

Neural language models and human linguistic knowledge

by

Jennifer Hu

B.A., Harvard University (2018)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Cognitive Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2023

© 2023 Jennifer Hu. This work is licensed under a CC BY 4.0.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Author.....

Department of Brain and Cognitive Sciences

May 5, 2023

Certified by

Roger P. Levy

Professor

Thesis Supervisor

Accepted by

Mark T. Harnett

Graduate Officer, Department of Brain and Cognitive Sciences

Neural language models and human linguistic knowledge

by

Jennifer Hu

Submitted to the Department of Brain and Cognitive Sciences
on May 5, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Cognitive Science

Abstract

Language is one of the hallmarks of intelligence, demanding explanation in a theory of human cognition. However, language presents unique practical challenges for quantitative empirical research, making many linguistic theories difficult to test at naturalistic scales. Artificial neural network language models (LMs) provide a new tool for studying language with mathematical precision and control, as they exhibit remarkably sophisticated linguistic behaviors while being fully intervenable. While LMs differ from humans in many ways, the learning outcomes of these models can reveal the behaviors that may emerge through expressive statistical learning algorithms applied to linguistic input.

In this thesis, I demonstrate this approach through three case studies using LMs to investigate open questions in language acquisition and comprehension. First, I use LMs to perform controlled manipulations of language learning, and find that syntactic generalizations depend more on a learner’s inductive bias than on training data size. Second, I use LMs to explain systematic variation in scalar inferences by approximating human listeners’ expectations over unspoken alternative sentences (e.g., “The bill was supported overwhelmingly” implies that the bill was not supported unanimously). Finally, I show that LMs and humans exhibit similar behaviors on a set of non-literal comprehension tasks which are hypothesized to require social reasoning (e.g., inferring a speaker’s intended meaning from ironic statements). These findings suggest that certain aspects of linguistic knowledge could emerge through domain-general prediction mechanisms, while other aspects may require specific inductive biases and conceptual structures.

Thesis Supervisor: Roger P. Levy

Title: Professor

Acknowledgments

This thesis – and my entire Ph.D. journey – would not have been possible without the support of my mentors, collaborators, friends, and family. First, I am deeply grateful for the mentorship of my advisor, Roger Levy. Roger constantly pushed me to take ownership of my ideas and pursue the most ambitious theoretical questions. Beyond his incredible scientific acumen, I also greatly admire Roger’s dedication to open, accessible science, and his passion for forging connections across traditional disciplinary boundaries. It has been such a privilege to work with and learn from him.

I have also benefitted tremendously from the support of my committee members, Joshua Tenenbaum, Evelina Fedorenko, and Christopher Potts. Josh inspired new perspectives that helped me situate my ideas within the broader context of cognition and artificial intelligence. Ev also broadened my scientific vision by encouraging me to engage with problems in brain representation and computation. Chris offered incisive insights that consistently challenged my assumptions about linguistic theory and computational modeling. To Chris I am immensely grateful, as my experience working with him as an undergraduate research intern ultimately led me to pursue a Ph.D. in cognitive science.

I am so fortunate to have met incredible friends and colleagues during my time in the Computational Psycholinguistics Lab: Helena Aparicio, Yevgeni Berzak, Veronica Boyce, Canaan Breiss, Thomas Clark, Reuben Cohn-Gordon, Tiwa Eisape, Jon Gauthier, Matthias Hofer, Carina Kauf, Suman Maity, Stephan Meylan, Peng Qian, MH Tessler, Tristan Thrush, Ethan Wilcox, Noga Zaslavsky, and Meilin Zhan. I will very fondly remember our lunch conversations, spontaneous discussions about science, and general office antics. I also had the opportunity to work with a wonderful undergraduate research assistant, Irene Zhou, as well as many external collaborators: Sammy Floyd, Ev Fedorenko, Ted Gibson, Olessia Jouravlev, Sebastian Schuster, and Judith Degen. I am also grateful to Raj Ammanabrolu, Xiang Ren, Yejin Choi, and the rest of the Mosaic team at AI2, where I spent a summer thinking about text adventure games, Dungeons and Dragons, and interesting problems in reinforcement learning and language.

Before joining the Computational Psycholinguistics Lab, I had the privilege of rotating

in Josh McDermott's and Ev Fedorenko's labs. Many thanks to Josh, James Traer, Maddie Cusimano, and other members of the McDermott Lab for taking me under their wing for my first rotation in graduate school. It was my first foray into research that wasn't about language, and I learned an enormous amount about experiment design, psychophysics, and probabilistic modeling. Thank you to Ev, Hannah Small, Hope Kean, Atsushi Takahashi, Leo Zelekman, Daniel Kleinman, Elizabeth Ryan, Alfonso Nieto-Castañón, Victor Ferreira, Anna Ivanova, and other members of EvLab for making my second rotation such a rewarding experience – and for making EvLab feel like a second home. Ev gave me the opportunity to pursue my own research project as a rotation student, for which I am incredibly grateful. Thank you also to Rebecca Saxe and Jacob Andreas, who were also wonderful resources for getting feedback on my ideas and talking informally about science.

I have been part of a brilliant cohort of students who entered the BCS graduate program in 2018: Yuan Bian, Daniel Cho, Josefina Correa, Gabrielle Drummond, Mila Halgren, Max Heinrich, Mackenzie Lee, Madison Leet, João Loula, Cecilia Pellegrini, Francis Reilly-Andujar, Omar Rutledge, Sugandha Sharma, Sara Kornfeld Simpson, Katya Tsimring, Dimitra Vardalaki, Lio Wong, and Qianli Xu. I am grateful to all of them for navigating graduate school together. It has been an incredible privilege to work in Building 46, where neuroscientists study proteins and mouse brains next to cognitive scientists studying language models and moral reasoning. Being immersed in such an intellectually diverse environment has been formative for my training as a scientist.

I would like to thank the BCS administrative staff, especially Julianne Ormerod and Sierra Vallin, for their tremendous efforts in keeping the department running smoothly and enriching the experience of students. Thank you also to Peter Child, Laura Jaye, and the MIT Music program, whose classes have brightened my later years of graduate school with the joy of music.

Thank you to Agnes and Phoebe for the many late-night conversations over hot pot. And thank you Christine, Kevin, and Molly, for your enduring friendship.

Finally, to Ben, to my brother Michael, and to my parents: thank you for your unconditional love and support.

Contents

1	Introduction	17
1.1	Practical challenges in the study of language	18
1.2	Artificial neural networks	21
1.2.1	ANNs and cognition	21
1.2.2	(Large) language models	23
1.2.3	Limitations of ANNs as cognitive models	24
1.3	Conceptual approach of thesis	26
1.4	Overview of thesis	26
2	A systematic assessment of syntactic generalization in neural language models	29
2.1	Introduction	29
2.2	Background	31
2.2.1	Perplexity	31
2.2.2	Targeted tests for syntactic generalization	31
2.2.3	Related work	33
2.3	Methods	33
2.3.1	Test suites	33
2.3.2	Model training data	36
2.3.3	Model classes	37
2.3.4	Off-the-shelf models	39
2.4	Results	39
2.4.1	Syntactic generalization and perplexity	40
2.4.2	Inductive bias and data scale	41

2.4.3	Circuit-level effects on SG score	42
2.4.4	Stability to modifiers	43
2.4.5	GPT-2 model performance	44
2.5	Discussion	45
3	Expectations over unspoken alternatives predict pragmatic inferences	47
3.1	Introduction	47
3.2	Background	51
3.2.1	Within-scale variation	51
3.2.2	Cross-scale variation (scalar diversity)	51
3.3	An expectation-based account of SI	52
3.3.1	String-based view of alternatives	54
3.3.2	Concept-based view of alternatives	55
3.4	Predicting variation within <i>⟨some, all⟩</i>	57
3.4.1	Human data	57
3.4.2	Model	57
3.4.3	Candidate alternatives	57
3.4.4	Results	57
3.5	Predicting variation across scales	58
3.5.1	Human data	58
3.5.2	Model	59
3.5.3	Candidate alternatives	60
3.5.4	Results	60
3.6	Related work	64
3.7	Discussion	65
3.7.1	How do listeners restrict the alternatives?	66
3.7.2	From alternatives to inference	67
3.7.3	Implications for NLP	69
4	A fine-grained comparison of pragmatic language understanding in humans and language models	71

4.1	Introduction	71
4.2	Related work	73
4.3	Evaluation materials	74
4.3.1	Overview of stimuli	74
4.3.2	Tested phenomena	76
4.4	Experiments	79
4.4.1	Evaluation paradigm	79
4.4.2	Models	80
4.5	Results	81
4.5.1	Do models choose the target pragmatic interpretation?	81
4.5.2	Do models and humans make similar types of errors?	82
4.5.3	Are models and humans sensitive to similar linguistic cues?	84
4.6	Discussion	86
5	Conclusion	89
5.1	Implications for cognitive science	90
5.2	Implications for natural language processing	95
A	Supplementary material for Chapter 2	101
A.1	Description of test suites	101
A.1.1	Notation	101
A.1.2	Center embedding	102
A.1.3	Pseudo-clefting	103
A.1.4	Filler–gap dependencies	105
A.1.5	Main-verb/reduced-relative garden-path disambiguation	107
A.1.6	Negative Polarity Licensing	108
A.1.7	NP/Z garden-path ambiguity	110
A.1.8	Subject–verb number agreement	111
A.1.9	Reflexive pronoun licensing	113
A.1.10	Subordination	114
A.2	Syntactic coverage of test suites	115

B	Supplementary material for Chapter 4	119
B.1	Example prompts	119
B.1.1	Deceits	119
B.1.2	IndirectSpeech	120
B.1.3	Irony	120
B.1.4	Maxims	120
B.1.5	Metaphor	121
B.1.6	Humor	121
B.1.7	Coherence	122
B.2	Timestamps of OpenAI model queries	122
B.3	No-context analysis	122
B.3.1	Details of human experiments	122
B.3.2	Raw accuracy scores	123
B.4	Sentence- and word-level scrambling	124
B.4.1	Sentence-level scrambled prompt	124
B.4.2	Word-level scrambled prompt	125

List of Figures

2-1	Average SG score by model class. Asterisks denote off-the-shelf models. Error bars denote bootstrapped 95% confidence intervals of the mean. . . .	40
2-2	Relationship between SG score and perplexity on our held-out BLLIP test set for each model.	41
2-3	Main results of our controlled evaluation of model class and dataset size. SG score varies more by model class (left) than by training dataset size (right).	42
2-4	Controlled evaluation results, split across test suite circuits. Circuit-level differences in SG score vary more by model class (left) than by training dataset size (right).	42
2-5	Evaluation results on all models, split across test suite circuits.	43
2-6	SG score on the pairs of test suites with and without intervening modifiers: Center Embedding, Cleft, MVRR, NPZ-Ambiguous, and NPZ-Object. . . .	44
3-1	(a) Distribution of human scalar inference (SI) ratings (on scale of 1-7) across instances of the $\langle some, all \rangle$ scale (reproduction of Fig. 1, Degen 2015). (b) Average SI rates across scales formed by different lexical items (reproduction of Fig. 2, van Tiel et al. 2016).	48
3-2	Relationship between human SI strength ratings within $\langle some, all \rangle$ scale (Degen, 2015) and BERT-derived predictors: (a) surprisal of scalemate <i>all</i> in the scalar construction, and (b) weighted average surprisal over the full set of candidate alternatives (Section 3.4.3). Each point represents a sentence. Shaded region denotes 95% CI.	58

3-3	Relationship between human SI rates and GPT-2-derived predictors across scales, for four datasets. Each point represents a single scale. Shaded region denotes 95% CI. (a) SI rate vs. surprisal of strong scalemate in the scalar construction. (b) SI rate vs. weighted average surprisal over the full set of candidate alternatives (Section 3.5.3).	61
3-4	GPT-2-derived surprisal of strong scalemate vs. accessibility rating of strong scalemates (Ronai and Xiang, 2022).	62
3-5	Probability assigned by GPT-2 to top 5 candidate strong alternatives (y-axis) for 3 example weak scalar items: <i>big</i> , <i>largely</i> , and <i>hard</i> (Ronai and Xiang, 2022). The full scalar construction is shown above each subplot, with the original tested strong scalemate underlined in <u>red</u>	63
4-1	Accuracy for each task. Error bars denote 95% CI. Dashed line indicates task-specific random baseline.	82
4-2	Mean accuracy vs. millions of parameters. Vertical dashed line indicates 1 billion parameters.	82
4-3	Response distributions across models and humans. Answer options for each task are shown on the x-axis. For models, y-axis denotes probability assigned to each answer option. For humans, y-axis denotes empirical frequency of each answer option being selected. Error bars denote 95% CI. Dashed line indicates random baseline.	83
4-4	Mean by-item difference in accuracy once story context was removed. . . .	85
4-5	Pearson correlation coefficients between by-item human accuracy and model probability of the correct answer. Cells are marked with significance codes.	86
5-1	Example interaction with ChatGPT, asking the model about probable candidates for the continuation of the sentence “the dog barked because”. Source: https://twitter.com/yoavgo/status/1598360581496459265	98
B-1	Proportion of items where humans and models select the correct pragmatic answer, on both original (shaded bars) and no-context (empty bars) versions.	124

B-2 Model performance across scrambling conditions (none = original, unmodified items). Error bars denote 95% CI. Dashed line indicates random baseline. 124

List of Tables

2.1	Statistics of training set for each corpus size.	36
2.2	Size of neural models in our controlled experiments.	38
2.3	Parameter counts for neural models in our controlled experiments.	38
2.4	Perplexity averages achieved by each controlled model on each corpus. Perplexity scores across training dataset sizes are not always strictly comparable (see Section 2.3.2).	39
3.1	Details of human data used in our analyses. An item is a unique (scale, context) combination.	50
3.2	Summary of full regression model predicting variation within $\langle some, all \rangle$, including original predictors from Degen (2015) (see the original study for a detailed description of each of the predictors).	59
3.3	Scalar construction templates for different parts of speech (for cross-scale variation).	59
3.4	Summary of full regression model (middle columns) and ANOVA comparing full model against intercept-only model (right columns) for each cross-scale variation dataset.	62
4.1	Sample item from each task in our evaluation. All items are originally curated by Floyd et al. (In prep).	75
4.2	Models tested in our experiments.	80
5.1	Contrast between traditional NLP benchmarking and targeted evaluation.	95
A.1	Test suite coverage of syntactic phenomena presented in Carnie (2012).	116

B.1 Timestamps of OpenAI API model queries. 122

Chapter 1

Introduction

Language is one of the hallmarks of human intelligence. Our societies are built upon relationships cultivated through collaboration and debate; our cumulative knowledge as a species is documented through written and oral histories; and the most fantastical reaches of our imagination are expressed through poetry and literature. Language learning, production, and comprehension seem to unfold effortlessly and successfully, supporting a rich set of behaviors that characterize human mental life. And yet, these abilities also pose significant computational challenges, requiring inference and generalization in the presence of noise and cognitive resource constraints. How, then, do children generalize from the linguistic stimulus with which they are presented? And how do interlocutors arrive at shared understanding so flexibly and efficiently? These questions touch upon foundational challenges in cognition, intertwined with the general problems of learning, reasoning, and collaboration.

For millennia, scholars across cultures and civilizations have been puzzled by the nature of language. Egyptian pharaohs and Holy Roman Emperors alike performed language deprivation experiments in order to uncover the language that children would know from birth (Herodotus, 440 B.C.; Coulton, 1972). In the Western tradition, philosophical investigations of language date as far back as Plato's *Cratylus* and Aristotle's *De Interpretatione*. Language became the focal point of analytical philosophy in the "linguistic turn" of the 1800s (Rorty, 1993), spawning a rich inquiry into meaning, reference, and the relationship between language and thought. The prominence of language in the study of the mind, coupled with the rise of experimental psychology in the 20th century, sparked a new gener-

ation of debate about the cognitive principles of language. In response to the behaviorist movement (Skinner, 1957), Chomsky (1959) proposed that language does not arise through reinforcement of behaviors. Instead, he argued, humans are biologically predisposed to learn language through genetically inherited mechanisms, and the core of language is knowledge of an underlying generative grammar (Chomsky, 1957, 1965). Chomsky’s proposal kindled a “cognitive revolution” in linguistics (McGilvray, 2014), establishing a tradition of inquiry that, for the first time, mirrored the goals and methods of the natural sciences.

Despite tremendous progress in the past century, many linguistic theories remain difficult to test with mathematical precision at naturalistic scales. In many ways, language presents unique practical challenges for quantitative empirical research. At the same time, recent advances in engineering have enabled computational models that use language with incredible sophistication. These models take on many different forms, but share the same underlying artificial neural network (ANN) architecture. Importantly, ANNs are fully intervenable – their training environments, learning mechanisms, and internal representations can be freely manipulated – and can provide predictions over arbitrary linguistic contexts. In this thesis, I argue that ANN language models are well-suited to address many of the unique challenges of studying language. Through three case studies focusing on syntactic generalizations and pragmatic inferences, I demonstrate how psycholinguistic evaluation of ANNs can offer new insights into human linguistic behaviors.

In the remainder of the Introduction, I discuss prominent challenges of studying language (Section 1.1), introduce ANNs as a tool for addressing these challenges (Section 1.2), and outline the conceptual framework for applying ANNs to investigate human language (Section 1.3). I then conclude with an outline of the thesis in Section 1.4.

1.1 Practical challenges in the study of language

Challenge 1: Controlled manipulations of language learning

Children in a particular language community receive varied and sparse input, but converge on the same generalizations about the permissible structures in the language (Chomsky, 1965). Do these generalizations arise from domain-general or language-specific learning

mechanisms? What are the relative contributions of input size and inductive biases?

Many positions in this debate have been based on logical arguments and toy grammars (e.g., Chomsky, 1980). Gaining empirical insights has been challenging, as controlled experiments on language learning are infeasible to perform in organisms. In other domains of cognition, such as visual perception, animal models have been critical for testing and advancing cognitive and neuroscientific theories (Hubel and Wiesel, 1962; Logothetis and Sheinberg, 1996; Zoccolan et al., 2009; Rajalingham et al., 2015). While animal models have been proposed to study low-level aspects of language such as speech (e.g., Fitch and Tallal, 2003; Helekar, 2013; Konopka and Roberts, 2016), it is unclear whether any non-human species can yield insights into higher-level phenomena such as the acquisition of syntax (see Hauser et al., 2002; Suzuki et al., 2016; Townsend et al., 2018; Suzuki and Zuberbühler, 2019; Schlenker et al., 2023, for discussion). The thinkers of the past – like Pharaoh Psamtik I and Holy Roman Emperor Frederick II – turned to human children as their experimental subjects, depriving them of linguistic interaction from birth. These manipulations clearly raise ethical and practical concerns. Furthermore, even with access to animal models, it would be unclear how to control key aspects of language learning, such as enforcing a preference for learning linear structures.

Challenge 2: Estimating expectations about the unsaid

One of the primary functions of language is to facilitate efficient information exchange (Hurford et al., 1998; Hurford, 2007; Kirby et al., 2015; Gibson et al., 2019; Hahn et al., 2020). However, the thoughts in speakers' minds are not simply transferred to listeners – noise, errors, and ambiguity are pervasive in everyday communication (Garrett, 1975; Levelt, 1983; Bock and Miller, 1991; Altmann, 1998). As such, the tools of probability theory have been productive in characterizing language comprehension (e.g., Jurafsky, 2002; Chater and Manning, 2006), mirroring the success of Bayesian models in other domains of cognition (Anderson, 1990; Griffiths et al., 2010; Tenenbaum et al., 2011). Probabilistic models of language understanding propose that listeners infer speakers' intended meanings by integrating expectations over multiple sources of information (e.g., Goodman and Frank, 2016; Degen, 2023). In many cases, a central component of this inference process is the

set of utterances that the speaker chose *not* to say – that is, the unspoken alternatives (e.g., Degen, 2013; Repp and Spalek, 2021).

Empirical studies suggest that listeners maintain context-driven probabilistic expectations over alternatives (Degen and Tanenhaus, 2015, 2016). However, it is difficult to estimate these expectations with enough precision to capture fine-grained variation in pragmatic inferences. While prior studies have measured expectations deployed in real-time reading using corpus-based relative frequency estimation (e.g., Levy, 2008; Smith and Levy, 2013), these methods may not work as well for capturing expectations over unspoken alternatives. Since alternatives are not produced, frequency-based models may be ill-suited to estimate the probability of an intended alternative, given the speaker’s utterance. Furthermore, some theories propose that alternatives operate at the conceptual level instead of at the level of linguistic forms (Buccola et al., 2021). Frequency-based models, which primarily use discrete token representations, might struggle to capture the conceptual similarity structures that listeners may be using to make inferences about alternatives.

Challenge 3: Isolating Theory of Mind in pragmatic comprehension

A large body of work has shown that humans consider other agents’ mental states when deciding how to produce or comprehend an utterance (Brennan et al., 2010a; Heller et al., 2012; Mozuraitis et al., 2018; Enrici et al., 2019; Clark and Marshall, 1981). However, reasoning about others’ beliefs, goals, and preferences has been characterized as a type of “System 2” reasoning, which is slow and cognitively effortful (Apperly and Butterfill, 2009). As a result, a prominent view is that these Theory of Mind (ToM) mechanisms are too cognitively demanding to underlie real-time processing (e.g., Gallagher, 2001; Pickering and Garrod, 2004). To address this tension, researchers have proposed a distinction between ToM abilities and “full-blown” ToM cognition (Butterfill and Apperly, 2013; Geurts and Rubio-Fernández, 2015), suggesting that ToM abilities can be approximated through a minimal set of heuristics (e.g., Heyes, 2014; Borg, 2018; Rubio-Fernández et al., 2019).

In the context of pragmatic language understanding, one specific proposal is that many implicatures are conventionalized and computed by default (e.g., Levinson, 2000). However, it is not straightforward to test the contributions of ToM and language experience in

pragmatic language understanding, as behavioral experiments cannot cleanly manipulate the presence of ToM reasoning in humans (or other living organisms). In principle, it would be possible to compare human pragmatic inference behaviors against statistical language models, which embody experience with linguistic forms and lack any form of explicit ToM (see Challenge 2). In practice, however, this is challenging, as many pragmatic understanding tasks (such as irony or metaphor) also require basic language processing, knowledge of cultural and social norms (Trosborg, 2010), and commonsense world knowledge – all of which are lacking even in relatively sophisticated computational models.

