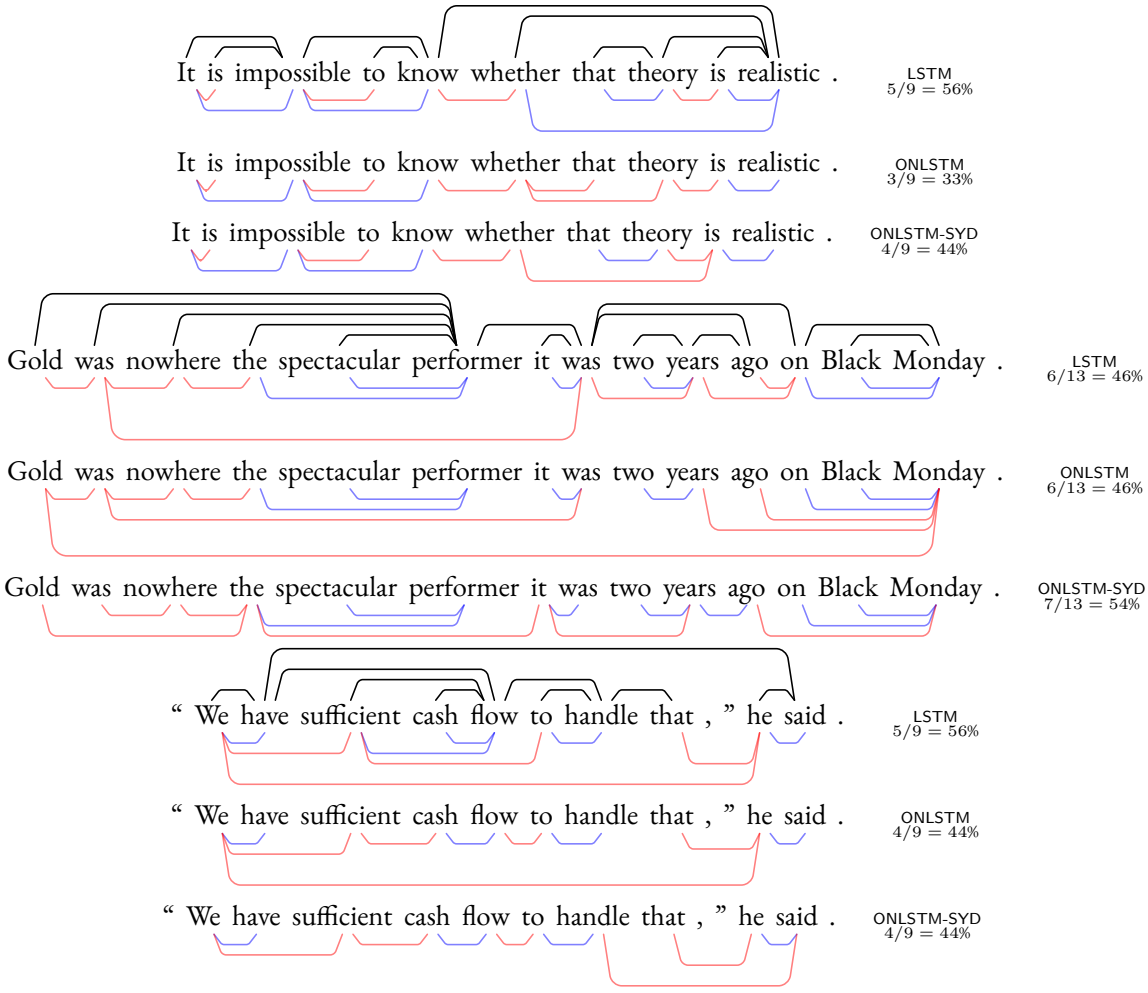


Figure C.9: Projective parses from the LSTM baseline and the ONSLTM and syntactic (ONSLTM-SYD) models for three example sentences. Matrices for the second sentence are in Figure C.8.