1.2 Artificial neural networks

This thesis demonstrates how artificial neural network (ANN) language models provide an important tool for addressing the challenges presented above. In the current section, I give an overview of ANNs and modern ANN-based language models. I highlight how ANNs can address the challenges described in Section 1.1, as well as their limitations for illuminating the human mind. In Section 1.3, I then discuss a conceptual framework for using ANNs to investigate questions in syntax acquisition and pragmatic language understanding.

1.2.1 ANNs and cognition

The significance of artificial neural networks traces back to early musings about the architecture of intelligence. Enlightenment philosophers were drawn to the idea that human thought could be seen as a logical sequence of rule-based operations on atomic units. Hobbes (1651) described reasoning as “nothing but reckoning (that is, adding and subtracting) of the consequences of general names agreed upon for the marking and signifying of our thoughts”; Leibniz argued that human reasoning could be described as the combination of an “alphabet of human thoughts” (Jourdain, 1916). As interest in artificial intelligence began to develop, this view became the basis of symbolism, or the idea that intelligence is based on knowledge abstractions that are manipulated through reasoning and learning (e.g., Newell and Simon, 1976; Simon, 1980; Fodor, 1975). On the other hand, the framework of connectionism argued that high-level cognitive phenomena can be explained by the activity of simultaneous,

distributed signals (Hinton et al., 1986), and the connections can be numerically modified based on experience (Rumelhart et al., 1987; McClelland and Rogers, 2003; Rogers and McClelland, 2004).

Artificial neural networks (ANNs) are a class of models that implement the connectionist hypothesis with mathematical precision and intervenability. ANNs are formed by interconnected computational units (Rosenblatt, 1958) typically arranged in a series of layers. Each unit receives real numbers as input, and outputs a non-linear function of the sum of the inputs. The weights of the connections between units get gradually updated according to a learning rule and error signal, typically through the backpropagation algorithm (Rumelhart et al., 1986). For example, an ANN might learn to perform visual object recognition by taking example images and updating its weights in a way that maximizes the probability of previously-labeled ground-truth classes. At the end of this training process, the weights of the ANN represent an approximately optimal solution for the object recognition task (given the appropriate distribution of inputs). These learned weights can then be used to classify new images that the model has never encountered before.

In this sense, one type of scientific insight that ANNs can provide is illustrating the behaviors and representations that emerge through task optimization in a given environment. In other words, ANNs allow us to test the hypothesis that a certain mental or neural phenomenon arises as an optimal solution to a specific computational problem that an organism may face (Kanwisher et al., 2023). This optimization-based view of ANNs has proven to be productive for testing theories of brain function and organization (e.g., Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Zhuang et al., 2017; Rajalingham et al., 2018; Kell et al., 2018; Schrimpf et al., 2021; Dobs et al., 2022; Doerig et al., 2022; Jain et al., 2023) and answering questions about *why* minds and brains look the way they do (Yamins and DiCarlo, 2016; Kanwisher et al., 2023). For example, deep convolutional neural networks (CNNs) trained on a visual object recognition task match primates' fine-grained behavioral patterns (Rajalingham et al., 2018) and neural activations (e.g., Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014), despite never having been trained to explicitly fit primate behaviors or brains. This suggests that optimization for visual classification is a computational principle that shapes the brain's visual algorithms and

organization.

1.2.2 (Large) language models

The traditional use of the term “language model” refers to a probability distribution over strings (i.e., sequences of tokens). Early n -gram language models estimated full-string probabilities by decomposing the joint probability of tokens into the product of conditional probabilities (Jurafsky and Martin, 2023). However, these models suffer from the curse of dimensionality: with large vocabularies, the number of possible token sequences increases exponentially. This poses significant data sparsity problems for estimating n -gram probabilities from corpora of natural text. Bengio et al. (2003) addressed this problem by proposing a language model based on ANNs and continuous word representations. Since then, ANN language models have achieved incredible success on next-word prediction (e.g., Brown et al., 2020) and are critical to applications such as machine translation (e.g., Brants et al., 2007; Gulcehre et al., 2017; Baziotis et al., 2020) and speech recognition (e.g., Bahl et al., 1989; Kuhn and De Mori, 1990; Jelinek et al., 1991; Toshniwal et al., 2018).

The nascent era of “large language models” (LLMs) has marked a dramatic shift in the size and capabilities of ANN language models (Bommasani et al., 2021).¹ While the architectures and learning algorithms underlying LLMs have existed for decades, the massive scale of modern language models has given rise to emergent abilities that are qualitatively different from what was possible before (Wei et al., 2022b). Beyond just generating fluent text, LLMs have been shown to learn from few-shot demonstrations (Brown et al., 2020), produce ordered outputs that mimic sequential reasoning (Nye et al., 2021; Wei et al., 2022c), synthesize scientific knowledge (Taylor et al., 2022), and generate effective code (Chen et al., 2021; Austin et al., 2021). These remarkable behaviors suggest that studying ANN models – the same way cognitive scientists might study other complex organisms – could

¹In the current AI landscape, the term “language model” is often loosely used to describe any model whose inputs and outputs are natural language. These models may perform a variety of tasks beyond incremental word prediction, such as masked language modeling (Devlin et al., 2019) or span infilling (Raffel et al., 2020). Importantly, many new “language models” are also fine-tuned on non-linguistic tasks and receive reinforcement from human feedback (Ouyang et al., 2022; Wei et al., 2022a). Whereas traditional language models implement prediction based on the distribution of linguistic forms, the connection between newer models and core cognitive motifs is less clear. Since my goal is to test ANNs as cognitive models, this thesis will primarily use autoregressive and bidirectional language models (except in Chapter 4).

potentially reveal the mechanisms supporting human linguistic knowledge.

In particular, I argue that ANN language models provide a useful tool for addressing the challenges discussed in Section 1.1. First, ANNs can be used to perform *in silico* experiments which manipulate the mechanisms supporting language acquisition. Researchers can manipulate ANN models' inductive biases, training inputs, and learning algorithms, enabling controlled experiments that would be infeasible to perform in humans or other living organisms (Chapter 2). Second, ANN-derived linguistic predictions are better aligned with human comprehension than probabilities derived from non-neural models (e.g., Goodkind and Bicknell, 2018). ANNs also use dense input representations, which could be useful for capturing conceptual similarity effects in language understanding. The predictive distributions of ANNs could therefore approximate human expectations over unspoken content in pragmatic inferences (Chapter 3). Finally, *large* language models display incredible knowledge about basic lexical meaning, social norms, and commonsense, while having no explicit representations of other agents' mental states. In this sense, they may capture many of the mechanisms that are involved in pragmatic language understanding, while lacking explicit Theory of Mind. This makes them uniquely suited to investigate the role of mentalizing in non-literal language understanding tasks (Chapter 4).

1.2.3 Limitations of ANNs as cognitive models

The central question of language acquisition is not whether language can be learned in the limit of infinite data, but rather how it arises in naturalistic settings. However, modern ANN language models are often trained on orders of magnitude more data (i.e., number of word tokens) than is typically available to a human child. For example, Warstadt and Bowman (2022) estimate that a ten-year-old child receives 100 million word tokens, whereas GPT-3 (Brown et al., 2020) receives *200 billion*. As such, one concern about the relevance of ANNs to linguistic theory is the discrepancy between ANNs' and children's input data. Several studies have sought to address this by performing systematic investigations of models trained on more realistic amounts of data (e.g., Warstadt et al., 2020b). However, the child-likeness of input data depends not only on size, but also in genre and modality – two

other dimensions where models differ from humans. ANN language models are typically trained through passive exposure to static text corpora, whereas human language learning involves social interaction (Baldwin et al., 1996) and perceptual grounding (e.g., Geeraerts, 2006; Vigliocco et al., 2014; Bisk et al., 2020; Bender and Koller, 2020) in addition to experience with linguistic forms. Furthermore, children typically observe continuous streams of child-directed speech (Rowe, 2008, 2012), whereas models typically learn from pre-segmented text scraped from books, encyclopedias, news archives, and internet forums (but see Huebner et al., 2021). This suggests that the task of generalizing from linguistic input is a different computational challenge for children than for models – although the extent of these differences, and how they affect the overall difficulty of the learning problem, remains unclear.

The issue of ecological validity is perhaps even more pronounced when using ANNs to study pragmatic language comprehension. In conversation, interaction is critical for supporting joint activities such as grounding (Clark and Brennan, 1991), alignment (Pickering and Garrod, 2004; Fusaroli et al., 2012), convention formation (Hawkins et al., 2021), and repair (Micklos and Woensdregt, 2022). Static ANN language models therefore are not a faithful approximation of human language comprehension, which typically unfolds as a cooperative act with a speaker, and relies on extra-linguistic shared context. Recent work has criticized ANN language models for their disconnect from a shared, interactive experience with other agents, with some arguing that this renders them unable to capture language understanding (Bisk et al., 2020; Bender and Koller, 2020) – and therefore limits their usefulness as cognitive models.

Clearly, ANNs differ in many ways from humans: they receive orders of magnitude more training data (Warstadt and Bowman, 2022), learn through biologically implausible mechanisms (Crick, 1989), and lack the rich grounded interactions that characterize human social life (Bisk et al., 2020; Bender and Koller, 2020). Given these limitations, how should one go about studying ANNs? What conclusions are researchers licensed to draw about humans, based on model behaviors? In Section 1.3, I briefly lay out a high-level conceptual approach to using ANNs to study human language, which will be the framework guiding the rest of the thesis.

1.3 Conceptual approach of thesis

This thesis takes the following approach to drawing conclusions about the human mind based on ANN behaviors. In each case study, I use ANN language models (see Section 1.2.2 for a note on terminology) trained on naturalistic English text as models of human language processing. I use psycholinguistic methods to measure the linguistic knowledge implicitly learned by ANNs. When ANN language models demonstrate positive behavioral results (e.g., human-like syntactic generalizations), this suggests a *lower bound* on the information that humans might learn through experience with language. Humans certainly integrate information from non-linguistic sources (see Section 1.2.3), but positive evidence from models can reveal what is, in principle, discoverable from linguistic signal without language-specific learning algorithms or symbolic representations. As such, the case studies in this thesis are designed to reveal models’ capacities and potential knowledge, instead of revealing models’ weaknesses when presented with adversarially-designed inputs (e.g., Nie et al., 2020b; Kiela et al., 2021).

More broadly, this approach treats ANNs as implementations of the hypothesis that linguistic knowledge arises as an emergent phenomenon through experience with language (e.g., Baroni, 2022; Warstadt and Bowman, 2022; Wilcox et al., 2022b). This connects to a long-standing debate about symbolic structures in language, and whether such representations might emerge through connectionist learning architectures. Indeed, Fodor and Pylyshyn (1988) themselves write that “a connectionist neural network can perfectly well implement a classical architecture at the cognitive level” (see also Lovering and Pavlick, 2022). Whether – and how – this occurs in language processing remains an empirical question that ANNs are well-suited to address.

1.4 Overview of thesis

In Chapter 2, I use ANNs to investigate how syntactic generalizations are acquired. We conduct a systematic evaluation of the syntactic knowledge of neural language models, testing 20 combinations of model types and data sizes on 34 English-language test suites.

We find substantial differences in syntactic generalization performance by model architecture, with sequential models underperforming architectures with hierarchical biases. This suggests that domain-general ANNs can learn syntactic generalizations through exposure to linguistic forms, contrary to strong nativist claims; however, inductive bias plays a more important role than size of input data, especially in small-data settings. This chapter is based on published materials from Hu et al. (2020b).

In Chapter 3, I study how ANN-derived linguistic expectations can explain systematic variation in scalar inferences (SI). Empirical studies have shown that human SI rates are highly variable, both within (Degen, 2015) and across scales (e.g., van Tiel et al., 2016). However, there have been few proposals explaining both cross- and within-scale variation. Furthermore, while it is generally assumed that SIs arise through reasoning about unspoken alternatives, it remains debated whether humans reason about alternatives as linguistic forms, or at the level of concepts. We test a shared mechanism explaining SI rates within and across scales: context-driven expectations about the unspoken alternatives (Degen and Tanenhaus, 2015). Using neural language models to approximate human predictive distributions, we find that SI rates are captured by the expectedness of the strong scalemate as an alternative. Crucially, however, expectedness robustly predicts cross-scale variation only under a meaning-based view of alternatives. Our results suggest that pragmatic inferences arise from context-driven expectations over alternatives, and these expectations operate at the level of concepts. This chapter is based on published materials from Hu et al. (2023b).

In Chapter 4, I use large language models to investigate the mechanisms underlying a broad set of non-literal language understanding tasks. We compare ANNs to humans on seven pragmatic phenomena, asking which pragmatic behaviors may arise through experience with language without explicit models of other agents or the world. We find that the largest models achieve high accuracy and match human error patterns: within incorrect responses, models favor the literal interpretation of an utterance over heuristic-based distractors. Through behavioral experiments, we also find preliminary evidence that models and humans are similarly sensitive to linguistic cues that make a non-literal interpretation more or less likely. Our results suggest that pragmatic behaviors can emerge in models without explicitly constructed representations of mental states. However, models

tend to struggle with phenomena relying on social expectation violations. This chapter is based on materials to appear at *ACL* (Hu et al., 2023a).

Finally, I conclude by discussing the implications of these findings for cognitive science and artificial intelligence, and highlight directions for future research (Chapter 5).

Chapter 2

A systematic assessment of syntactic generalization in neural language models

2.1 Introduction

A growing body of work advocates that assessment of neural language models should include both information-theoretic metrics, such as perplexity, as well as targeted linguistic evaluation. Benchmarks such as GLUE (Wang et al., 2019a,b) have demonstrated that neural language models trained on naturalistic corpora for next-word prediction learn representations that can yield remarkable performance on many semantic tasks. Targeted syntactic evaluations have shown that these models also implicitly capture many **syntactic generalizations**, ranging from subject–verb agreement to long-distance filler–gap dependencies (Linzen et al., 2016; Marvin and Linzen, 2018; Futrell et al., 2018; Wilcox et al., 2019b). This paper aims to bring targeted evaluations of syntactic performance to scale, complementing similar developments in semantic evaluation (McCoy et al., 2019).

Because the most widespread currency of evaluation for language models is perplexity—how well, on average, a model predicts a word in its context—a primary focus of this paper is the relationship between a model’s perplexity and its performance on targeted syntactic evaluations. As perplexity improves, can we expect more human-like syntactic generalization? How do training dataset size and model architecture jointly affect syntactic generalization? And what picture of models’ syntactic generalization emerges when evaluation is brought to

scale, across dozens of controlled syntactic tests?

In this paper we offer initial answers to these questions, systematically assessing the syntactic generalization abilities of neural language models on 34 targeted test suites (33 adapted from previously published work, and 1 novel) covering a wide range of syntactic phenomena. Test suites are written using a standard format that allows for flexible predictions which more closely resemble those used in psycholinguistic studies, specifically allowing for predictions about interactions among multiple testing conditions. Performance on each test suite is reported as a Syntactic Generalization (SG) score. We group test suites into six syntactic circuits based on the linguistic representations needed to achieve high performance on each suite.

We train four classes of neural models and one baseline n -gram model on four datasets derived from a newswire corpus, consisting of 1, 5, 14, and 42 million tokens. While previous work has compared model architectures for a fixed dataset size (e.g. Wilcox et al., 2019b) and network sizes for a fixed architecture (e.g. van Schijndel et al., 2019), our controlled regime allows us to make an apples-to-apples comparison across model architectures on a range of sizes. In addition, we evaluate several off-the-shelf models which were trained on datasets ranging up to 2 billion tokens.

Our results address the three questions posed above: First, for the range of model architectures and dataset sizes tested, we find a substantial dissociation between perplexity and SG score. Second, we find a larger effect of model inductive bias than training data size on SG score, a result that accords with van Schijndel et al. (2019). Models afforded explicit structural supervision during training outperform other models: One structurally supervised model is able to achieve the same SG scores as a purely sequence-based model trained on ~ 100 times the number of tokens. Third, we find that architectures have different relative advantages across types of syntactic tests, suggesting that the tested syntactic phenomena tap into different underlying processing capacities in the models.

2.2 Background

2.2.1 Perplexity

Standard language models are trained to predict the next token given a context of previous tokens. Language models are typically assessed by their *perplexity*, the inverse geometric mean of the joint probability of words w_1, \dots, w_N in a held-out test corpus C :

$$\text{PPL}(C) = p(w_1, w_2, \dots, w_N)^{-\frac{1}{N}} \quad (2.1)$$

Models with improved perplexity have also been shown to better match various human behavioral measures, such as gaze duration during reading (Frank and Bod, 2011; Fossum and Levy, 2012; Goodkind and Bicknell, 2018; Wilcox et al., 2020). However, a broad-coverage metric such as perplexity may not be ideal for assessing human-like syntactic knowledge for a variety of reasons. In principle, a sentence can appear with vanishingly low probability but still be grammatically well-formed, such as *Colorless green ideas sleep furiously* (Chomsky, 1957). While perplexity remains an integral part of language model evaluation, fine-grained linguistic assessment can provide both more challenging and more interpretable tests to evaluate neural models.

2.2.2 Targeted tests for syntactic generalization

Alternatively, a language model can be evaluated on its ability to make human-like generalizations for specific syntactic phenomena (Linzen et al., 2016; Lau et al., 2017; Gulordava et al., 2018). The targeted syntactic evaluation paradigm (Marvin and Linzen, 2018; Futrell et al., 2019) incorporates methods from psycholinguistic experiments, designing sentences which hold most lexical and syntactic features of each sentence constant while minimally varying features that determine grammaticality or surprise characteristics of the sentence. For example, given the two strings *The keys to the cabinet are on the table* and **The keys to the cabinet is on the table*, a model that has learned the proper subject–verb number agreement rules for English should assign a higher probability to the grammatical plural verb in the first sentence than to the ungrammatical singular verb in the second (Linzen

et al., 2016).

Although some targeted syntactic evaluations, such as the example discussed above, involve simple comparisons of conditional probabilities of a word in its context, other evaluations are more complex. We can demonstrate this with an evaluation of models’ “garden-pathing” behavior (Futrell et al., 2019). For example, the sentence *The child kicked in the chaos found her way back home* yields processing disruption for humans at the word *found*. This is because, up to right before that word, the part-of-speech ambiguous *kicked* is preferentially interpreted as the main verb of the sentence, whereas it turns out to be a passive participle in a reduced relative clause modifying *child*. This garden-path disambiguation effect is ameliorated by replacing *kicked* with *forgotten*, which is not part-of-speech ambiguous (B below; Trueswell et al., 1994) or by using an unreduced relative clause (C below; Ferreira and Clifton, 1986). In probabilistic language models, these garden-path disambiguation effects are well captured by word negative log probabilities, or SURPRISALS (Hale, 2001): $S(w|C) = -\log_2 p(w|C)$, which are independently well-established to predict human incremental processing difficulty over several orders of magnitude in word probability (Smith and Levy, 2013). A targeted syntactic evaluation for garden-pathing is provided by comparing surprisals at the disambiguating word *found* in the set of four examples below (Futrell et al., 2019):

- (A) The child kicked in the chaos **found** . . .
- (B) The child forgotten in the chaos **found** . . .
- (C) The child who was kicked in the chaos **found** . . .
- (D) The child who was forgotten in the chaos **found** . . .

Successful human-like generalization involves three criteria: (i) *found* should be less surprising (i.e., more probable) in B than A; (ii) *found* should be more probable in C than A; (iii) the C–D surprisal difference should be smaller than the A–B surprisal difference—a 2×2 *interaction effect* on surprisal—because the syntactic disambiguation effect of not reducing the relative clause was achieved by using a part-of-speech unambiguous verb.

We will use these controlled tests to help us describe and test for human-like syntactic

knowledge in language models.

2.2.3 Related work

The testing paradigm presented here differs in several crucial ways from recent, related syntactic assessments and provides complementary insights. Unlike Warstadt et al. (2019), our approach does not involve fine-tuning, but rather assesses what syntactic knowledge is induced from the language modeling objective alone. The most closely related work is the Benchmark of Linguistic Minimal Pairs (Warstadt et al., 2020a), which is a challenge set of automatically-generated sentence pairs also designed to test language models on a large set of syntactic phenomena. Our approach differs in important ways: we compare critical sentence regions instead of full-sentence probabilities, and employ a 2×2 paradigm with a strict, multi-fold success criterion inspired by psycholinguistics methodology. This allows us to factor out as many confounds as possible, such as the lexical frequency of individual tokens and low-level n -gram statistics.

2.3 Methods

We designed a controlled paradigm for systematically testing the relationship between two design choices — model class and dataset size — and two performance metrics — perplexity and syntactic generalization capacity. Section 2.3.1 describes the test suites collected for our evaluation, and Sections 2.3.2 and 2.3.3 describe the datasets and model classes investigated.

2.3.1 Test suites

We assemble a large number of test suites inspired by the methodology of experimental sentence-processing and psycholinguistic research. Each test suite contains a number of ITEMS (typically between 20 and 30), and each item appears in several CONDITIONS: across conditions, a given item will differ only according to a controlled manipulation designed to target a particular feature of grammatical knowledge. Each test suite contains at least one PREDICTION, which specifies inequalities between surprisal values at pairs

of regions/conditions that should hold if a model has learned the appropriate syntactic generalization.¹

We expect language models which have learned the appropriate syntactic generalizations from their input to satisfy these inequalities without further fine-tuning. We compute accuracy on a test suite as the proportion of items for which the model’s behavior conforms to the prediction. Most of our test suites involve 2×2 designs and a success criterion consisting of a conjunction of inequalities across conditions, as in the garden-pathing example described in Section 2.2.2.² Random baseline accuracy varies by test suite and is $\sim 25\%$ overall. Most of these test suites and criteria are designed so that n -gram models cannot perform above chance for $n = 5$ (sometimes greater).

Syntactic coverage In order to assess the coverage of our test suites, we manually inspected the phenomena covered in Carnie (2012), a standard introductory syntax textbook. Of the 47 empirical phenomena reviewed in the summary sections at the end of each chapter, our tests target 16 ($\sim 34\%$). These are evenly distributed across the whole range of subject matter, with tests targeting phenomena in 11 of the 15 chapters ($\sim 73\%$).³

Modifiers Five test suites include paired modifier versions, where extra syntactically irrelevant (but semantically plausible) content, such as a prepositional phrase or relative clause, is inserted before the critical region being measured. We use these paired test suites to evaluate models’ stability to intervening content within individual syntactic tests.

Circuits The test suites are divided into 6 syntactic circuits, based on the type of algorithm required to successfully process each construction. We give a brief overview of each circuit below.

- **Agreement** is a constraint on the feature values of two co-varying tokens. For example, the number feature of a verb must agree with the number feature of its

¹A full overview of our test suites is given in Appendix A.1.

²The exception is Center Embedding, which features a 2-condition design with a single-inequality criterion.

³For more details on this analysis, see Appendix A.2.

upstream subject. We include 3 *Subject-Verb Number Agreement* suites from Marvin and Linzen (2018).

- **Licensing** occurs when a particular token must exist within the scope of an upstream licensor token. Scope is determined by the tree-structural properties of the sentence. Test suites include *Negative Polarity Item Licensing (NPI)* (4 suites) and *Reflexive Pronoun Licensing* (6 suites), both from Marvin and Linzen (2018).
- **Garden-Path Effects** are well-studied syntactic phenomena that result from tree-structural ambiguities that give rise to locally-coherent but globally implausible syntactic parses. Garden-path test suites include *Main Verb / Reduced Relative Clause (MVRR)* (2 suites) and *NP/Z Garden-paths (NPZ)* (4 suites), both from Futrell et al. (2018).
- **Gross Syntactic Expectation** is a processor’s expectation for large syntactic chunks such as verb phrases or sentences, and are often set up by subordinating conjunctions such as *while*, *although* and *despite*. Our tests for gross syntactic expectation include *Subordination* (4 suites) from Futrell et al. (2018).
- **Center Embedding** sentences are sentences recursively nested within each other. Subject and verbs must match in a first-in-last-out order, meaning models must approximate a stack-like data-structure in order to successfully process them. Our 2 suites of *Center Embedding* sentences come from the items presented in Wilcox et al. (2019a).
- **Long-Distance Dependencies** are co-variations between two tokens that span long distances in tree depth. Test suites include *Filler-Gap Dependencies (FGD)* (6 suites) from Wilcox et al. (2018) and Wilcox et al. (2019b), and 2 novel *Cleft* suites, described in detail below.

Novel test suite: Cleft We introduce one novel test suite that assesses models’ ability to process pseudo-cleft constructions, which are used to put a particular syntactic constituent into focus via passive transformation. Consider Example (1):

BLLIP sizes:	XS	SM	MD	LG
# sentences	40K	200K	600K	1.8M
# tokens	1M	4.8M	14M	42M
# non-UNK types	24K	57K	100K	170K
# UNK types	68	70	71	74

Table 2.1: Statistics of training set for each corpus size.

- (1) a. What he did after coming in from the rain was **eat a hot meal**. [DO/VP]
 b. *What he devoured after coming in from the rain was **eat a hot meal**. [LEX/VP]
 c. *What he did after coming in from the rain was **a hot meal**. [DO/NP]
 d. What he devoured after coming in from the rain was **a hot meal**. [LEX/NP]

When this constituent is a verb, it must be replaced in the wh-clause that heads the sentence with the DO verb, as in (1a), below. However, when it is a noun, the lexical verb for which it serves as an object must be preserved, as in (1d). If models have properly learned the pseudo-cleft construction, then DO verbs should set up expectations for VPs (the region in bold should have a lower surprisal in (1a) than in (1b)) and lexicalized verbs should set up expectations for NPs (the region in bold should have a lower surprisal in (1d) than in (1c)).

2.3.2 Model training data

Corpora We train and evaluate models on English newswire corpora of four different sizes, obtained by randomly sampling sections from the Brown Laboratory for Linguistic Information Processing 1987-89 Corpus Release 1 (BLLIP; Charniak et al., 2000). The corpora are sampled such that the training set of each corpus is a proper subset of each larger corpus. We call these four corpora BLLIP-XS (40K sentences, 1M tokens); BLLIP-SM (200K sentences, 5M tokens); BLLIP-MD (600K sentences, 14M tokens); and BLLIP-LG (2M sentences, 42M tokens). Table 2.1 summarizes statistics of the training set for each corpus.

To ensure consistency in perplexity evaluation across datasets, we report perplexity scores achieved by the models on a shared held-out test set. We additionally use a shared held-out validation for tuning and early stopping.

We use the NLTK implementation of the Penn Treebank tokenizer to process all datasets (Bird and Loper, 2004; Marcus et al., 1993).

Out-of-vocabulary tokens For each corpus, we designate a token as OOV if the token appears fewer than two times in the training set. Our larger training datasets thus contain larger vocabularies than our smaller training datasets. This allows larger-training-set models to learn richer word-specific information, but may also harm perplexity evaluation because they have vocabulary items that are guaranteed to not appear in the BLLIP-XS test set. This means that perplexity scores across training dataset sizes will not be strictly comparable: if a larger-training-set model does better than a smaller-training-set model, we can be confident that it has meaningfully lower perplexity, but the reverse is not necessarily the case. The exception to the above is GPT-2, which uses sub-words from byte-pair encoding and has no OOVs (see also Footnote 6).

Unkification We follow the convention used by the Berkeley parser (Petrov and Klein, 2007), which maps OOVs to UNK classes which preserve fine-grained information such as orthographic case distinctions and morphological suffixes (e.g. UNK-ed, UNK-ly). Before training, we verified that the UNK classes in the test and validation sets were all present in the training set.

2.3.3 Model classes

In order to study the effects of model inductive bias and dataset size, we trained a fleet of models with varying inductive biases on each corpus. Because many of our test suites exploit ambiguities that arise from incremental processing, we restrict evaluation to left-to-right language models; future work could involve evaluation of bidirectional models (Devlin et al., 2019; Yang et al., 2019) on an appropriate subset of our test suites, and/or adaptation of our suites for use with bidirectional models (Goldberg, 2019). Training ran until convergence of perplexity on a held-out validation set. Wherever possible, we trained multiple seeds of each model class and corpus size. We use the model sizes and training hyperparameters reported

	# layers	# hidden units	Embedding size
LSTM	2	256	256
ON-LSTM	3	1150	400
RNNG	2	256	256
GPT-2	12	768	768

Table 2.2: Size of neural models in our controlled experiments.

BLLIP sizes:	XS	SM	MD	LG
LSTM	13.4M	30.5M	52.2M	88.1M
ON-LSTM	30.8M	44.2M	61.2M	89.2M
RNNG	22.8M	48.4M	81.1M	134.9M
GPT-2	124.4M	124.4M	124.4M	124.4M

Table 2.3: Parameter counts for neural models in our controlled experiments.

in the papers introducing each model (Table 2.2).⁴ The full parameter counts and perplexity scores for each model \times corpus combination are given in Tables 2.3 and 2.4, respectively.

LSTM Our baseline neural model is a vanilla long short-term memory network (LSTM; Hochreiter and Schmidhuber, 1997) based on the boilerplate PyTorch implementation (Paszke et al., 2017).

Ordered-Neurons We consider the Ordered-Neurons LSTM architecture (ON-LSTM; Shen et al., 2019), which encodes an explicit bias towards modeling hierarchical structure.

RNNG Recurrent neural network grammars (RNNG; Dyer et al., 2016) model the joint probability of a sequence of words and its syntactic structure. RNNG requires labeled trees that contain complete constituency parses, which we produce for BLLIP sentences with an off-the-shelf constituency parser (Kitaev and Klein, 2018).⁵ To compute surprisals from RNNG, we use word-synchronous beam search (Stern et al., 2017) to approximate the conditional probability of the current word given the context.

⁴Due to computational constraints, we performed only minimal tuning past these recommended hyperparameters.

⁵While the BLLIP corpus already contains Treebank-style parses, we strip the terminals and re-parse in order to obtain more accurate, up-to-date syntactic parses.

BLLIP sizes:	XS	SM	MD	LG
LSTM	98.19	65.52	59.05	57.09
ON-LSTM	71.76	54.00	56.37	56.38
RNN	122.46	86.72	71.12	69.57
GPT-2	529.90	183.10	95.03	60.40
<i>n</i> -gram	240.21	158.60	125.58	106.09

Table 2.4: Perplexity averages achieved by each controlled model on each corpus. Perplexity scores across training dataset sizes are not always strictly comparable (see Section 2.3.2).

Transformer Transformer models (Vaswani et al., 2017) have recently gained popularity in language processing tasks. We use GPT-2 (Radford et al., 2019) as a representative Transformer model and train it from scratch on our BLLIP corpora.⁶

***n*-gram** As a baseline, we consider a 5-gram model with modified Kneser-Ney smoothing.

2.3.4 Off-the-shelf models

We also test five off-the-shelf models: GRNN, trained on 90M tokens from Wikipedia (Gulordava et al., 2018); JRNN, trained on 800M tokens from the 1 Billion Word Benchmark (Jozefowicz et al., 2016); Transformer-XL, trained on 103M tokens from WikiText-103 (Dai et al., 2019); and the pre-trained GPT-2 and GPT-2-XL, trained on 40GB of web text (Radford et al., 2019). These models are orders of magnitude larger than our controlled ones in parameter count and/or training set size.

2.4 Results

Figure 2-1 shows the average accuracy of all models on the complete set of SG test suites. Asterisks denote off-the-shelf models. With sufficient training data, all neural models achieve a SG score significantly greater than a random baseline (dashed line). However, the range within neural models is notable, with the best-performing model (GPT-2-XL)

⁶Our GPT-2 code is based on nshepperd/gpt-2. The model vocabulary consists of byte-pair encoded sub-words extracted from the GPT-2 pre-trained model, not from the BLLIP training corpora. To calculate GPT-2 perplexities, we divide the sum of all sub-word conditional log-probabilities by the total number of words in the corpus.

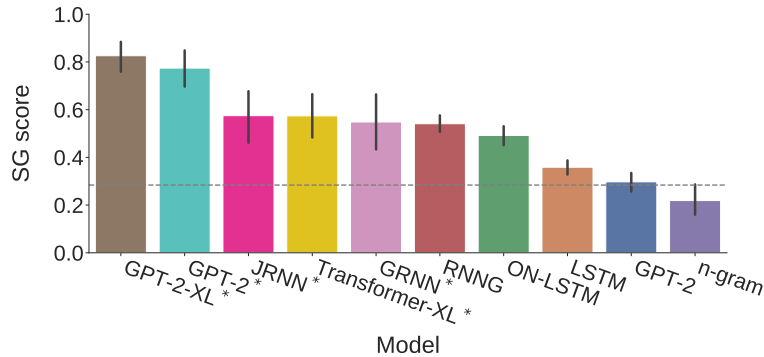


Figure 2-1: Average SG score by model class. Asterisks denote off-the-shelf models. Error bars denote bootstrapped 95% confidence intervals of the mean.

scoring over twice as high as the worst-performing model (LSTM). Also notable is the controlled RNNNG model, which achieve comparable performance to Transformer-XL and JRNN, despite being trained on significantly smaller data sizes.

We now return to the three major issues presented in Section 2.1. In 2.4.1 we present evidence that SG score is dissociated from perplexity. In 2.4.2 we argue that model architecture accounts for larger gains in SG score than amount of training data. And in 2.4.3 we show that this cross-architecture difference is due largely to variance on a handful of key test suites.

2.4.1 Syntactic generalization and perplexity

Figure 2-2 shows the relationship between SG score and perplexity on the BLLIP test set across models and training set sizes. As expected, *n*-gram models never rise appreciably above chance in SG score. Among neural models which exceed chance performance, there is no simple relationship between perplexity and SG score, especially once training dataset size is controlled for (comparing points in Figure 2-2 of the same color). For example, there is a remarkable amount of variance in the SG score of models trained on BLLIP-LG not explained by perplexity. This suggests that targeted syntactic evaluation can reveal information that may be orthogonal to perplexity.

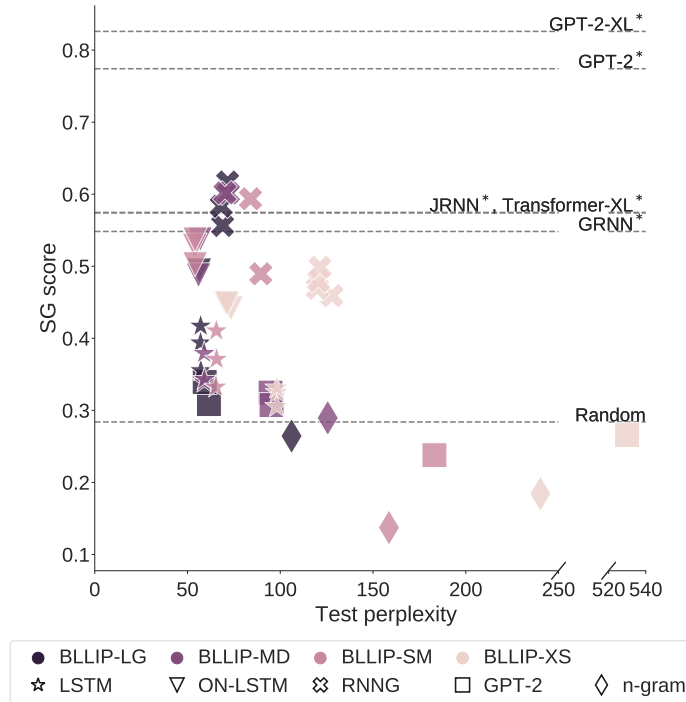


Figure 2-2: Relationship between SG score and perplexity on our held-out BLLIP test set for each model.

2.4.2 Inductive bias and data scale

In order to decouple the effects of model class and data scale from test suite difficulty, we represent a particular trained model’s performance on each test suite as a delta relative to the average performance of all models on this test suite. Unless noted otherwise, the remainder of the figures in this section plot a score delta, aggregating these deltas within model classes or corpus types.

Figure 2-3 tracks the influence of model class and data scale across the model types tested in our experiments, with SG score deltas on the y-axis. The left-hand panel shows the difference in SG score by model class. We find that model class clearly influences SG score: for example, the error bars (bootstrapped 95% confidence intervals of the mean) for RNNG and LSTM do not overlap. The right-hand panel shows the difference in SG score delta by training dataset, and shows a much more minor increase in mean SG score as training data increases.

We tested the influence of these factors quantitatively using a linear mixed-effects regression model, predicting suite-level performance as a feature of model architecture and

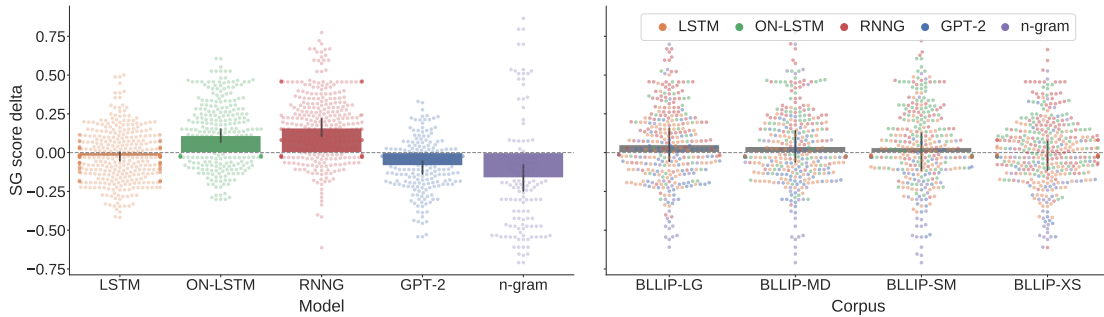


Figure 2-3: Main results of our controlled evaluation of model class and dataset size. SG score varies more by model class (left) than by training dataset size (right).

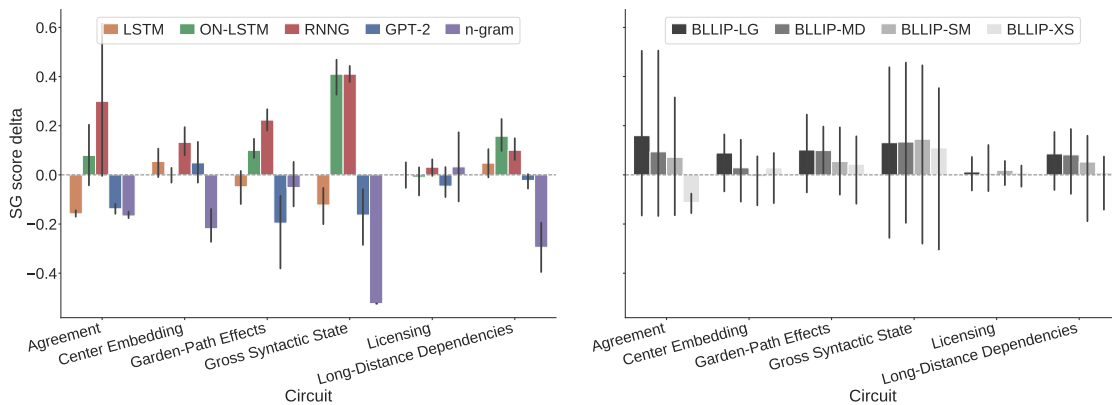


Figure 2-4: Controlled evaluation results, split across test suite circuits. Circuit-level differences in SG score vary more by model class (left) than by training dataset size (right).

training dataset size (represented as log-number of words). Both features made statistically significant contributions to SG score (both $p < 0.001$). However, predictor ablation indicates that architecture affects regression model fit more (AIC=-809 when dataset size is ablated; AIC=-822 when architecture is ablated).⁷

2.4.3 Circuit-level effects on SG score

Figure 2-4 shows the breakdown at the circuit level by model architecture (left) and training dataset size (right). The right panel demonstrates little effect of dataset size on SG score delta

⁷ n -grams and/or GPT-2 could arguably be expected to have qualitatively different sensitivity to training dataset size (the latter due to byte-pair encoding), so we repeated the analyses here and in Section 2.4.3 excluding both architectures individually as well as simultaneously. In all cases the same qualitative patterns described in the main text hold.

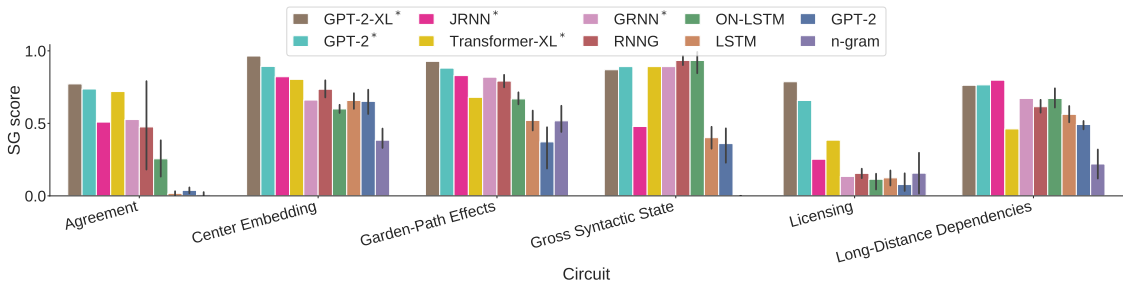


Figure 2-5: Evaluation results on all models, split across test suite circuits.

within most circuits, except for Agreement, on which the models trained on our smallest dataset fare poorly. In the left panel we find substantial between-circuit differences across architectures. Linear mixed-effects analyses support this finding: interactions with circuit are significant for both training dataset size and model architecture, but stronger for the latter (AIC=-890 and AIC=-868 when size and architecture are respectively ablated).

While model inductive biases separate clearly in performance on some circuits, they have little effect on performance on Licensing. This minimally suggests that Licensing taps into a distinct syntactic process within language models. One potential explanation for this is that the interactions tested by Licensing involve tracking two co-varying tokens where the downstream token is optional (see e.g. Hu et al., 2020a).

We show the circuit-level breakdown of absolute SG scores for all models (including off-the-shelf) in Figure 2-5. In general, the models that obtain high SG scores on average (as in Figure 2-1) also perform well across circuits: pre-trained GPT-2 and GPT-2-XL outperform all other models on each circuit, including Licensing, on which JRNN, GRNN, and most of our custom-trained models perform particularly poorly. Again, we highlight the impressive performance of RNNNG: it achieves comparable average performance to GRNN on all circuits, despite being trained on a fraction of the data size.

2.4.4 Stability to modifiers

We separately investigate the degree to which models’ syntactic generalizations are robustly stored in memory. For five test suites (Center Embedding, Cleft, MVRR, NPZ-Ambiguous,

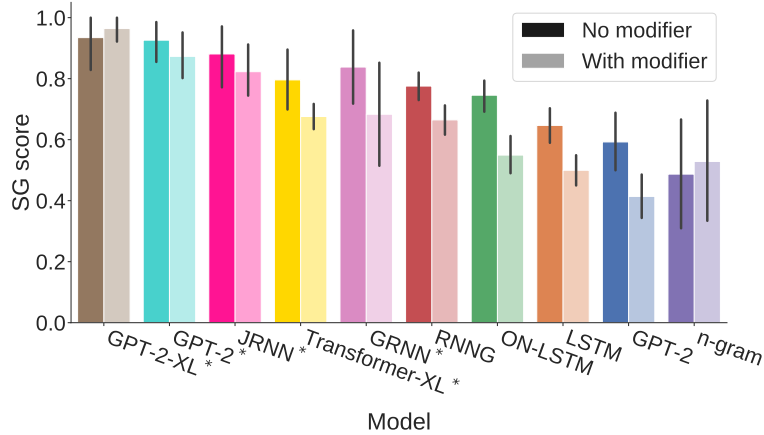


Figure 2-6: SG score on the pairs of test suites with and without intervening modifiers: Center Embedding, Cleft, MVRR, NPZ-Ambiguous, and NPZ-Object.

NPZ-Object), we designed minimally edited versions where syntactically irrelevant intervening content was inserted before the critical region. An ideal model should robustly represent syntactic features of its input across these modifier insertions.

In Figure 2-6 we plot models’ average scores on these five test suites (dark bars) and their minimally edited versions (light bars), evaluating how robust each model is to intervening content. Among models in our controlled experiments, we see that model class clearly influences the degree to which predictions are affected by intervening content (compare e.g. the stability of RNN to that of ON-LSTM). Some off-the-shelf models, such as GPT-2-XL, perform near ceiling on the original five test suites and are not affected at all by intervening content.

2.4.5 GPT-2 model performance

The GPT-2 models trained from scratch in these experiments exhibit an especially strong disconnect between SG score and perplexity. Comparing the models trained on BLLIP-MD and BLLIP-LG, we see a 36% improvement in average perplexity with negligible improvement (3% on average) in SG score.

This trend does not hold globally for the GPT-2 model, however: we have evidence that changes in both data scale and model inductive bias can improve GPT-2’s performance.

First, the pretrained GPT-2 model (equal in parameters to our own GPT-2 models) achieves near state-of-the-art performance in SG score. This demonstrates that, between the data scale of our experiments and the extreme data scale of the pretrained GPT-2 model, further improvements in perplexity likely correlate with improvements in SG score. Second, there appears to be evidence that small modifications to the GPT-2 model’s inference regime can yield a correlated SG score–perplexity relationship. While our GPT-2 models were trained with input context windows of 1024 tokens, Qian et al. (2021) trained a GPT-2 architecture with input context windows containing only single sentences. Their models achieve both good perplexity and high SG score, with average SG score 0.665 and perplexity 49.0 for models trained on BLLIP-LG, and average SG score 0.666 and perplexity 67.6 for models trained on BLLIP-MD. This suggests that constraining the context window size may help Transformer models generalize when the data scale is relatively small.

2.5 Discussion

This work addresses multiple open questions about syntactic evaluations and their relationship to other language model assessments. Our results dissociate model perplexity and performance in syntactic generalization tests, suggesting that the two metrics capture complementary features of language model knowledge. In a controlled evaluation of different model classes and datasets, we find model architecture plays a more important role than training data scale in yielding correct syntactic generalizations. Our circuit-level analysis reveals consistent failure on Licensing but inconsistent behavior on other circuits, suggesting that different syntactic circuits make use of different underlying processing capacities. In addition to the insight these results provide about neural NLP systems, they also bear on questions central to cognitive science and linguistics, putting lower bounds on what syntactic knowledge can be acquired from string input alone.

Targeted syntactic evaluation is just one in a series of complementary methods being developed to assess the learning outcomes of neural language processing models. Other methods include classifying sentences as grammatical or ungrammatical (Warstadt et al., 2019), decoding syntactic features from a model’s internal state (Belinkov et al., 2017;

Giulianelli et al., 2018), or transfer learning to a strictly syntactic task such as parsing or POS tagging (Hewitt and Manning, 2019). As each task brings an explicit set of assumptions, complementary assessment methods can collectively provide greater insight into models' learning outcomes.

Although this paper, together with Warstadt et al. (2020a), report what is to our knowledge the largest-scale targeted syntactic evaluations to date, we emphasize that they are only first steps toward a comprehensive understanding of the syntactic capabilities of contemporary language models. This understanding will be further advanced by new targeted-evaluation test suites covering a still wider variety of syntactic phenomena, additional trained models with more varied hyperparameters and randomization seeds, and new architectural innovations. Humans develop extraordinary grammatical capabilities through exposure to natural linguistic input. It remains to be seen to just what extent contemporary artificial systems do the same.

Chapter 3

Expectations over unspoken alternatives predict pragmatic inferences

3.1 Introduction

Much of the richness of linguistic meaning arises from what is left unsaid (e.g., Grice, 1975; Sperber and Wilson, 1986; Horn, 1989). For example, if Alice says “Some of the students passed the exam”, Bob can infer that Alice means *not all* students passed the exam, even though Alice’s utterance would still be logically true if all students had passed. One explanation of this inference is that Bob reasons about the unspoken **alternatives** that were available to the speaker. Under the assumptions that (1) speakers generally try to be informative, (2) Alice has full knowledge of the situation, and (3) it would have been relevant and more informative for Alice to say “All of the students passed the exam”, Alice’s choice to say “some” suggests that she believes the sentence with “all” is false. This inference pattern is more generally known as **scalar inference** (SI), which arises from orderings between linguistic items (scales).

SI has often been treated as a categorical phenomenon: when a speaker utters a weaker (less informative) item on a scale, a listener rules out the meaning of stronger (more informative) items on that scale (e.g., Levinson, 2000). However, empirical studies have demonstrated substantial variability in the rates at which humans draw SIs, both within instances of a single scale (Degen, 2015; Eiteljoerge et al., 2018; Li et al., 2021) and across

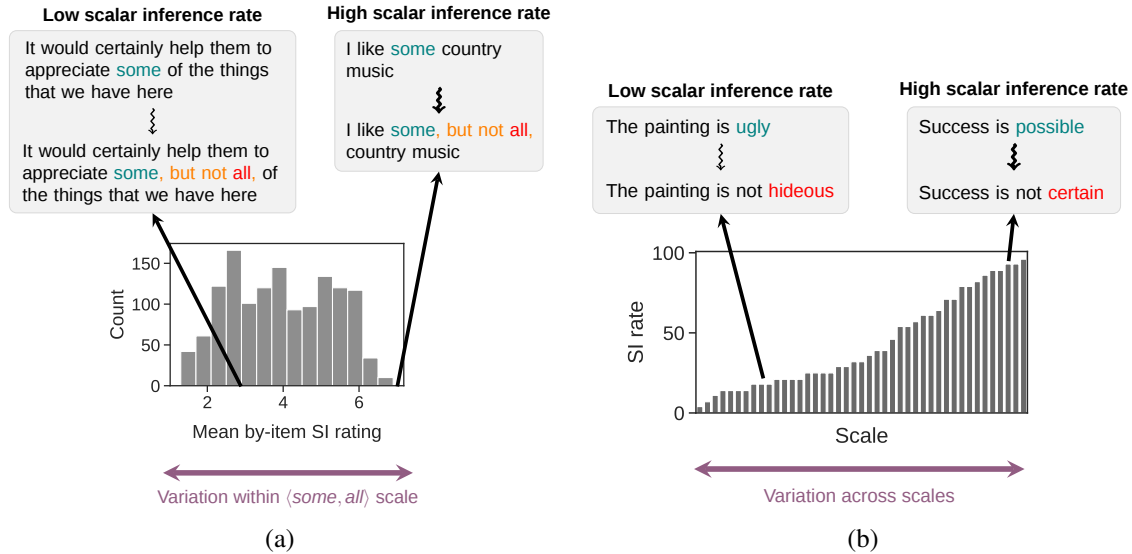


Figure 3-1: (a) Distribution of human scalar inference (SI) ratings (on scale of 1-7) across instances of the $\langle \text{some}, \text{all} \rangle$ scale (reproduction of Fig. 1, Degen 2015). (b) Average SI rates across scales formed by different lexical items (reproduction of Fig. 2, van Tiel et al. 2016).

scales formed by different lexical items (e.g., Doran et al., 2009; Beltrama and Xiang, 2013; van Tiel et al., 2016; Gotzner et al., 2018; Pankratz and van Tiel, 2021; Ronai and Xiang, 2022). For example, consider the following instances of the scale $\langle \text{some}, \text{all} \rangle$:

- (1)
 - a. I like some country music.
 - b. I like some, but not all, country music.
- (2)
 - a. It would certainly help them to appreciate some of the things that we have here.
 - b. It would certainly help them to appreciate some, but not all, of the things that we have here.

Degen (2015) finds that humans are highly likely to consider (1a) as conveying a similar meaning as (1b), but unlikely to consider (2a) as conveying a similar meaning as (2b) (Figure 3-1a). Similarly, consider the following instances of the scales $\langle \text{possible}, \text{certain} \rangle$ and $\langle \text{ugly}, \text{hideous} \rangle$, which both consist of adjectives ordered by entailment:

- (3)
 - a. Success is possible.
 - b. Success is not certain.

- (4) a. The painting is ugly.
- b. The painting is not hideous.

van Tiel et al. (2016) find that humans are highly likely to conclude that (3a) implies (3b), but unlikely to conclude that (4a) implies (4b) (Figure 3-1b).

While cross-scale and within-scale variation have typically been studied as distinct empirical phenomena, they both reflect gradedness in listener inferences based on alternatives and context. It therefore seems desirable to explain these empirical findings with a shared account, but there have been few proposals that quantitatively explain both within- and cross-scale variation. For example, cross-scale variation can be explained by intrinsic properties of the scale (e.g., whether the strong scalemate refers to an extreme endpoint; van Tiel et al., 2016), but these factors cannot explain variation within instances of a single scale. On the other hand, many factors explaining within-scale variance are scale-specific (e.g., the partitive “of the” for *⟨some, all⟩*; Degen, 2015) and may not generalize to new scales.

Here, we investigate a shared account of SI rates within and across scales. Since the alternatives are not explicitly produced (by definition), the listener has uncertainty over which alternatives the speaker could have used – and therefore, which strong scalemates ought to be negated through SI. Building upon constraint-based accounts of human language processing (Degen and Tanenhaus, 2015, 2016), we test the hypothesis that SIs depend on the availability of alternatives, which depend on context-driven expectations maintained by the listener. For example, if a speaker says “The movie was good”, the listener might predict that *amazing* is a more likely alternative than *funny* to the weak term *good*. An expectation-based view predicts that the listener would be thus be more likely to infer that the movie is not amazing (according to the speaker), and less likely to infer that the movie is not funny. However, while Degen and Tanenhaus (2015, 2016) have argued that listeners maintain context-driven expectations over alternatives, these studies have primarily investigated a single scale (*⟨some, all⟩*) in small domains, arguing from qualitative patterns and in the absence of a formal theory.

Furthermore, while it is generally assumed that SIs arise based on reasoning about unspoken alternatives, it remains debated whether humans reason about alternatives as

Dataset	Type of variation	# participants	# scales	# contexts per scale	# data points per item
Degen (2015)	Within-scale	243	1	1363	~ 10
Ronai and Xiang (2022)	Cross-scale	40	57	1	40
Pankratz and van Tiel (2021)	Cross-scale	1970	50	1	~ 40
Gotzner et al. (2018)	Cross-scale	220	67	1	40
van Tiel et al. (2016)	Cross-scale	28	39	3	10

Table 3.1: Details of human data used in our analyses. An item is a unique (scale, context) combination.

linguistic structures (e.g., Katzir, 2007; Fox and Katzir, 2011), or at the level of concepts (e.g., Gazdar, 1979; Buccola et al., 2021). Returning to the earlier example, if the weak scalemate is *good*, listeners may reason about a concept like VERYGOOD instead of a specific linguistic expression like *amazing*. In this sense, the listener’s uncertainty about alternatives might arise from uncertainty about both the scale itself (*Is the speaker implying the plot wasn’t amazing, or that the jokes weren’t funny?*), as well as the exact word forms under consideration by the speaker (*Is the speaker implying the movie wasn’t amazing, fantastic, or wonderful?*). Despite theoretical debates about the nature of alternatives, however, the role of concept-based alternatives in SI has not been tested in a systematic, quantitative way.

We provide a formalization of an expectation-based account of alternatives and test it on both string-based and concept-based views of alternatives. Instead of empirically estimating human expectations over alternatives (cf. Ronai and Xiang, 2022), we use neural language models as an approximation, which allows us to generate predictions for arbitrary sentences and contexts. We test the account’s predictions on human SI rates within the $\langle \textit{some}, \textit{all} \rangle$ scale (Degen, 2015), and across 148 scales from four datasets (van Tiel et al., 2016; Gotzner et al., 2018; Pankratz and van Tiel, 2021; Ronai and Xiang, 2022). We find support for the expectation-based account, and also provide the first evidence that concept-based alternatives may be underlying a wide range of SIs. Our results suggest that pragmatic inferences may arise from context-driven expectations over unspoken alternatives, and these expectations operate at the level of concepts.

3.2 Background

3.2.1 Within-scale variation

Within-scale variation refers to the variation in SI rates across instances of a single scale, such as $\langle \textit{some}, \textit{all} \rangle$. To explore SI variation within the scale $\langle \textit{some}, \textit{all} \rangle$, we use the dataset collected by Degen (2015), which features 1363 naturalistic sentences containing a “some”-NP from the Switchboard corpus of telephone dialogues (Godfrey et al., 1992) (Table 3.1). For each sentence, SI rates were measured using a sentence-similarity paradigm. On each trial, participants saw two sentence variants: the original sentence containing “some”, and a minimally differing sentence where “, but not all,” was inserted directly after “some”. Participants were asked, “How similar is the statement with ‘some, but not all’ to the statement with ‘some’?” and indicated responses (similarity judgments) on a seven point Likert scale. If the speaker’s originally intended meaning clearly includes an implicature, then making the implicature explicit by inserting “, but not all,” should not change the meaning of the sentence, so similarity judgments should be high. Thus, a higher similarity judgment indicates a stronger SI.

Degen (2015) finds substantial variation in SI rates across contexts, challenging the idea that the “some, but not all” inference arises reliably without sensitivity to context (Horn, 1989; Levinson, 2000). She also reports several features that predict SI rates, such as whether “some” occurs with the partitive “of the”, or whether the “some”-NP is in subject position. However, these features may be highly specific to the $\langle \textit{some}, \textit{all} \rangle$ scale, and it is unclear whether a more general mechanism may also explain variation within or across other scales.

3.2.2 Cross-scale variation (scalar diversity)

Cross-scale variation refers to the variation in SI rates across scales formed by different lexical items. To explore this, we use SI rates across 148 unique scales from four datasets, summarized in Table 3.1. Each scale involves a pair of English words (adjectives, adverbs, or verbs) of the form $\langle [\text{WEAK}], [\text{STRONG}] \rangle$, where $[\text{WEAK}]$ is less informative than $[\text{STRONG}]$

(e.g., *intelligent, brilliant*).¹ For each dataset, SI rates were measured through a binary choice task. Participants saw a character make a short, unembedded statement consisting of a simple noun phrase subject and a predicate with a weak scalar item (e.g., “John says: This student is intelligent.”). Their task was to indicate (*Yes* or *No*) whether they would conclude that the speaker believes the negation of a strong scalar item (e.g., “Would you conclude from this that, according to John, she is not brilliant?”). The SI rate for a scale is the proportion of *Yes* responses.

This method has revealed large variation in SI rates, ranging from 4% (*ugly, hideous*) to 100% (*sometimes, always*) (van Tiel et al., 2016). van Tiel et al. (2016) test two classes of factors that might predict SI rates: the availability of the strong scalemate given the weak scalemate, and the degree to which scalemates can be distinguished from each other. They find SI rates are predicted by measures of scalemate distinctness (e.g., whether the strong scalemate forms a fixed endpoint on the scale), but not by availability (but see Westera and Boleda, 2020; Ronai and Xiang, 2022). Other studies have proposed additional scale-intrinsic factors (e.g., Gotzner et al., 2018; Sun et al., 2018; Pankratz and van Tiel, 2021). However, structural properties of a scale cannot explain variability in SI rates *within* a scale, as these properties do not change across contexts.

While others have proposed context-dependent factors – which could, in principle, explain both cross- and within-scale variation – these factors often lack explanatory power in practice. For example, Ronai and Xiang (2021) find that the prominence of the Question Under Discussion (Roberts, 2012) is correlated with SI rates, but only for unbounded scales (i.e., scales where neither scalemate has a fixed, extreme meaning).

3.3 An expectation-based account of SI

Theoretically, it is the set of alternative utterances – utterances that the speaker could have used, but didn’t – that drive scalar implicature, and in principle every possible utterance in a language might be an alternative to every other. However, at an algorithmic level (Marr, 1982), it would be intractable for listeners to perform inference over this entire set.

¹We excluded scales where one of the items was formed by a multi-word expression (e.g., *may, have to*).

Furthermore, the signature pattern of SI would not arise without restrictions on the alternatives: otherwise, “[WEAK], but not [STRONG]” and “[STRONG]” would both be alternatives to “[WEAK]”, leading to contradictory inferences without a mechanism for breaking symmetry (Kroch, 1972; Katzir, 2007; Breheny et al., 2018).

To solve this symmetry problem, some approaches restrict alternatives based on structural complexity through grammar-internal mechanisms (e.g., Katzir, 2007; Fox and Katzir, 2011). However, these theories do not capture the uncertainty that listeners maintain, and are difficult to test quantitatively. Here, we test the view that listeners form probabilistic expectations over alternatives, given information from their interaction with the speaker. In the remainder of this section, we first discuss the conceptual predictions of an expectation-based account of SI, and then describe how we operationalize these predictions using neural language models.

Suppose that a listener hears a sentence with a weak scalar term [WEAK] (e.g., “This student is intelligent”). To rule out the meaning of a particular strong scalemate [STRONG] (e.g., the student is not *brilliant*), the listener must have reason to believe that the speaker would have said [STRONG] if they had intended to convey the strong meaning. However, since the alternatives are not explicitly produced, the listener has some degree of uncertainty over what alternatives were considered by the speaker. If it is likely that the speaker would have said [STRONG] to convey the strong meaning, then their choice to say [WEAK] suggests that they did not have grounds to say [STRONG] – and thus, an SI should be more likely to arise.

The key question, then, is how listeners estimate which alternatives are likely to be considered by the speaker. An expectation-based account proposes that listeners integrate contextual and grammatical cues to maintain probabilistic expectations over these alternatives. A scalemate that is more probable (given these cues) should be more likely to enter the scalar inference computation. Thus, this account predicts that the more expected the strong scalemate is as an alternative to the weak scalemate, the higher SI rates should be.

3.3.1 String-based view of alternatives

When an alternative is likely to be a strong scalemate, listeners should be more likely to rule out its meaning, resulting in higher SI rates. Conditioned on the context and the speaker’s choice to use [WEAK], the listener must estimate the probability of [WEAK] and [STRONG] being contrasted in a scalar relationship. Since it is difficult to directly estimate this probability, we construct a sentence frame where the probability of [STRONG] – at the level of forms – approximates the probability of [STRONG] being in a scalar relationship with a weak scalemate [WEAK]. This approach allows us to re-frame the problem of estimating listeners’ expectations over strong scalemates into a word prediction problem.

To do this, we use the scalar construction “*X, but not Y*”, which in many cases suggests that *Y* is a strong scalemate to *X* (Hearst, 1992; de Melo and Bansal, 2013; van Miltenburg, 2015; Pankratz and van Tiel, 2021). For a given utterance [CONTEXT] [WEAK] [CONTEXT] and hypothesized scale ⟨[WEAK], [STRONG]⟩, we form a sentence that explicitly states the SI:

$$\text{[CONTEXT] } \underbrace{\text{[WEAK], but not [STRONG]}}_{\text{scalar construction}} \text{, [CONTEXT]} \quad (3.1)$$

To test how expected [STRONG] is as an alternative to [WEAK], we need to estimate how likely a human would predict [STRONG] to appear in the [STRONG] position in (3.1).² Instead of attempting to directly measure these predictions (cf. Ronai and Xiang, 2022, see (3.3)), we approximate this with neural language models. We measure how unexpected [STRONG] is by computing its surprisal (negative log probability) under a language model, conditioned on the rest of the sentence. Since surprisal measures *unexpectedness*, we predict a negative relationship between SI rate and the surprisal of the strong scalemate.

This predictor is closely related to the notion of an SI’s “relevance” (Pankratz and van Tiel, 2021). Under usage-based theories of language (e.g., Tomasello, 2003; Bybee and Beckner, 2015), if a weak scalar term is encountered frequently in a scalar relationship with

²Another approach would be to measure the expectedness of [STRONG] in the template [CONTEXT] [STRONG] [CONTEXT] – that is, by replacing [WEAK] with [STRONG] in the speaker’s original utterance. This template would instantiate the theory that listeners determine alternatives based on the context. In contrast, the template we use in (3.1) instantiates the theory that listeners form expectations over alternatives based on the context as well as the speaker’s usage of [WEAK]. We return to this topic in Section 3.7.1.

a particular strong term, then the scalar relationship between these items will be enforced. Thus, Pankratz and van Tiel (2021) measure the relevance of an SI by counting corpus frequencies of the scalemates in the string “[WEAK], but not [STRONG].” This is conceptually aligned with our setup, where we might expect higher corpus frequencies to correspond to lower surprisal under a language model. However, our predictor differs from Pankratz and van Tiel’s in an important way: they aim to measure the “general relevance” of an SI, which they define as “relevance even in the absence of a situated context.” It is unclear how general relevance can explain variation in SI rates within instances of a scale. By using context-conditioned probabilities from a language model, our predictor could account for both the general frequency of “[WEAK], but not [STRONG]” as well as expectations driven by the context in which the scale occurs.

3.3.2 Concept-based view of alternatives

The method described above implicitly treats linguistic forms as the alternatives driving scalar inferences. However, recent proposals have advanced the view that alternatives are not linguistic objects, but instead operate at the level of more general reasoning preferences (Buccola et al., 2021). On this view, alternatives are constructed by replacing primitives of the concept expressed by the speaker with primitives of equal or less complexity.

Here, we test a generalization of this concept-based view of alternatives. Suppose, for example, a speaker uses the weak scalar term *big*. On a concept-based view, the listener may infer that the speaker is contrasting *big* with a concept like VERYBIG instead of a particular linguistic expression like *enormous*. However, in the experiments mentioned in Section 3.2.2, the SI process likely needs to be grounded in linguistic forms before the listener makes a judgment about a particular strong scalemate (in string form). One hypothesis is that upon hearing an expression with a weak scalemate, a stronger conceptual alternative is activated, which in turn probabilistically activates all the strings that could reflect it. Returning to our earlier example, if the conceptual alternative is VERYBIG, and *huge*, *massive*, and *enormous* are string-based realizations of that alternative, they may be assigned a high likelihood. When asked about a specific string-form alternative (e.g., “The

elephant is big. Would you conclude that it is not enormous?”), humans may endorse the SI if the probability of conceptually similar linguistic alternatives is sufficiently high, even if the probability of the tested alternative (here, *enormous*) is low.

If SIs involve reasoning about conceptual alternatives, then surprisal values estimated from assumed string-form alternatives may be poor estimates of the true relevant surprisal, as a single concept could be expressed with multiple forms. Therefore, in addition to assessing whether expectedness of specific linguistic forms predicts SI rates (Section 3.3.1), we also test a second predictor which approximates the expectedness of conceptual alternatives. To do this, we need a set of alternatives \mathcal{A} that could serve as potential linguistic scalemates. As described in more detail in Sections 3.4.3 and 3.5.3, we obtain \mathcal{A} by taking a fixed set of words with the same part of speech as the weak scalemate, inspired by grammatical theories of alternatives (e.g., Rooth, 1985; Katzir, 2007).³

Using this alternative set \mathcal{A} , we compute the weighted average surprisal of \mathcal{A} using weights determined by the conceptual similarity between each alternative and the tested strong scalemate. We use GloVe embeddings (Pennington et al., 2014) as an approximation for conceptual representations of scalar items, and cosine similarity between GloVe vectors to approximate conceptual similarity.

For each scale $\langle [\text{WEAK}], [\text{STRONG}] \rangle$, we obtain weights by computing the cosine similarity between the GloVe embeddings for $[\text{STRONG}]$ ($v_{[\text{STRONG}]}$) and each potential alternative a (v_a) in the alternative set \mathcal{A} . We compute the weighted average probability over \mathcal{A} using these weights, and then take the negative log to obtain the weighted average surprisal:

$$-\log \left(\frac{\sum_{a \in \mathcal{A}} P(a) \cdot \text{cossim}(v_{[\text{STRONG}]}, v_a)}{\sum_{a \in \mathcal{A}} \text{cossim}(v_{[\text{STRONG}]}, v_a)} \right) \quad (3.2)$$

If there are many conceptually similar alternatives with low surprisal, then the weighted average surprisal will be low, even if the surprisal of the tested scalemate is high. Therefore, weighted average surprisal forms a proxy for concept-based surprisal, which we compare to string-based surprisal.

³We adopt a liberal view of alternatives to avoid undergeneration. However, an important open question is how alternatives are determined, which we leave for future work.

3.4 Predicting variation within $\langle \textit{some}, \textit{all} \rangle$

3.4.1 Human data

To investigate variation within the scale $\langle \textit{some}, \textit{all} \rangle$, we use human SI strength ratings collected by Degen (2015). These ratings were measured by asking participants to rate the similarity (1-7) between a sentence with “some” and a minimally differing sentence with “some, but not all”. See Section 3.2.1 for details.

3.4.2 Model

Following the experiment conducted by Degen (2015), we construct scalar templates by inserting “, but not all,” after the occurrence of “some” in each sentence from the dataset. Since this scalar construction (“some, but not all,”) often occurs in the middle of the sentence, we use the bidirectional language model BERT (Devlin et al., 2019) to measure model expectations at the position of the strong scalemate. Concretely, we replace “all” with the [MASK] token and measure BERT’s probability distribution at that token. All models in our study are accessed via the Huggingface transformers library (Wolf et al., 2020).

3.4.3 Candidate alternatives

For our string-based surprisal predictor (Section 3.3.1), we are only concerned with the surprisal of the alternative *all* in the [STRONG] position in (3.1). However, to compute our concept-based surprisal predictor (Section 3.3.2), we need a set of candidate alternatives that could potentially serve as the strong scalemates implied by the speaker. Since the alternatives to *some* are highly constrained by the grammar, we manually constructed a set of English quantifiers that can be used in contrast to *some*: *each*, *every*, *few*, *half*, *much*, *many*, *most*, and *all*.

3.4.4 Results

Figure 3-2 shows the relationship between our predictors and human SI ratings for Degen’s (2015) dataset of variation within $\langle \textit{some}, \textit{all} \rangle$. We find that both string-based and concept-

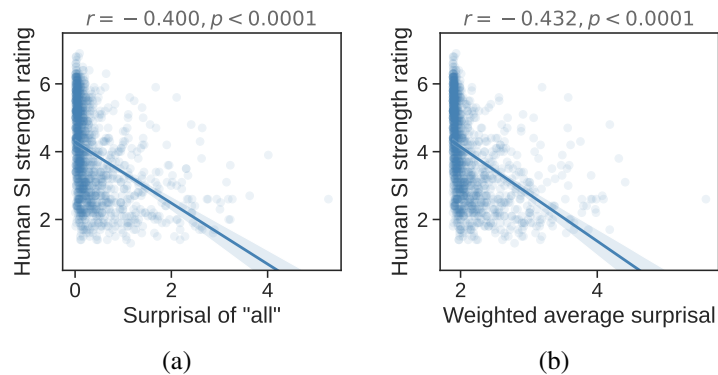


Figure 3-2: Relationship between human SI strength ratings within $\langle some, all \rangle$ scale (Degen, 2015) and BERT-derived predictors: (a) surprisal of scalemate *all* in the scalar construction, and (b) weighted average surprisal over the full set of candidate alternatives (Section 3.4.3). Each point represents a sentence. Shaded region denotes 95% CI.

based surprisal are indeed negatively correlated with human similarity judgments (string-based: Figure 3-2a, Pearson $\rho = -0.400$, $p < 0.0001$; concept-based: Figure 3-2b, $\rho = -0.432$, $p < 0.0001$).⁴

We additionally conducted a multivariate analysis including our two new predictors (string- and concept-based surprisal) among the predictors investigated in Degen’s original study. We centered and transformed all variables according to Degen’s original analyses. The results are summarized in Table 3.2. We find that the original predictors remain statistically significant, and that concept-based surprisal (but not string-based surprisal) is a significant predictor in the full model. This suggests that listeners draw stronger scalar inferences when *all* – or a conceptually similar alternative – is more expected in a given context.

3.5 Predicting variation across scales

3.5.1 Human data

To investigate variation across scales, we use human SI rates collected by four studies (Ronai and Xiang, 2022; Pankratz and van Tiel, 2021; Gotzner et al., 2018; van Tiel et al., 2016). SI

⁴We note that the relationship between surprisal and SI ratings appears highly non-linear in Figure 3-2a. We expect this is due to the fact that the scalemate *all* is highly expected in most contexts, so the surprisal values of *all* are concentrated near zero. There is a stronger linear relationship between SI ratings and raw probabilities ($\rho = 0.482$, $p < 0.0001$).

Predictor	β	p
DEGEN (2015) PREDICTORS		
Partitive	0.658	< 0.0001
Strength	-0.470	< 0.0001
Mention	0.287	< 0.0001
Subjecthood	0.495	< 0.0001
Modification	0.157	< 0.01
Log sentence length	0.189	< 0.0001
OUR PREDICTORS		
String-based surprisal	0.008	0.960
Concept-based surprisal	-0.782	< 0.001

Table 3.2: Summary of full regression model predicting variation within $\langle some, all \rangle$, including original predictors from Degen (2015) (see the original study for a detailed description of each of the predictors).

POS	# unique	Form of original sentence	Form of scalar construction	Example
Adj	120	[NP] is [WEAK]	[NP] is [WEAK], but not [STRONG]	The elephant is big , but not enormous
Adv	12	[NP] is [WEAK] [ADJ]	[NP] is [WEAK] [ADJ], but not [STRONG]	The director is some-times late, but not always
Verb	16	[NP] [WEAK]-ed	[NP] [WEAK]-ed, but did not [STRONG]	The runner started , but did not finish

Table 3.3: Scalar construction templates for different parts of speech (for cross-scale variation).

rates were measured by showing participants a sentence with the weak scalemate (e.g., “The student is intelligent”), and asking whether they would endorse the negation of the strong scalemate (e.g., “The student is not brilliant”). See Section 3.2.2 for details.

3.5.2 Model

We construct scalar templates following the pattern summarized in Table 3.3. Since in each case the strong scalemate is the final word in the sentence,⁵ we use an autoregressive language model to measure expectations over potential scalemates in the [STRONG] position. We use the base GPT-2 model (Radford et al., 2019) via Huggingface and obtain model surprisals through the SyntaxGym command-line interface (Gauthier et al., 2020).

⁵For a small number of verbal scales, the strong scalemate is followed with the pronoun “it” to make the sentence grammatical. We don’t expect this to matter for our purposes.

3.5.3 Candidate alternatives

Recall from Section 3.3.2 that we need a set of potential linguistic alternatives to compute the weighted average surprisal. We take this set of alternatives to be a set of words with the same part of speech (POS) as the weak scalemate and obtain these candidate alternative sets by extracting lists of English adjectives, adverbs, and verbs from WordNet (Miller, 1995). We then used NLTK (Loper and Bird, 2002) to find the words satisfying finer-grained POS tags (JJ for adjectives, RB for adverbs, and VB for verbs), and sorted each POS set according to word frequencies from the OpenSubtitles corpus (Lison and Tiedemann, 2016).^{6,7} We excluded words in the POS sets that were not in the frequency corpus, resulting in 3204 adjectives, 1953 adverbs, and 226 verbs. We restricted each POS set to its 1000 highest-frequency words, and performed some manual exclusions (e.g., removing “do” and “be” from the verb set, which are unlikely to form scales with any of the tested items and follow different syntactic rules). This finally resulted in our three alternative sets: 1000 adjectives, 960 adverbs, and 224 verbs.⁸

3.5.4 Results

String-based analyses

Figure 3-3a shows our results for cross-scale variation, under a string-based view of alternatives. We find that surprisal is a significant predictor only for Ronai and Xiang’s dataset (Pearson $\rho = -0.361$, $p = 0.006$).⁹

Model surprisal vs. human completions. For the dataset where we do find a relationship between surprisal and SI rates, we ask whether model surprisals are correlated with human-derived measurements of how “accessible” the strong scalemate is. If model surprisals and human accessibility scores are strongly linked, this would suggest that models and humans

⁶<https://github.com/hermitdave/FrequencyWords>

⁷<http://www.opensubtitles.org>

⁸Most words in the alternative sets occur with low frequency, but we chose to be liberal when including alternatives to ensure broad coverage over potential scalemates.

⁹We repeated this analysis after removing an outlier from Gotzner et al.’s dataset, and again found a lack of relationship between SI rate and surprisal ($\rho = -0.0452$, $p = 0.719$).

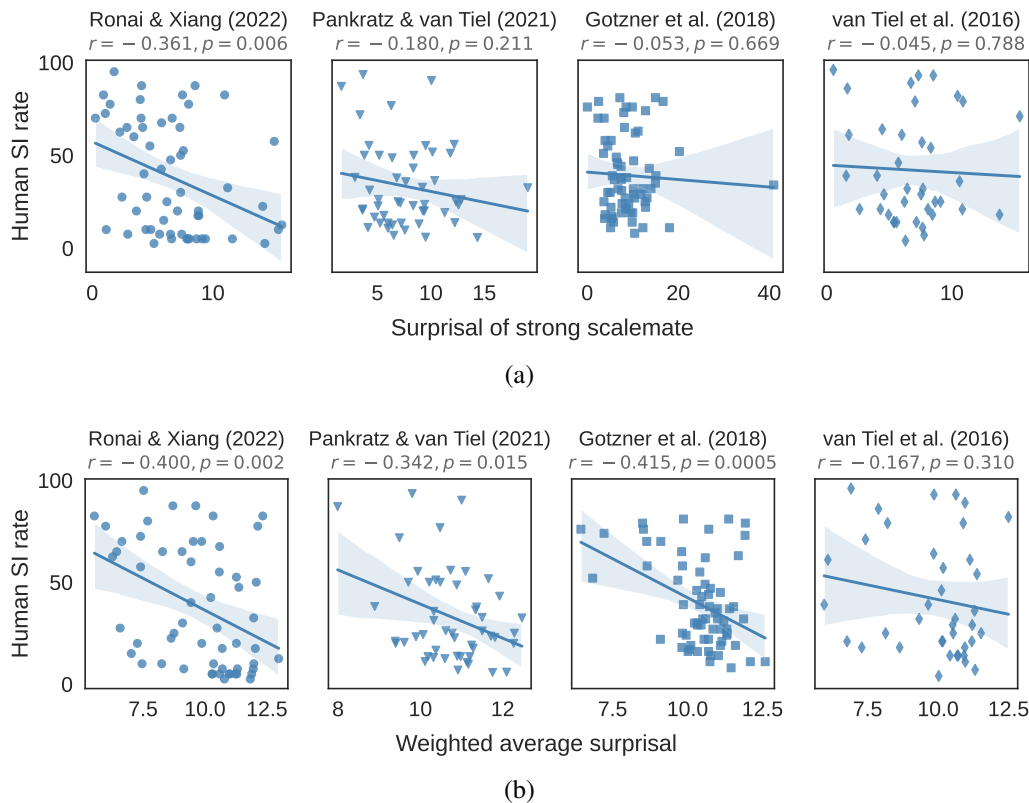


Figure 3-3: Relationship between human SI rates and GPT-2-derived predictors across scales, for four datasets. Each point represents a single scale. Shaded region denotes 95% CI. (a) SI rate vs. surprisal of strong scalemate in the scalar construction. (b) SI rate vs. weighted average surprisal over the full set of candidate alternatives (Section 3.5.3).

are aligned at the level of predictive distributions over alternatives, validating our approach of using language models to approximate human predictions.

To this end, we use data from Ronai and Xiang’s Experiment 2, which measured the accessibility of scalemates through a Cloze task. Humans were presented with a short dialogue featuring a sentence with the weak scalemate, as in (3.3), and then asked to generate a completion of the dialogue in the blank. The “accessibility” of the strong scalemate is taken to be the frequency with which it is generated in this paradigm.

Sue: The movie is good. (3.3)
 Mary: So you mean it’s not _____.

We find that model surprisals are negatively correlated with accessibility scores (Figure 3-

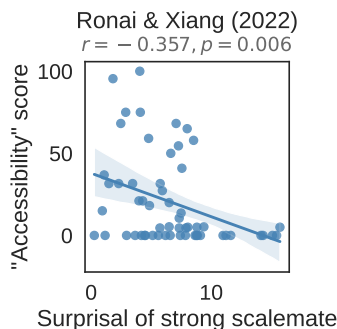


Figure 3-4: GPT-2-derived surprisal of strong scalemate vs. accessibility rating of strong scalemates (Ronai and Xiang, 2022).

Dataset	Full model			ANOVA	
	Predictor	β	p	F	p
Ronai and Xiang (2022)	String-based surprisal	-1.538	0.215	3.247	0.012
	Concept-based surprisal	-4.503	0.065		
Pankratz and van Tiel (2021)	String-based surprisal	0.460	0.694	3.198	0.050
	Concept-based surprisal	-9.491	0.036		
Gotzner et al. (2018)	String-based surprisal	0.384	0.545	2.751	0.019
	Concept-based surprisal	-8.010	0.0005		
van Tiel et al. (2016)	String-based surprisal	0.293	0.858	1.016	0.422
	Concept-based surprisal	-3.340	0.291		

Table 3.4: Summary of full regression model (middle columns) and ANOVA comparing full model against intercept-only model (right columns) for each cross-scale variation dataset.

4; $\rho = -0.357, p = 0.006$), suggesting that our method of estimating expectations over alternatives using artificial language models aligns with direct measurements in humans.

Concept-based analyses

Turning to a conceptual view of alternatives, Figure 3-3b shows the relationship between human SI rates and weighted average surprisals (Equation 3.2). We find a significant negative correlation for all but one of the tested datasets (Ronai and Xiang: $\rho = -0.400, p = 0.002$; Pankratz and van Tiel: $\rho = -0.342, p = 0.015$; Gotzner et al.: $\rho = -0.415, p = 0.0005$; van Tiel et al.: $\rho = -0.167, p = 0.310$), demonstrating that similarity-weighted surprisal captures more variation than raw surprisal (cf. Figure 3-3a; Section 3.5.4).

We additionally included both (centered) string-based and concept-based surprisal as predictors in a multivariate model, summarized in Table 3.4 (middle columns). As in the

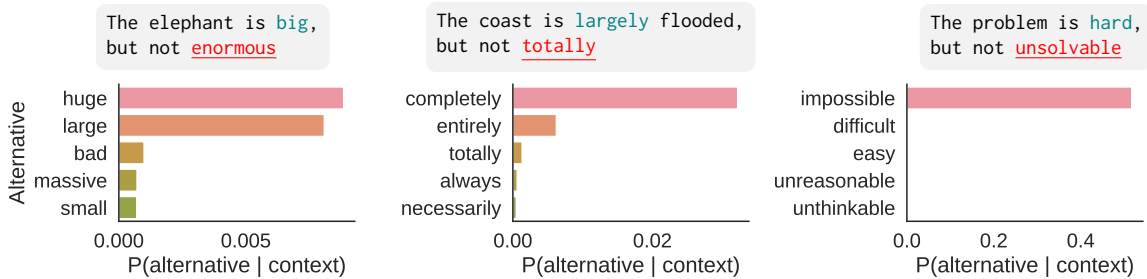


Figure 3-5: Probability assigned by GPT-2 to top 5 candidate strong alternatives (y-axis) for 3 example weak scalar items: *big*, *largely*, and *hard* (Ronai and Xiang, 2022). The full scalar construction is shown above each subplot, with the original tested strong scalemate underlined in red.

within-scale analysis, for three of the four datasets we find that concept-based surprisal is a stronger predictor than string-based surprisal. With that said, we find only a marginal effect of concept-based surprisal in Ronai and Xiang’s data, and no effect of either predictor in van Tiel et al.’s data. However, for Ronai and Xiang’s data, this does not mean that there is no value in either predictor – rather, the predictors are too closely correlated to definitively favor one over the other. To demonstrate this, for each dataset we performed an analysis of variance (ANOVA) comparing the full model to a null intercept-only model (Table 3.4, right columns). We find that for all datasets except that of van Tiel et al., the model with both surprisal predictors explains significantly more variance than the null model. In sum, our results suggest that the expectedness of the strong scalemate can capture significant cross-scale SI variation, but these expectations may operate over groups of semantically similar linguistic forms instead of individual strings.

Qualitative analysis. As a follow-up analysis, we identified cases where GPT-2 assigns low probability to the tested strong scalemate, but high probability to near synonyms. We analyzed the top 5 alternatives from the full alternative set (Section 3.5.3) that were assigned highest probability as strong scalemates under GPT-2. Figure 3-5 shows three examples from Ronai and Xiang’s dataset. The title of each subplot shows the scalar construction, with the weak scalemate highlighted in teal and the tested strong scalemate underlined in red. The y-axis shows the top 5 candidate scalemates, and the x-axis shows the probability assigned by the model. For the weak scalemate *big* (left), GPT-2 assigns

highest probability to the alternative *huge*, which semantically conveys similar information to the empirically tested alternative *enormous*. We see a similar pattern for weak scalemate *largely* and alternatives *completely* and *totally* (middle), as well as for weak scalemate *hard* and alternative *impossible* (right). This is consistent with the hypothesis that surprisal of a specific string may not capture surprisal of the underlying concept.

Taken together, these analyses suggest that a concept-based view of alternatives is better aligned with human inferences than treating alternatives as specific linguistic forms. Testing additional ways of operationalizing concept-based alternatives is a promising direction for future work.

3.6 Related work

Prior work has evaluated the ability of computational models to capture scalar inferences. For example, the IMPPRES benchmark (Jeretic et al., 2020) frames SI as a natural language inference problem: the weak scalar expression (e.g., “Jo ate some of the cake”) is the premise, and the negated strong scalar expression (e.g., “Joe didn’t eat all of the cake”) is the hypothesis. Under this setup, an interpretation consistent with the strictly logical reading would assign a *neutral* relationship between the premise and hypothesis, whereas a pragmatic reading would assign an *entailment* relationship. Models are evaluated based on how often they assign the entailment label across items, which treats SIs as a homogeneous phenomenon and does not capture SI variation.

Another line of work has attempted to predict within-scale SI variation through a supervised approach (Schuster et al., 2020; Li et al., 2021). This approach takes a sentence with a weak scalar item, and attempts to directly predict the human SI strength through a prediction head on top of a sentence encoder. This differs from our approach in that it requires training directly on the SI-rate-prediction task, whereas we probe the predictive distribution that emerges from language modeling with no task-specific representations. This allows us to compare model probability distributions to the expectations deployed by humans during pragmatic inferences, building upon a literature linking language models to predictive processing (e.g., Frank and Bod, 2011; Smith and Levy, 2013; Wilcox et al.,

2020; Merkx and Frank, 2021).

There have also been several studies extracting scalar orderings from corpora or language model representations. For example, de Marneffe et al. (2010) use distributional information from a web corpus to ground the meanings of adjectives for an indirect question answering task. Similarly, Shivade et al. (2015) use scalar constructions like “*X, but not Y*” to identify scales from a corpus of biomedical texts. Others have found that adjectival scale orderings can be derived from static word embeddings (Kim and de Marneffe, 2013) and contextualized word representations (Garí Soler and Apidianaki, 2020, 2021).

3.7 Discussion

We tested a shared mechanism explaining variation in SI rates across scales and within $\langle \textit{some}, \textit{all} \rangle$, based on the hypothesis that humans maintain context-driven expectations about unspoken alternatives (Degen and Tanenhaus, 2015, 2016). We operationalized this in two ways using neural language models: the expectedness of a linguistic alternative as a scale-mate (string-based surprisal), and the expectedness of a conceptual alternative (weighted average surprisal). We found that for both within-scale and cross-scale variation, expectedness captures human SI rates. Crucially, however, expectedness of the strong scalemate is a robust predictor of cross-scale variation only under a conceptual view of alternatives (Buccola et al., 2021). Our results support the idea that the strength of pragmatic inferences depends on the availability of alternatives, which depends on in-context predictability.

One open question is the source of variability across the tested human behavioral datasets – in particular, the lack of surprisal effect for van Tiel et al.’s data (Section 3.5.4). While we cannot be certain about why the results vary, we identified a few differences that might affect data quality across datasets (see Table 3.1). van Tiel et al.’s study has the smallest number of participants (28), smallest number of ratings per scale (10), and smallest number of scales (39). In addition, their experiments presented multiple sentence contexts per scale, whereas the other experiments only presented one sentence per scale. Other experimental factors, such as participant recruitment and exclusion criteria, may have also contributed to differences in data reliability.

3.7.1 How do listeners restrict the alternatives?

We now return to the issue raised in Footnote 2: what information do listeners use to form expectations about alternatives? To illustrate potential hypotheses, consider the item “The soup is warm/hot” from van Tiel et al.’s experimental materials. In our framework described in Section 3.3.1, [CONTEXT] = “The soup is”, [WEAK] = “warm”, and [STRONG] = “hot”. One hypothesis is that listeners form expectations over relevant scalar expressions given [CONTEXT] alone. On this view, expectations over strong scalemates could be measured by computing the probability of [STRONG] in the template [CONTEXT][STRONG]; i.e., “The soup is [STRONG]”. In contrast, in this paper we test expectations of [STRONG] in the template “The soup is warm, but not [STRONG]”, which instantiates an alternate theoretical position: that listeners use not only the context, but also [WEAK] as information for forming expectations over alternatives.

We adopt this view for several reasons. First, it could be the case that the context does not provide enough information for the listener to narrow down alternatives. Returning to the running example, “The soup is” could be followed by many continuations, some potentially relating to the taste or size of the soup in addition to its temperature. Taking the weak scalar term “warm” into account allows the listener to restrict the relevant alternatives to a smaller, more tractable set, which presents an algorithmic solution to the computationally challenging inference problem. However, the underinformativity of the context may be a problem unique to the simple stimuli used in the behavioral experiments. It is plausible that listeners could sufficiently restrict alternative sets given more naturalistic contexts, which likely provide more cues to the Question Under Discussion (Roberts, 2012).

In addition, there could be cues from [WEAK] that provide information about likely alternatives, independent of the context. For example, listeners might prefer strong scalemates that match [WEAK] in register or formality, or in shared phonological features. This motivates why we chose template (3.1) to measure expectations over alternatives, instead of [CONTEXT][STRONG]. However, the extent to which listeners tune their predictions based on [WEAK] above and beyond the context remains an open empirical question.

3.7.2 From alternatives to inference

Conceptually, computing an SI involves two steps: (1) determining the suitable alternatives, and (2) ruling out the meaning of alternatives to arrive at a strengthened interpretation of the weak scalar term. Our results primarily shed light on the first step, providing evidence that expectations play a role in determining alternatives, and that alternatives are likely based on meanings in addition to linguistic forms.

When considering the higher-level reasoning process, many factors beyond alternatives play a causal role in SI. One view is that humans use alternatives in a cooperative reasoning process, such as that formalized by the Rational Speech Act framework (RSA; Frank and Goodman, 2012; Goodman and Frank, 2016). In an RSA model, a pragmatic listener $L_1(m | u)$ uses a speaker’s utterance u to update their prior beliefs $P(m)$ over which meaning m the speaker is trying to convey. The listener does this by computing the likelihood of a pragmatic speaker S_1 producing u given each potential meaning. The pragmatic S_1 speaker corresponds to the utility U of the utterance u to convey m , relative to the utility of the alternative utterances in the set of alternatives \mathcal{A} :

$$L_1(m | u) = \frac{S_1(u | m)P(m)}{\sum_{m'} S_1(u | m')P(m')} \quad (3.4)$$

$$S_1(u | m) = \frac{U(u, m)}{\sum_{u' \in \mathcal{A}} U(u', m)} \quad (3.5)$$

Our findings appear compatible with RSA: listeners reason about a speaker that normalizes over alternatives. However, it remains an open question how variable expectations over alternatives should be operationalized in an RSA model. One option, as recently proposed by Zhang et al. (2023), is that the pragmatic speaker is conditioned on the alternative set \mathcal{A} . The pragmatic listener has beliefs over different sets of \mathcal{A} and marginalizes over these beliefs when drawing an inference:

$$L_1(m | u) = \sum_{\mathcal{A}} P(\mathcal{A}) \frac{S_1(u | m, \mathcal{A})P(m)}{\sum_{m'} S_1(u | m', \mathcal{A})P(m')} \quad (3.6)$$

Another possibility is that the variable expectations are not inputs to the model, but instead fall out of reasoning about how likely speakers are to use the weaker versus stronger terms, given variable contextual priors over meanings and questions under discussion (see, e.g., Goodman and Lassiter, 2015; Qing et al., 2016). We leave a detailed exploration of such a model to future work.

The role of priors. Pragmatic influences are influenced by the prior probabilities of the world states compatible with the weak and strong meanings Degen et al. (2015); Sikos et al. (2021). For example, consider the scale $\langle start, finish \rangle$. If a human were asked “The movie started at 2:30. Would you conclude that the movie did not finish at 2:30?”, they would likely answer *Yes*. This *Yes* response would count as an SI under the experimental paradigm, but does not reflect pragmatic reasoning over scalar alternatives: it is simply implausible for a movie to start and finish at the same time, given our knowledge of the world.¹⁰

These priors have an important connection to our analyses. As outlined in Section 3.3.1, we approximate the expectedness of a strong scalemate by measuring the expectedness of its linguistic form. This approach can be seen as reflecting an implicit assumption that the more likely a certain meaning is, the more likely it is to be expressed linguistically. This is likely to be wrong in certain cases – for example, if a certain meaning is so likely that it is obvious without being said, then speakers may avoid the effort of explicitly producing the linguistic expression (and thus, the linguistic expression would have low probability). This could potentially be the case for relatively common SIs. For example, a speaker might be able to get away with only saying *some* and expecting a listener to recover the meaning *some but not all*.

With that said, we believe our estimation method may minimize this issue, as we measure expectations conditioned on an explicit scalar contrast with the weak scalemate (i.e., “[WEAK], but not”). Thus, our approach can be seen as approximating listeners’ expectations about upcoming linguistic material, given that the speaker has *already chosen* to produce a scalar contrast. Nevertheless, a complete account of scalar inferences will need to account for the influence of the prior probabilities over world states, which may explain

¹⁰This example is due to Lassiter (2022).

some of the variance not captured by our expectedness predictors.

3.7.3 Implications for NLP

While the main role of language models in our analyses was to systematically test a cognitive theory, we believe this work also has implications for NLP evaluation. A growing body of work uses controlled assessments to evaluate the linguistic knowledge of NLP models. Many studies test whether models exhibit a categorical pattern of behavior that reflects a particular linguistic generalization. For example, in syntactic evaluations, a model is successful if it satisfies certain inequality relationships between grammatical and ungrammatical sentences (e.g., Linzen et al., 2016; Futrell et al., 2019; Hu et al., 2020b). SI (and other types of implicatures) have largely been treated the same way (see Section 3.6).

In contrast, we do not evaluate whether language models exhibit a categorical pattern of behavior (“*Do models interpret SIs pragmatically?*”). Instead, based on the empirical evidence for scalar variation, we test whether models capture systematic variability in human inferences (“*Are models sensitive to the factors that modulate human pragmatic inferences?*”). We urge other NLP researchers to consider variability in human behaviors instead of relying on categorical generalizations (see also Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022; Baan et al., 2022; Webson et al., 2023). Through this approach, we can build models that capture the rich variability of human language, and use these models to refine our theories about the human mind.

Chapter 4

A fine-grained comparison of pragmatic language understanding in humans and language models

4.1 Introduction

Non-literal language understanding is an essential part of communication. For example, in everyday conversations, humans readily comprehend the non-literal meanings of metaphors (*My new coworker is a block of ice*), polite deceits (*I love the gift*), indirect requests (*It's a bit cold in this room*), and irony (*Classy pajamas, dude!*). These phenomena fall under the broad label of **pragmatics**, which encompasses the aspects of meaning that go beyond the literal semantics of what is said (Horn, 1972; Grice, 1975; Yule, 1996; Levinson, 2000).

A long-standing challenge for NLP is to build models that capture human pragmatic behaviors. The remarkable abilities of modern language models (LMs) have triggered a recent effort to investigate whether such models capture pragmatic meaning, both through philosophical arguments (Bisk et al., 2020; Bender and Koller, 2020; Potts, 2020; Michael, 2020) and empirical evaluations (Jeretic et al., 2020; Zheng et al., 2021; Tong et al., 2021; Liu et al., 2022; Ruis et al., 2022; Stowe et al., 2022). However, prior empirical studies have primarily evaluated LMs based on a binary distinction between pragmatic and non-pragmatic

responses, providing limited insights into models’ weaknesses. A model could fail to reach the target pragmatic interpretation in multiple ways – for example, by preferring a literal interpretation, or by preferring a non-literal interpretation that violates certain social norms. Understanding these error patterns can suggest specific directions for improving the models, and foreshadow where pragmatics might go awry in user-facing settings (e.g., Saygin and Cicekli, 2002; Dombi et al., 2022; Kreiss et al., 2022).

From a cognitive perspective, understanding the pragmatic abilities of LMs could also offer insights into humans. Human pragmatic language comprehension involves a variety of mechanisms, such as basic language processing, knowledge of cultural and social norms (Trosborg, 2010), and reasoning about speakers’ mental states (Brennan et al., 2010b; Enrici et al., 2019; Rubio-Fernández, 2021). However, it remains an open question when language understanding relies on explicit mentalizing – which may be cognitively effortful – versus lower-cost heuristics (e.g., Butterfill and Apperly, 2013; Heyes, 2014). Since LMs lack explicit, symbolic representations of mental states, they can serve as a tool for investigating whether pragmatic phenomena arise without full-blown mentalizing (e.g., belief updates in the Rational Speech Act framework; Frank and Goodman, 2012).

In this paper, we perform a fine-grained comparison of humans and LMs on pragmatic language understanding tasks. Adopting the approach of targeted linguistic evaluation (e.g., Linzen et al., 2016; Futrell et al., 2019; Hu et al., 2020b), our analysis serves two goals: assessing the pragmatic capabilities of modern LMs, and revealing whether pragmatic behaviors emerge without explicitly constructed mental representations. Our test materials are a set of English multiple-choice questions curated by expert researchers (Floyd et al., In prep), covering seven diverse pragmatic phenomena. We use zero-shot prompting to evaluate models with varying sizes and training objectives: GPT-2 (Radford et al., 2019), *Tk*-Instruct (Wang et al., 2022), Flan-T5 (Chung et al., 2022), and InstructGPT (Ouyang et al., 2022).

Through model analyses and human experiments, we investigate the following questions: (1) Do models recover the hypothesized pragmatic interpretation of speaker utterances? (2) When models do not select the target response, what errors do they make – and how do these error patterns compare to those of humans? (3) Do models and humans use

similar cues to arrive at pragmatic interpretations? We find that Flan-T5 (XL) and OpenAI’s text-davinci-002 achieve high accuracy and mirror the distribution of responses selected by humans. When these models are incorrect, they tend to select the incorrect literal (or straightforward) answer instead of distractors based on low-level heuristics. We also find preliminary evidence that models and humans are sensitive to similar linguistic cues. Our results suggest that some pragmatic behaviors emerge in models without explicitly constructed representations of agents’ mental states. However, models perform poorly on humor, irony, and conversational maxims, suggesting a difficulty with social conventions and expectations.

4.2 Related work

Prior work has evaluated LMs’ ability to recognize non-literal interpretations of linguistic input, such as scalar implicature (Jeretic et al., 2020; Schuster et al., 2020; Li et al., 2021) or figurative language (Tong et al., 2021; Liu et al., 2022; Gu et al., 2022; Stowe et al., 2022). In a broad-scale evaluation, Zheng et al. (2021) test five types of implicatures arising from Grice’s (1975) conversational maxims, and evaluate their models after training on the task. In our work, we consider Gricean implicatures as one of seven phenomena, and we evaluate pre-trained LMs without fine-tuning on our tasks.

Similar to our work, Ruis et al. (2022) also use prompting to evaluate LMs on pragmatic interpretation tasks. They formulate implicature tests as sentences ending with “yes” or “no” (e.g., “Esther asked “Can you come to my party on Friday?” and Juan responded “I have to work”, which means no.”). A model is considered pragmatic if it assigns higher probability to the token that makes the sentence consistent with an implicature. In our work, models must select from multiple interpretations, enabling a detailed error analysis and comparison to humans. Ruis et al.’s materials also focus on indirect question answering as an implicature trigger, whereas we consider a broader range of pragmatic phenomena and utterance types.

Since pragmatic language understanding often draws upon knowledge of social relations, our tasks are conceptually related to benchmarks for evaluating social commonsense (e.g.,

Sap et al., 2019; Zadeh et al., 2019). These evaluations focus on the interpretation of actions and events, whereas we focus on the interpretation of speaker utterances. Another hypothesized component of pragmatics is Theory of Mind (ToM; Leslie et al., 2004; Apperly, 2011), or the ability to reason about others’ mental states. Benchmarks for evaluating ToM in models (e.g., Nematzadeh et al., 2018; Le et al., 2019; Sap et al., 2022) primarily focus on false-belief tasks (Baron-Cohen et al., 1985), which assess whether a model can represent the beliefs of another agent that are factually incorrect but consistent with that agent’s observations. LMs have been shown to succeed on some ToM tests (Kosinski, 2023) while failing on others (Sap et al., 2022; Ullman, 2023).

4.3 Evaluation materials

4.3.1 Overview of stimuli

Our evaluation materials are taken from Floyd et al.’s (In prep) experiments,¹ covering seven phenomena. Each item is a multiple choice question, with answer options representing different types of interpretation strategies. For most of the tasks, the question has three parts: a short story context (1-3 sentences), an utterance by one of the characters, and a question about what the character intended to convey.² Table 4.1 shows an example item for each task, with annotated answer options. Green labels indicate the target pragmatic interpretation.³ Blue labels indicate the literal interpretation. Red labels indicate incorrect non-literal interpretations, which are based on heuristics such as lexical similarity to the story, thus serving as distractor options.

Each task has 20-40 items, which were manually curated by expert researchers to cover a broad range of non-literal phenomena and elicit individual differences among humans. The stimuli were not specifically designed to require Theory of Mind reasoning (ToM). However, behavioral and neural evidence suggests that many of the tested phenomena rely

¹https://osf.io/6abgk/?view_only=42d448e3d0b14ecf8b87908b3a618672

²The exceptions are Humor and Coherence.

³We refer to these answer options as “Correct” throughout the paper. However, these answers are only “correct” in the sense of a normative evaluation. We acknowledge the wide variation in individual humans’ abilities and tendencies to use non-literal language, which is not captured in our analyses. We thank an anonymous reviewer for highlighting this point.

Task	Example query	Example answer options
Deceits	Henry is sitting at his desk and watching TV, and reluctantly switches off the TV with the remote control and picks up a text-book. Shortly after, his mother comes in the room and asks, "What have you been doing up here?" Henry responds: "Reading." Why has Henry responded in such a way?	<ol style="list-style-type: none"> 1. Correct He does not want to get into trouble for not studying. 2. Literal He has been reading for some time. 3. DistractorLexicalOverlap He does not want to offend his mom by not reading the books that she gave him. 4. DistractorSocialConvention He wants his mom to believe that he has been watching TV.
Indirect speech	Nate is about to leave the house. His wife points at a full bag of garbage and asks: "Are you going out?" What might she be trying to convey?	<ol style="list-style-type: none"> 1. Correct She wants Nate to take the garbage out. 2. Literal She wants to know Nate's plans. 3. DistractorAssociative She wants Nate to bring his friends over. 4. DistractorLexicalOverlap She wants Nate to spend more time with the family.
Irony	It is a holiday. Stefan and Kim are sitting in the backseat of the car. They are fighting all the time. Their father says: "Oh, it is so pleasant here." What did the father want to convey?	<ol style="list-style-type: none"> 1. Correct He does not want to listen to his kids' arguments. 2. Literal He enjoys listening to his kids fighting. 3. DistractorAssociative AC gives them some needed cool. 4. DistractorNonSequitur He remembers about his wife's birthday.
Maxims	Leslie and Jane are chatting at a coffee shop. Leslie asks, "Who was that man that I saw you with last night?" Jane responds, "The latte is unbelievable here." Why has Jane responded like this?	<ol style="list-style-type: none"> 1. Correct She does not want to discuss the topic that Leslie has raised. 2. Literal She thinks that it is the best latte in the town. 3. DistractorAssociative The man who Leslie saw makes unbelievable lattes. 4. DistractorNonLiteral A coffee break is not a good time to discuss men.
Metaphor	Andrew and Bob were discussing the investment company where Andrew works. Bob said: "The investors are squirrels collecting nuts." What does Bob mean?	<ol style="list-style-type: none"> 1. Correct They buy stocks hoping for future profit. 2. Literal Squirrels were hired to work in the company. 3. DistractorNonLiteral The investors dress and eat well. 4. DistractorNonSequitur Bob is allergic to nuts. 5. DistractorPlausibleLiteral The investors enjoy picking nuts as much as squirrels do.
Humor	Martha walked into a pastry shop. After surveying all the pastries, she decided on a chocolate pie. "I'll take that one," Martha said to the attendant, "the whole thing." "Shall I cut it into four or eight pieces?" the attendant asked.	<ol style="list-style-type: none"> 1. Correct Martha said, "Four pieces, please; I'm on a diet." 2. Literal Martha said: "Well, there are five people for dessert tonight, so eight pieces will be about right." 3. DistractorAssociative Martha said, "You make the most delicious sweet rolls in town." 4. DistractorFunny Then the attendant squirted whipped cream in Martha's face. 5. DistractorNeutral Martha said, "My leg is hurting so much."
Coherence	Cleo brushed against a table with a vase on it. She decided to study harder to catch up.	<ol style="list-style-type: none"> 1. Correct Incoherent 2. Incorrect Coherent

Table 4.1: Sample item from each task in our evaluation. All items are originally curated by Floyd et al. (In prep).

on mentalizing processes. In Section 4.3.2, we briefly describe the role of ToM for each tested phenomenon, and how LMs’ training corpora may provide linguistic cues to perform the tasks.

4.3.2 Tested phenomena

Deceits. Humans produce polite deceits (“white lies”) in the service of social and personal relationships (e.g., Camden et al., 1984). Behavioral studies in young children suggest that understanding white lies requires interpretive ToM, or the ability to allow different minds to interpret the same information in different ways (Hsu and Cheung, 2013). Furthermore, the tendency to produce white lies is linked to emotional understanding abilities, (Demedardi et al., 2021), and moral judgments about white lies are linked to second-order false-belief understanding (Vendetti et al., 2019).

The Deceits task presents a story with a white lie, and asks why the speaker has used this utterance. The underlying intentions behind polite deceits are rarely explicitly explained in text. As a result, it is unlikely that LMs learn a direct connection between the utterance and the speaker’s intention during training on static texts. However, instances of polite deceits in text corpora may be accompanied by descriptions of characters’ emotional states, which may indicate that speakers’ intentions differ from what is literally conveyed by their utterance. This highlights the importance of context in interpreting deceits, which we return to in Section 4.5.3.

Indirect speech. Humans often use language in a performative sense, such as indirectly requesting an action from other individuals (e.g., Austin, 1975; Searle, 1975). Indirect or polite speech comprehension has been captured by Rational Speech Act (RSA; Frank and Goodman, 2012) models, which characterize listeners as performing Bayesian inference about a speaker who chooses utterances based on a tradeoff between epistemic and social utility (Brown and Levinson, 1987; Yoon et al., 2016, 2020; Lumer and Buschmeier, 2022).

The IndirectSpeech task presents a story with an indirect request, and asks what the speaker intends to convey. Like deceits, it’s unlikely that indirect speech acts are explained in text data. However, indirect requests may be followed by descriptions of the completion of

the implied request – for example, that someone closed a window after hearing the utterance “It’s cold in here”. Therefore, models may learn relationships between the utterances and desired outcomes through linguistic experience.

Irony. Humans use irony to convey the opposite of the semantic content of their utterance (Booth, 1974; Wilson and Sperber, 1992; Attardo, 2000; Wilson and Sperber, 2012). As such, irony has long been hypothesized to rely on social reasoning and perspective-taking (e.g., Happé, 1993; Andrés-Roqueta and Katsos, 2017). Indeed, human irony comprehension behaviors are captured by Bayesian reasoning models that take into account speakers’ affective goals (Kao and Goodman, 2014). In addition, neuroimaging studies suggest that irony interpretation relies on brain regions that are implicated in classic ToM tasks (Spotorno et al., 2012).

The Irony task presents a story with an ironic statement, and asks what the character intends to convey. While ironic statements are also rarely explained in text, models could leverage accompanying cues such as descriptions of characters’ emotional states or a mismatch in sentiment.

Maxims of conversation. Grice (1975) proposes that communication follows a set of *maxims*: be truthful; be relevant; be clear, brief, and orderly; and say as much as needed, and no more. A prevailing theory is that listeners derive implicatures by expecting speakers to be cooperative (i.e., abide by the maxims) and reasoning about speakers’ beliefs and goals. Indeed, there is extensive evidence for RSA models capturing these implicatures, such as those arising from the maxims of *quantity* (Potts et al., 2016; Frank et al., 2018; Degen, 2023) and *manner* (Bergen et al., 2016; Franke and Jäger, 2016; Tessler and Franke, 2018).

The Maxims task presents a story with a character flouting one of Grice’s maxims, and asks why the character has responded in such a way. Based on linguistic input, it may be easy for LMs to recognize when a speaker is flouting a maxim – for example, if an utterance is particularly long, features an uncommon syntactic construction, or diverges semantically from the context. However, it is unclear whether LMs will be able to recover the speaker’s underlying intentions.

Metaphor. Metaphors (Lakoff and Johnson, 1980) are used to draw comparisons between entities in a non-literal sense. Metaphor understanding has been hypothesized to require mentalizing (Happé, 1993), and fine-grained metaphor comprehension behaviors are captured by RSA models where listeners and speakers reason about each others’ beliefs and goals (Kao et al., 2014).

The Metaphor task presents a story with a metaphor, and asks what the speaker intends to convey. For models, the challenges of metaphor comprehension include accessing world knowledge and forming abstract relationships between domains. However, it’s possible that the relevant properties of the entities under comparison could emerge through linguistic experience.

Humor. Humor is one of the most distinctive aspects of human conversation, reflecting communicative goals with complex social function (Veatch, 1998; Martin and Ford, 2018). Neuroimaging studies suggest that joke understanding is supported by regions in the ToM brain network (Kline Struhl et al., 2018). Behavioral tests also reveal associations between ToM and humor abilities (Aykan and Nalçacı, 2018; Bischetti et al., 2019).

The Humor task presents a joke and asks which punchline makes the joke the funniest.⁴ Some theories argue that humor is triggered by linguistic incongruity effects (e.g., Deckers and Kizer, 1975), which might be straightforward for LMs to detect. Recent work has also shown that LMs can explain certain jokes (Chowdhery et al., 2022). However, some of Floyd et al.’s Humor items require complex world knowledge – for example, that slicing a pie into four versus eight pieces does not change the total amount of pie (see Table 4.1). As such, selecting the funniest punchline is a nontrivial task.

Coherence inferences. Humans also make pragmatic inferences beyond the sentence level – for example, by assuming that consecutive sentences form a logical or sequential relationship. Moss and Schunn (2015); Jacoby and Fedorenko (2020) find that constructing these discourse relationships loads on regions of the ToM brain network, suggesting a role of ToM in coherence inferences.

⁴Unlike the other tasks, there is no speaker utterance.

The Coherence task presents a pair of sentences, and asks whether the pair forms a coherent story.⁵ We assume that LMs’ training data, which consists of naturalistic text, is primarily coherent. Therefore, we expect LMs to be able to distinguish between coherent and incoherent sentence pairs (for an in-depth study, see Beyer et al., 2021).

4.4 Experiments

4.4.1 Evaluation paradigm

Our evaluation paradigm uses *zero-shot prompting*. Prompting can easily be adapted to all of our seven tasks, allowing us to compare performance across tasks within a model. Prompting also allows us to present models with inputs that are nearly identical to the stimuli seen by humans in Floyd et al.’s experiments, whereas other methods would require converting the stimuli into task-specific formats. We choose zero-shot prompts in order to evaluate the knowledge that emerges through training, and not through in-context adaptation to the task.

Prompt structure. Each prompt consisted of two parts: task instructions, and a query. The instructions were nearly identical to the instructions presented to humans in Floyd et al.’s experiments, prepended with the keyword “Task:”. The only other modification was that the original instructions had a final sentence of “Please answer as quickly as possible”, which we replaced with a sentence like “The answer options are 1, 2, 3, or 4”.⁶

For all tasks except Humor, the query consisted of the scenario (prepended with keyword “Scenario:”) and question, and then the numbered answer options (prepended with “Options:”).⁷ The prompt concluded with the keyword “Answer:”. Full example prompts are given in Appendix B.1.

Evaluation. To evaluate a model on a given item, we feed the prompt to the model, and measure the model’s probability distribution over tokens conditioned on the prompt. We

⁵This task differs from the others in that there is no speaker utterance, and the answer options are identical across items (“Coherent” or “Incoherent”).

⁶The exact answer options changed according to the task.

⁷For the Humor task, the joke was prepended with “Joke:”, and the answer options were prepended with “Punchlines:”.

Model	# parameters	Training
GPT-2	117M	Autoregressive LM
Tk-Instruct (3B)	3B	Multitask
Tk-Instruct (11B)	11B	Multitask
Flan-T5 (base)	250M	Multitask
Flan-T5 (XL)	3B	Multitask
InstructGPT-3 (ada)	350M (est.)	Multitask, human feedback
text-davinci-002	Unknown	FeedME

Table 4.2: Models tested in our experiments.

compare the probabilities of each answer token (e.g., “1”, “2”, “3”, or “4”) under this distribution. The model is considered correct on a given item if it assigns highest probability to the correct answer token, among all the possible answer tokens for that item.

We generated 5 versions of each item by randomizing the order of answer options. This was done to control for the base probabilities of the answer tokens. Since we do not analyze generated text, the model results themselves are deterministic.

4.4.2 Models

We test seven models across four model families, summarized in Table 4.2.⁸ As a baseline, we first test a base **GPT-2** model (117M parameters; Radford et al., 2019), which is trained on an autoregressive language modeling objective.

Second, we test a set of models which are based on T5 (Raffel et al., 2020) and instruction-finetuned on a diverse collection of tasks (Wei et al., 2022a). This set of models consists of two **Tk-Instruct** models (3B and 11B; Wang et al., 2022), which were fine-tuned on 1.6K tasks, and two **Flan-T5** models (base: 250M parameters; XL: 3B parameters; Chung et al., 2022), which were fine-tuned on 1.8K tasks. The fine-tuning tasks cover a wide range of categories, such as commonsense reasoning, translation, mathematics, and programming.

Finally, we test two **InstructGPT**-based models (Ouyang et al., 2022) via the OpenAI API: text-ada-001 (350M parameters), which we refer to as InstructGPT-3 (ada); and text-davinci-002, which comes from the GPT-3.5 family of models.^{9,10} These models are

⁸All non-OpenAI models were accessed via Huggingface (Wolf et al., 2020) and run on a single NVIDIA A100 GPU.

⁹Parameter estimates come from <https://blog.eleuther.ai/gpt3-model-sizes/>. Although the size of text-davinci-002 is unknown, we assume that it is larger than InstructGPT-3 (ada).

¹⁰The OpenAI models are not fully reproducible, but timestamps of model runs can be found in Appendix B.2.

fine-tuned to follow instructions and align with human feedback.

We compare models to a baseline from 374 humans, collected by Floyd et al. (In prep). Their experiments presented multiple choice questions to humans in nearly identical format to our prompts.

4.5 Results

We now return to the three questions posed in the Introduction, in each of the following subsections.

4.5.1 Do models choose the target pragmatic interpretation?

Figure 4-1 shows the proportion of trials where models and humans select the pragmatic answer. The smallest models (GPT-2, Flan-T5 (base), InstructGPT-3 (ada)) fail to perform above chance. The largest models (T_k -Instruct (11B), Flan-T5 (XL), text-davinci-002) perform above chance on all tasks (except T_k -Instruct (11B) on Maxims), and in some cases near human-level. Overall, models perform worst at the Humor, Irony, and Maxims tasks. Interestingly, these phenomena involve speakers violating listeners' expectations in some way: producing a funny punchline to a mundane story (Humor), stating the direct opposite of the speaker's belief (Irony), or disobeying one of the assumed rules of conversation (Maxims). It may be that models fail to represent certain social expectations that are maintained by human listeners.

Next, we investigated the relationship between model size and accuracy. Figure 4-2 shows the mean accuracy achieved by each model (averaged across tasks) vs. millions of parameters.¹¹ The line and error bars denote the mean and 95% CIs, while points represent individual models. We find a coarse effect of model size: there is a stark jump in accuracy after 1B parameters (dashed line). However, model size does not fully explain variance in accuracy: all models with <1B parameters achieve similar accuracy, and Flan-T5 (XL) outperforms T_k -Instruct (3B), despite both having 3B parameters.

¹¹text-davinci-002 was excluded from this analysis, as the number of parameters is unknown.

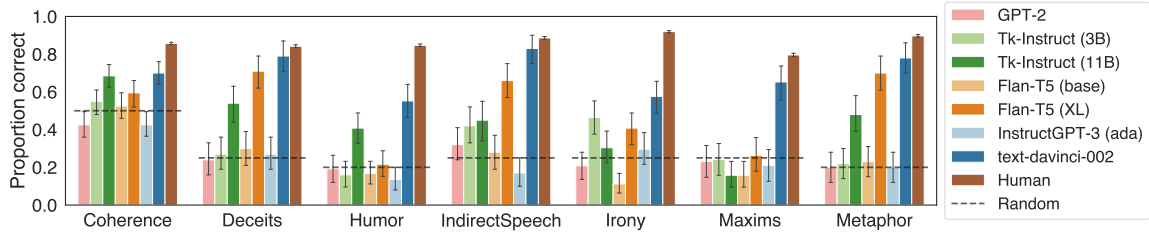


Figure 4-1: Accuracy for each task. Error bars denote 95% CI. Dashed line indicates task-specific random baseline.

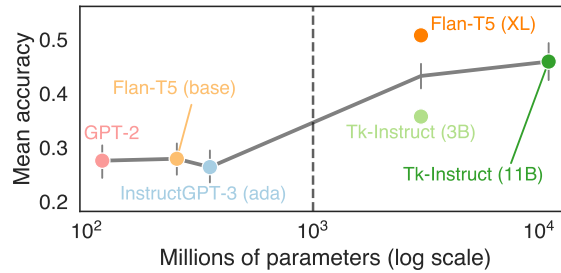


Figure 4-2: Mean accuracy vs. millions of parameters. Vertical dashed line indicates 1 billion parameters.

4.5.2 Do models and humans make similar types of errors?

Recall from Section 4.3 that each item has a set of answer options that correspond to different strategies (Table 4.1).¹² In addition to the target pragmatic answer (Correct), each item also has a plausible but unlikely literal answer (Literal), as well as distractors based on lexical overlap or semantic associations (Distractor*). For each item, we computed the human empirical distribution over answer choices, and compared it to models’ probability assigned to the answer tokens (e.g., “1”, “2”, “3”, and “4”).

Figure 4-3 shows the answer distributions across models and humans for each task. Across tasks, humans primarily select the Correct option, occasionally select the Literal option, and rarely select the distractors. We find a similar pattern for text-davinci-002, although the model is more likely to select the Literal option in general. The other large models (Tk-Instruct (11B) and Flan-T5 (XL)) also generally assign highest probability to the Correct and Literal options, although the distribution looks less human-like. The next-largest models (Tk-Instruct (3B) and Flan-T5 (base)) prefer the Literal option, and the remaining

¹²The exception is Coherence, which is excluded here.

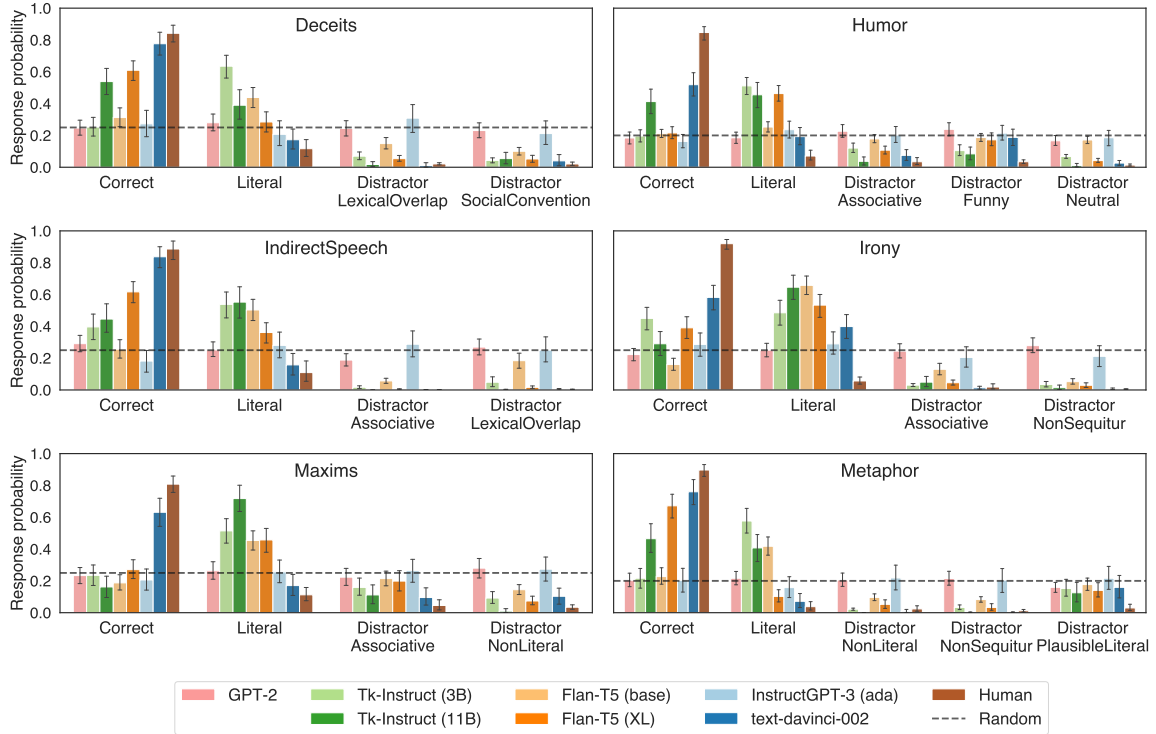


Figure 4-3: Response distributions across models and humans. Answer options for each task are shown on the x-axis. For models, y-axis denotes probability assigned to each answer option. For humans, y-axis denotes empirical frequency of each answer option being selected. Error bars denote 95% CI. Dashed line indicates random baseline.

models (GPT-2 and InstructGPT-3 (ada)) are at chance. These results suggest that, for our test items, larger models satisfy the basic language processing component of pragmatic comprehension: when these models fail, they strongly prefer the literal interpretation over semantically and lexically related distractors.

However, even highly performing models occasionally do select the distractor answers, revealing interesting behaviors. For example, in the Metaphor task, text-davinci-002 and Flan-T5 (XL) prefer the DistractorPlausibleLiteral option – which is a figurative reading of the utterance – over the Literal option – which is completely non-figurative. Similarly, in the Humor task, text-davinci-002 is much more likely to select the DistractorFunny option over the other (non-humorous) distractors. This suggests a coarse sensitivity to humor, even if the model selects the human-preferred punchline only 55% of the time (see Figure 4-1). We take this analysis to illustrate the value of looking beyond binary pragmatic/non-pragmatic response distinctions, and using controlled distractor items to evaluate models’ abilities

(e.g., McCoy et al., 2019).

4.5.3 Are models and humans sensitive to similar linguistic cues?

Having found qualitatively similar response patterns between humans and models, we now ask *how* models and humans arrive at pragmatic interpretations, and whether they use similar types of information. We begin with a broad evaluation of the extent to which models and humans rely on linguistic context (Section 4.5.3). Finally, we take a more granular approach and ask whether model and human performance is correlated at the item level – i.e., if models and humans exhibit similar sensitivity to the cues that make a non-literal interpretation more or less likely (Section 4.5.3).

The role of context

Many cues for enriched language understanding come from the context in which the speaker makes their utterance. However, some aspects of non-literal comprehension might arise given the utterance in isolation, while others are highly sensitive to specific contextual details (e.g., Levinson, 2000). Therefore, we expect that the degree to which humans rely on context to select non-literal interpretations will vary across the tested tasks.

To investigate this variation, we created a new set of stimuli by removing the context stories, leaving only the speaker utterance and final question (e.g., *Dan says, “The dog knocked it over.” Why has Dan responded in such a way?*).¹³ We re-ran the human experiment on 30 participants, following the protocols of Floyd et al. (In prep)’s original experiment using the no-context modified materials.¹⁴ We also re-ran the three models that achieved highest accuracy on the original items: *Tk*-Instruct (11B), Flan-T5 (XL), and text-davinci-002.

Figure 4-4 shows the mean accuracy difference on the original versus no-context versions of each item.¹⁵ We find that models and humans exhibit a similar qualitative pattern:

¹³This manipulation is not compatible with the Humor and Coherence tasks, so they are excluded from this analysis.

¹⁴Details can be found in Appendix B.3.1.

¹⁵See Figure B-1 in Appendix B.3.2 for comparison of raw accuracy scores on the original and no-context items.

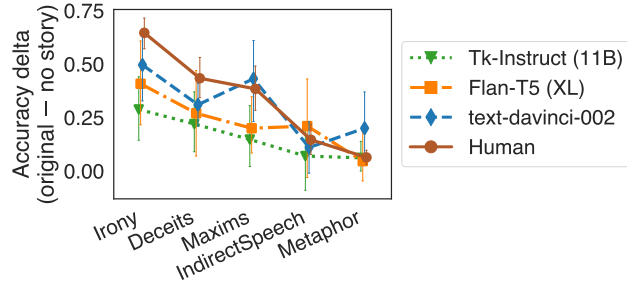


Figure 4-4: Mean by-item difference in accuracy once story context was removed.

removing the story leads to the largest degradation for Irony, followed by Deceits and Maxims. This aligns with our intuitions, because in these cases, speakers’ utterances can be interpreted either literally or as the complete opposite, based on the specific social situation (e.g., “It is so pleasant here”). In contrast, there are smaller degradations for IndirectSpeech and Metaphor. This suggests that some indirect requests are conventionalized (e.g., “I am getting cold”), although their interpretations may be facilitated by context (e.g., Gibbs, 1979). Similarly, this suggests that metaphor interpretation may draw more upon global knowledge than local context.

Scrambling. Next, we tested whether models rely on syntactic and discourse-level information from the context, or whether they can perform the tasks when ordering cues are removed. We constructed two scrambled versions of each item by randomizing the order of sentences and words. In both versions, the instructions, final question (e.g., *Why has Dan responded in such a way?*), and answer options were unmodified and remained in their original positions. Again, we only tested the best-performing models on these items.

We found that the models maintain reasonable performance for most tasks, with the notable exception of Metaphor (Figure B-2; Appendix B.4). This accords with prior work showing that models often rely on lexical information without human-like compositionality (e.g., Dasgupta et al., 2018; Nie et al., 2019; McCoy et al., 2019). We expect that scrambling, especially at the word-level, would likely disrupt human performance, but this remains an open empirical question. We leave an in-depth investigation of human performance to future work.

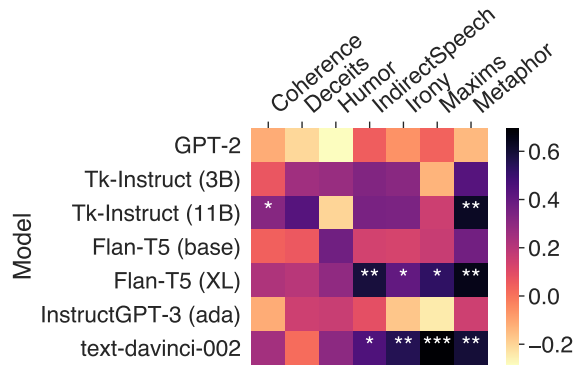


Figure 4-5: Pearson correlation coefficients between by-item human accuracy and model probability of the correct answer. Cells are marked with significance codes.

Item-level alignment

Up to this point, we analyzed differences across phenomena by averaging over items. However, there is also variance *within* each phenomenon in the types of cues that suggest how the utterances should be interpreted. For example, some items contain explicit descriptions of characters’ emotional states (e.g., “Sarah becomes angry”). If models and humans leverage these cues in similar ways, then we would expect to see correlations between model and human performance at the item level.

For each task and model, we compute the Pearson correlation between by-item mean accuracy achieved by humans and by-item mean probability that models assigned to the correct answer (Figure 4-5). In general, the larger models (Tk-Instruct (11B), Flan-T5 (XL), text-davinci-002) are better aligned with humans, and the strongest correlations occur for IndirectSpeech, Irony, Maxims, and Metaphor. This suggests that for those tasks, models and humans are similarly sensitive to cues that make a non-literal interpretation likely.

4.6 Discussion

We used an expert-curated set of materials (Floyd et al., In prep) to compare LMs and humans on seven pragmatic phenomena. We found that Flan-T5 (XL) and text-davinci-002 achieve high accuracy and match human error patterns: within incorrect responses, these models tend to select the literal interpretation of an utterance over heuristic-based distractors.

We also found preliminary evidence that LMs and humans are sensitive to similar linguistic cues: model and human accuracy scores correlate at the item-level for several tasks, and degrade in similar ways when context is removed.

Our results suggest that language models can consistently select the pragmatic interpretation of a speaker’s utterance – but how? The models tested in our experiments reflect a variety of learning processes through which pragmatic knowledge could emerge. GPT-2 is trained to learn the distribution of linguistic forms; the *Tk*-Instruct and Flan-T5 models are pre-trained on a denoising task and fine-tuned on thousands of instruction-based tasks; and the OpenAI models receive signal from human feedback. Our experiments are not designed to tease apart the contributions of these training procedures to models’ behaviors. Therefore, we do not intend to make strong claims about the mechanisms by which models learn pragmatics.

A shared feature of our tested models is the lack of explicitly constructed mental state representations. In this sense, our results are potentially compatible with two hypotheses. One possibility is that the models do not have an ability that can be considered an analog of Theory of Mind (ToM). This view is supported by evidence that language models perform poorly on social commonsense and false-belief tasks (Sap et al., 2022), and are remarkably brittle to small perturbations of classic tests (Ullman, 2023). If models truly lack ToM, then their pragmatic behaviors might be explained by inferences based on low-level linguistic cues. Taken a step further, this could potentially suggest that certain human pragmatic behaviors arise through inferences based on language statistics, with no need for mental state representations.

A second possibility is that models do have a heuristic version of ToM, which is not explicitly engineered but instead emerges as a by-product of optimizing for other objectives (such as linguistic prediction). Since language contains many descriptions of agents’ beliefs, emotions, and desires, it may be beneficial – perhaps even necessary – to induce representations of these mental states in order to learn a generative model of linguistic forms. Indeed, Andreas (2022) argues that whereas language models have no explicit representation of communicative intents, they can infer approximate representations of the mental states of the agents that produce a given linguistic context. If this hypothesis

is true, however, it would still remain unclear whether ToM is *necessary* to support the pragmatic behaviors tested in our evaluation materials.

Our experiments do not differentiate between these two hypotheses. However, fine-grained behavioral evaluations – such as those presented in this work – are important for revealing models’ capabilities and weaknesses, and offer a first step toward understanding how pragmatic behaviors could emerge. A promising direction for future work is to test models with a wider range of training objectives, or even new architectures, such as distinct language and reasoning modules (see Mahowald et al., 2023). In addition, while there is evidence for the role of mentalizing in our tested pragmatic phenomena (see Section 4.3.1), one limitation of our stimuli is that they were not specifically designed to require ToM. New datasets that perform targeted manipulations of ToM alongside tests of language comprehension could help shed light on how linguistic experience and ToM jointly support pragmatic behaviors.

Chapter 5

Conclusion

This thesis presents three case studies using artificial neural networks (ANNs) to investigate questions about language learning and comprehension. In particular, we use ANNs to test the idea that experience with linguistic forms – and the induced probabilistic expectations – can support the emergence of complex linguistic abilities. Our findings demonstrate that ANNs capture many human language behaviors, which adds plausibility to the idea that certain aspects of linguistic competence could, in principle, emerge through domain-general probabilistic learning. However, our experiments also reveal areas where optimizing for word-prediction might not be enough: for example, hierarchical inductive biases appear to support human-like syntactic generalizations in small-data settings (Chapter 2), and scalar inference rates are better captured by treating unspoken alternatives as concepts instead of raw string forms (Chapter 3).

Importantly, these studies do not provide direct evidence that humans are acquiring or processing language in any particular way. As discussed in Section 1.2.3, models and humans differ drastically in many respects, such as the quantity, genre, and modality of their training data. As such, positive evidence that a model possess a particular linguistic ability does not license the conclusion that models and humans achieve that outcome in similar ways. Nonetheless, the question posed above – whether experience with linguistic forms can support the emergence of complex linguistic abilities – remains relevant from a cognitive perspective. For example, it is not obvious whether experience with linguistic forms enables a learner to formulate distributional categories over abstract categories of

words, or whether speaker intentions can be inferred at all without built-in representations of communicative goals. Furthermore, if it is the case that human linguistic abilities do emerge through statistical learning and prediction, this would support the broader views that humans are tuned to their environments (Anderson, 1990; Anderson and Schooler, 1991; Tomasello, 2003; Bybee and Beckner, 2015) and that prediction serves as a core motif in the mind and brain (Bar, 2009; Bubic et al., 2010; Den Ouden et al., 2012). Orthogonally, ANNs enable us to systematically investigate these questions in a way that would be infeasible in humans. In this sense, even if our results do not provide a mechanistic explanation of human linguistic abilities, the type of evidence offered by our experiments – for example, favoring a certain inductive bias over another (Chapter 2) – marks a methodological contribution in approaching long-standing questions about language learning and understanding.

In the remainder of this chapter, I discuss broader implications of ANN language models for understanding the human mind (Section 5.1), as well as for developing artificial intelligence systems with language abilities (Section 5.2).

5.1 Implications for cognitive science

LMs as scientific theories. A growing movement proposes that ANN language models (LMs) should be treated as implemented theories of language acquisition and processing (e.g., Baroni, 2022; Wilcox et al., 2022a; Contreras Kallens et al., 2023; Piantadosi, 2023). Many of these studies focus on the implications of LMs for nativism, or arguments that human linguistic knowledge relies on innate, language-specific learning mechanisms (e.g., Chomsky, 1965). At face value, LMs appear to challenge the predictions made by this theoretical approach: they use language with remarkable fluency and sophistication, while having very different primitives from those postulated in generative linguistics. For example, LMs have continuous and gradient representations, typically lack built-in language-specific machinery, and optimize for objectives related to distribution learning (e.g., next-word prediction, masked language modeling, or infilling, occasionally with additional supervision from human preferences). Indeed, Piantadosi (2023) goes so far as to claim that LMs “refute” Chomsky’s foundational assumptions about the way language is acquired and

represented. The mechanisms of LMs, Piantadosi argues, reflect a deep integration of syntax and semantics, as well as a central role for probabilistic prediction, in contrast to prior proposals about the autonomy of syntax and irrelevance of probabilities (e.g., Chomsky, 1957, 1968; Adger, 2018).

Piantadosi’s argument (and similar perspectives) has been criticized from several angles, such as its conflation of likelihood and grammaticality (Katzir, 2023), its misinterpretation of Chomsky’s key claims (Milway, 2023), and even its logical basis (Rawski and Baumont, 2023; Milway, 2023). Perhaps the most prominent critique is that LMs are fundamentally limited in their relevance to scientific theory, as they do not explain key traits of the human language faculty. For example, Milway (2023) and Chomsky¹ argue that LMs do not differentiate between natural languages and “impossible languages”, or languages with grammatical features that would not be easily acquired by humans. Similarly, LMs do not explain linguistic “universals”, such as the regularity of phonological processes and hierarchical dependencies in syntax (Katzir, 2023). In this sense, LMs fall short as linguistic theories, as they do not predict which features are (un)attested in human languages, let alone *why* these features are (dis)preferred.

Another criticism is that LMs conflate competence and performance, which linguists have argued to be distinct components of human language (Yngve, 1960; Chomsky, 1965; Dupre, 2021; Katzir, 2023). For example, Katzir (2023) claims that “[LMs’] behavior directly reflects their competence, and when they fail it is their competence that is at fault”; furthermore, he argues, “it is never the case” that an LM’s competence supports the correct underlying generalization, while its behavior reflects a “failure” due to performance constraints. However, many researchers have highlighted a functional distinction between performance and competence in LMs (e.g., Miracchi, 2019; Firestone, 2020). For both models and humans, any method of evaluating linguistic knowledge will pose its own set of task demands, which will necessarily influence the behaviors measured under that paradigm. Indeed, in many cases, LMs’ performance failures can be ameliorated with evaluation methods that better “motivate” the model or expand its contextual resources, such as few-shot or chain-of-thought prompting (e.g., Wei et al., 2022c; Lampinen, 2023;

¹https://www.youtube.com/watch?v=PBdZi_JtV4c; see pages 2-3 of Milway (2023) for details.

Moghaddam and Honey, 2023). In this sense, contrary to Katzir's claim, a model could represent the correct generalization, while producing behaviors that depend on task demands and environmental pressures.

Clearly, it remains debated which aspects of generative linguistics are challenged by LMs. It may be fruitful to consider hybrid perspectives, combining algebraic linguistic formalisms with data-driven, bottom-up statistical learning. For example, behavioral tests (such as those in Chapter 2) have highlighted the promise of neuro-symbolic language models, such as recurrent neural network grammars (Dyer et al., 2016) and Transformer grammars (Sartran et al., 2022). Indeed, while one potential conclusion from Piantadosi's argument is that LMs support connectionist views of language (e.g., Rumelhart and McClelland, 1986; Rumelhart et al., 1987), the empirical landscape is compatible with alternate views: for example, that structural inductive biases are necessary early during learning, but certain syntactic computations become amortized and can later be gleaned from language statistics. Another consideration is that it may be misguided to treat LMs as blank-slate learners, as is done in many studies to provide a theoretical counterpoint to nativist approaches (see Baroni, 2022, for examples). Instead, Baroni (2022) argues that LMs embody specific inductive biases, which should be seen as testable scientific theories. On this view, a particular model architecture (e.g., LSTM or Transformer) can be seen as defining a space of possible grammars, and a model trained on data from a particular language can be seen as a system that predicts whether an input utterance is acceptable in that language.

Regardless of one's theoretical position, the surprising capabilities of LMs warrant at least a re-examination of traditional assumptions regarding the inductive biases, cognitive architectures, and learning processes that support human language. Furthermore, LMs allow us to generate and test predictions through systematic, large-scale simulations given precise architectural and environmental assumptions. In this sense, the type of evidence contributed by LMs marks a qualitative shift in the way we can test theories of language acquisition and processing. It appears promising to continue probing LMs as scientific theories, combining psycholinguistic methods and linguistic theory to advance our understanding of human language.

LMs as quantity estimation tools. While many recent studies have focused on LMs and language acquisition, another way that LMs can inform linguistic theory is by estimating quantities that enable theory-testing at new scales. One of the most successful examples of this approach is the theory of expectation-based language comprehension. A body of empirical work has demonstrated a tight link between human reading times and LM-derived next-word probabilities (e.g., Levy, 2008; Smith and Levy, 2013; Shain et al., 2022). In addition, Chapter 3 of this thesis shows that context-sensitive expectations also explain variation in pragmatic inferences. Together, these findings provide large-scale evidence for predictive mechanisms during language comprehension – a type of evidence that, crucially, relies on models that can estimate string probabilities in arbitrary contexts with fine granularity.

A relatively underexplored way of using LMs to operationalize theories at naturalistic scales is in the domain of pragmatics. Many researchers have proposed computational accounts of pragmatics that formalize Grice’s (1975) idea of rational communication using game theory and Bayesian inference. These proposals could have profound implications for describing language with broader cognitive principles, framing communication as a rational, inferential process. However, the proposed modeling frameworks are often difficult to test beyond toy domains. One challenge is that these models are highly sensitive to quantities such as costs and prior likelihoods of world states, which are typically specified by hand. Recent work has leveraged ANNs to address this challenge – for example, by estimating utterance costs in pragmatic speaker models through statistical learning (Nie et al., 2020a).

Another major challenge for many pragmatics models is specifying the literal meanings of linguistic expressions, or a function that describes how semantically compatible a certain expression is with a certain meaning (the “lexicon” function). Prior work has addressed this challenge by training ANNs to parameterize this function (e.g., Andreas and Klein, 2016; Monroe et al., 2017), providing new support for pragmatic reasoning in scaled up, non-toy domains. This demonstrates how ANNs can enable pragmatic-theory-testing in naturalistic settings, but unfortunately requires large amounts of labeled data and new models for each new task. A promising future direction is to derive lexicon functions from large language models (LLMs), which can output vast semantic knowledge without the need for

additional fine-tuning datasets or gradient updates. For example, LLMs can be prompted to provide semantics in color and spatial domains through few-shot in-context learning (Patel and Pavlick, 2022). Using unstructured data-driven knowledge from LMs as the semantic inputs to a Bayesian reasoning process (such as the Rational Speech Act model; Frank and Goodman, 2012; Goodman and Frank, 2016) could allow us to test the hypothesis that humans engage in a hybrid reasoning process, where some information is implicit in data-driven statistical knowledge, and other information is derived through probabilistic symbolic reasoning. This also connects to prior work using LMs to provide primitives in a probabilistic language of thought (Goodman et al., 2015) to solve structured reasoning problems (Lew et al., 2020).

There are many promising ways that ANNs could inform theories about human language and cognition. With that said, there are still many open questions about how existing models work the way they do. For example, *why* are models trained on next-word prediction so effective at capturing human language comprehension behaviors? One approach (known as “probing”) has investigated this question by analyzing the information encoded in model representations using simple readout functions. Many studies have demonstrated that LMs represent syntactic parse trees (Hewitt and Manning, 2019; Manning et al., 2020) and incremental syntactic parse states (Eisape et al., 2022), suggesting that incremental syntactic inferences may underlie next-word prediction. But to what extent do these representations of symbolic structures play a causal role in model dynamics, versus simply reflecting a property of the probing functions? Another open puzzle is that LMs capture human behavioral trends in targeted syntactic assessments, and yet they consistently underpredict the magnitude of human processing difficulties (Wilcox et al., 2021; van Schijndel and Linzen, 2021). This tension – predicting qualitative processing difficulties, while systematically misestimating the magnitude of syntactic violations – reveals a misalignment between the way humans and ANNs process incremental linguistic content. If models are misaligned with humans in this way, then how do they produce language that sounds so fluent and human-like? These questions merely scratch the surface of a trove of mysteries about the successes and limitations of modern LMs. Reaching a deeper understanding of these models – through a combination of behavioral and representational analyses – can help us gain even sharper

	Benchmarking	Targeted evaluation
Design	What are the downstream tasks that we want our model to succeed at?	What are the behavioral or representational signatures of the knowledge that we want our model to learn?
Methods	When we probe/prompt the model, what type of information falls out?	Which analysis method is best suited to measure the information of interest?
Interpretation of results	Does the model improve upon state-of-the-art task performance?	Which hypotheses are supported by the model results?

Table 5.1: Contrast between traditional NLP benchmarking and targeted evaluation.

insights into the mechanisms of the mind.

5.2 Implications for natural language processing

In this thesis, the primary role of ANNs is to investigate theoretical questions about human language learning and understanding. However, the studies presented in Chapters 2 to 4 also contribute new tests for evaluating human-like linguistic knowledge in ANNs. As one of the goals of natural language processing (NLP) is to build models that use language like humans do, a critical enterprise for NLP research is to develop evaluation paradigms that measure a model’s ability to capture key features of human language.

Currently, the dominant paradigm for NLP model evaluation is *benchmarking*. Benchmarks use large collections of items to measure how well a model performs on a certain task (e.g., Williams et al., 2018; Wang et al., 2019b,a; Nie et al., 2020b; Srivastava et al., 2022). Typically, models are pre-trained on a general objective (such as language modeling), and then fine-tuned and evaluated on the downstream task of interest. Many popular benchmarks have garnered criticism for being solvable based on spurious heuristics (e.g., McCoy et al., 2019), being quickly saturated by modern models (Kiela et al., 2021), over-rewarding low-bias architectures (Linzen, 2020), and harboring biases and poor design (Bowman and Dahl, 2021). This has inspired new approaches such as adversarially constructed items (Kiela et al., 2021) and massive collections of tasks (Srivastava et al., 2022), but the general paradigm remains fundamentally the same.

This thesis demonstrates a complementary approach: using controlled assessments

to measure whether models exhibit behavioral signatures of human linguistic knowledge. Table 5.1 highlights the primary contrasts between benchmarking and targeted evaluation, which I discuss in more detail in the remainder of the section.

Task design. As NLP technologies are widely deployed in consumer-facing products (e.g., Dale, 2019; Le Glaz et al., 2021), there is broad interest in evaluating NLP models on tasks that are “downstream” (i.e., closer to an end-point application) from a general training objective like language modeling. This task-driven approach also influences benchmarks that are meant to measure more abstract linguistic abilities. For example, the *GLUE family of benchmarks – which are intended to measure “general-purpose language understanding” (Wang et al., 2019b,a) – consist of tasks like question answering, natural language inference, and paraphrasing.

Instead of asking which tasks a model should succeed on, targeted evaluation asks: what are the behavioral or representational signatures of the knowledge we’re interested in evaluating? For example, suppose we are interested in evaluating whether a model captures generalizations about English syntax. One approach would be to define a benchmark based on an acceptability classification task, training classifier models on top of sentence encoders to output the probability of a sentence being grammatical or ungrammatical (e.g., Warstadt et al., 2019). In contrast, a targeted evaluation approach would identify the signatures of the knowledge that the model should learn: a model that has captured the relevant generalizations should assign higher probability to grammatical strings than minimally different ungrammatical counterparts (see Chapter 2; also Warstadt et al., 2020a). This approach measures the knowledge implicit in the model, without direct experience with the evaluation setup.

Another important aspect of task design is specifying how model outputs should be compared to human behavior. Many benchmarks compare models’ task performance to a human-level baseline that averages performance over a large number of human crowdworkers. However, in many cases, there is systematic *variation* in human behaviors that may complicate such categorical generalizations (Pavlick and Kwiatkowski, 2019; Jiang and de Marneffe, 2022; Baan et al., 2022; Webson et al., 2023), or should be *directly captured*

by the model (see Chapter 3).

Methods & linking functions. Like humans and other intelligent systems, modern NLP models produce sophisticated patterns of behavior, but their underlying mechanisms cannot be directly observed. The past decades have produced a variety of methods for analyzing the knowledge and abilities of NLP models (for review, see Belinkov and Glass, 2019). Many NLP studies select an analysis method – such as prompting, behavioral testing, or probing – and apply it to a phenomenon of interest, without explicitly stating (1) the type of knowledge that, in principle, can be revealed by that method; and (2) assumptions that the method makes about the relationship between behavior and representations. This issue of linking functions is central in cognitive science, where a major research goal is to link observable behaviors to underlying mental models (e.g., Jasbi et al., 2019) or neurobiological representations (e.g., Ivanova et al., 2022). However, few NLP benchmarks explicitly describe the hypothesized link between model outputs and underlying representations. In contrast, targeted evaluation should clearly specify the assumptions made by the analysis method, and choose a method that is best suited to measure the behavioral or representational phenomenon of interest.

As an example, suppose we are interested in whether LMs exhibit human-like linguistic predictions. One could design a benchmark that queries models’ internal expectations through prompts. Figure 5-1 shows an example interaction with ChatGPT (circa December 2022) with the following prompt: “what are probable candidates for the next word of the sentence ‘the dog barked because?’” While the model clearly has the ability to produce probable candidates for the next word of the sentence – that is precisely the task it has been pre-trained on – it responds to the prompt by stating that it is “not able to provide a list of ‘probable’ candidates”.² In contrast, evaluating the model’s output logits after consuming the sentence prefix would allow us to directly measure probable candidates for continuations under the distribution learned by the model. Clearly, this second approach is a more robust way to measure LMs’ expectations than the prompting method. But it still isn’t as simple as it seems. If we want to compare model-derived probabilities to human linguistic predictions, then we would need to specify a *link* between probabilities and human

²Source: <https://twitter.com/yoavgo/status/1598360581496459265>

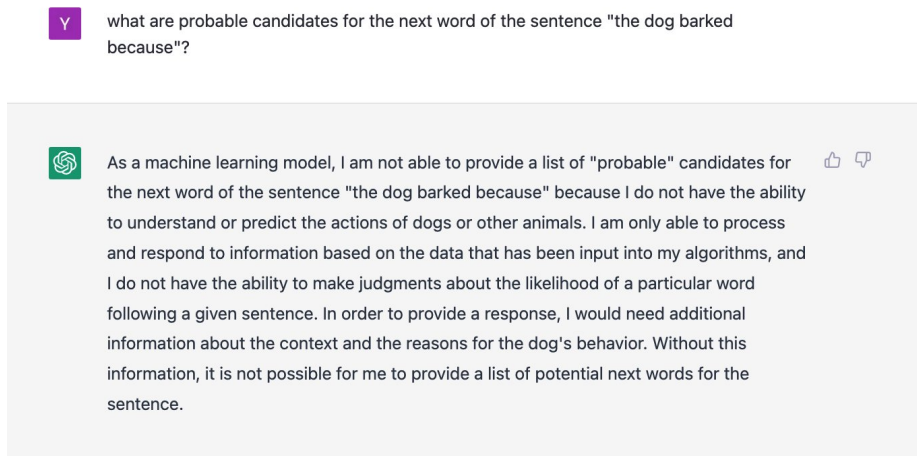


Figure 5-1: Example interaction with ChatGPT, asking the model about probable candidates for the continuation of the sentence “the dog barked because”. Source: <https://twitter.com/yoavgo/status/1598360581496459265>

behaviors. For example, if we want to compare model-derived next-word probabilities with human word-by-word reading times, then we need to justify the assumed functional relationship between these quantities. If we want to compare model sentence-completions with human sentence-completions (e.g., in a Cloze task), then we need to justify the assumed decoding method that produces samples from the underlying distribution. More broadly, NLP researchers should motivate the choice of analysis method, as well as any assumed links between variables of interest, in order to frame the interpretation of the empirical findings.

Interpretation & statistical analysis. The goal of the targeted evaluation paradigm is to use model data to adjudicate between competing hypotheses, whereas benchmarking is typically used to test whether a model improves upon the state-of-the-art on a certain task. This distinction has connections to a broader theme in the changing NLP landscape: the shift from engineering to science. Now that NLP researchers and practitioners have created models that are capable of incredible feats and tasks, there has been an outpouring of attempts to reverse-engineer the inner workings of these models, spawning new subfields like “BERT-ology” (e.g., Rogers et al., 2021). These changes are making the research goals of NLP – to understand the mechanisms of a complex system by observing its

behaviors and representations – increasingly aligned with the research goals of (cognitive) science. However, the *methods* of NLP – especially with respect to experiment design and interpretation of results – haven’t shifted to align with the methods of science in the same way. Many proposals have advocated for using the tools of natural and psychological sciences to improve the scientific rigor of NLP research. For example, van Miltenburg et al. (2021) argue for using preregistration in NLP research, which allows researchers to distinguish between exploratory hypothesis generation and confirmatory hypothesis testing (Nosek et al., 2018). Others have advocated for designing evaluations with sufficient statistical power (Card et al., 2020) and performing significance testing for appropriate interpretation of results (Yeh, 2000; Koehn, 2004; Riezler and Maxwell, 2005; Berg-Kirkpatrick et al., 2012; Søgaard et al., 2014; Dror et al., 2018; Sadeqi Azer et al., 2020).

Returning to the broader picture, targeted evaluation is not a replacement for benchmarking – instead, it should be a complementary tool. There is certainly value in having massive amounts of test items (e.g., to control for lexical confounds), or being able to easily compare models using a single metric for specific downstream applications. Indeed, the points highlighted above – carefully designing tasks, methods, and statistical analyses that measure the signatures of human-like linguistic knowledge – can also be integrated into traditional benchmarking paradigms. Nevertheless, recent benchmarks seem to primarily follow the trend of getting larger (e.g., Srivastava et al., 2022), which is orthogonal to the suggestions described in this section. Continuing to develop targeted evaluations, grounded in what we know about the human language faculty, is a promising direction for building artificial models with human-like linguistic abilities.

Appendix A

Supplementary material for Chapter 2

A.1 Description of test suites

In this work we have assembled a large number of test suites inspired by the methodology of experimental sentence-processing and psycholinguistic research. Each test suite contains a number of ITEMS, and each item appears in several CONDITIONS: across conditions, a given item will differ only according to a controlled manipulation designed to target a particular feature of grammatical knowledge. For each suite we define a SUCCESS CRITERION, which stipulates inequalities among conditional probabilities of sentence substrings.

In the main paper, a model’s accuracy for a test suite is computed as the percentage of the test suite’s items for which it satisfies the criterion. In this appendix, we briefly describe each test suite and the criterion used to determine whether a given model succeeds on each item of the test suite.

A.1.1 Notation

Sentence status

Following and building on linguistic traditions, we annotate examples as follows. Examples marked with a * violate a well-established grammatical constraint, and are ungrammatical. Examples marked with ? or ?? are not necessarily ungrammatical, but are marginal: for example, they may require an unusual interpretation of a word in order for the sentence

to be grammatical. (More ?'s is roughly intended to indicate more severe marginality). Examples marked with ! are not ungrammatical, but induce severe processing difficulty that is measurable in real-time human sentence processing. For all test suites, we include references to established literature on the relevant grammatical and/or sentence-processing phenomena.

Success criteria

Criteria involve inequalities among conditional probabilities of sentence substrings given the complete sentence context preceding the substring. In describing criteria, we use $P(\cdot)$ for raw probabilities and $S(\cdot)$ for surprisals (negative log-probabilities), and leave the conditioning on preceding context implicit. For concision, we use subscripts on P and S to indicate the variant of the sentence within the test suite that we are referring to. In the first described test suite, CENTER EMBEDDING (Appendix A.1.2), we show the criterion in both concise and fully spelled-out forms, to help clarify the conventions we are using in the concise form. All items within a given test suite share the same criterion for success.

We provide chance accuracy on the assumption that the order of probabilities among conditions for a given item is random. In some cases, exactly determining chance accuracy may require further assumptions about the distribution of these probabilities; in this case we provide an upper bound on chance accuracy.

A.1.2 Center embedding

Center embedding, the ability to embed a phrase in the middle of another phrase of the same type, is a hallmark feature of natural language syntax. Center-embedding creates NESTED SYNTACTIC DEPENDENCIES, which could pose a challenge for some language models. To succeed in generating expectations about how sentences will continue in the context of multiple center embedding, a model must maintain a representation not only of what words appear in the preceding context but also of the order of those words, and must predict that upcoming words occur in the appropriate order. In this test suite we use verb transitivity and subject-verb plausibility to test model capabilities in this respect. For example, A below is a

correct center-embedding, but B is not:

- (A) The painting_{N₁} that the artist_{N₂} painted_{V₂} deteriorated_{V₁}. [correct]
- (B) ??The painting_{N₁} that the artist_{N₂} deteriorated_{V₁} painted_{V₂}. [incorrect]

Here, N_i and V_i correspond to matched subject–verb pairs.

In the WITH-MODIFIER version of the test suite, we postmodify N_2 with a relative clause to increase the linear distance over which the nested dependencies must be tracked, potentially leading to a harder test suite:

- (A) The painting_{N₁} that the artist_{N₂} who lived long ago painted_{V₂} deteriorated_{V₁}. [correct]
- (B) #The painting_{N₁} that the artist_{N₂} who lived long ago deteriorated_{V₁} painted_{V₂}. [incorrect]

Criterion The probability of the verb sequence in the correct variant should be higher than the probability of the verb sequence in the incorrect variant:

$$P_A(V_2V_1) > P_B(V_1V_2)$$

In full form, this criterion for the example item in the no-modifier version of this test suite would be:

$$P(\text{painted deteriorated}|\text{The painting that the artist}) > \\ P(\text{deteriorated painted}|\text{The painting that the artist})$$

Chance performance on these center-embedding test suites would be 50%.

References Miller and Chomsky (1963); Wilcox et al. (2019a)

A.1.3 Pseudo-clefting

The pseudo-cleft construction involves (i) an extraction of a TARGETED CONSTITUENT from a sentence and (ii) a constituent that provides the semantic contents of the targeted

constituent and must match it in syntactic category, where (i) and (ii) are linked by the copula. The pseudo-cleft construction can target both NPs and VPs; in the latter case, the VP of the free relative becomes an inflected form of *do*. This means that a free relative subject plus the copula can set up a requirement for the syntactic category that comes next. If the free relative clause has a *do* VP without a direct object, then the main-clause postcopular predicate can be a VP (A below). Otherwise, the postcopular predicate must be an NP (C below):

- (A) What the worker did was $\overbrace{\text{board the plane}}^{\text{VP}}$.
- (B) ?What the worker did was $\overbrace{\text{the plane}}^{\text{NP}}$.
- (C) What the worker repaired was $\overbrace{\text{the plane}}^{\text{NP}}$.
- (D) *What the worker repaired was $\overbrace{\text{board the plane}}^{\text{VP}}$.

Criterion The postcopular predicate should be more surprising when its syntactic category mismatches the cleft, averaging across VP and NP postcopular predicates:

$$S_D(\text{VP}) + S_B(\text{NP}) > S_C(\text{NP}) + S_A(\text{VP})$$

Chance is 50%. A more stringent criterion would be to apply this requirement separately for each of NP and VP postcopular predicates:

$$S_D(\text{VP}) > S_A(\text{VP}) \wedge S_B(\text{NP}) > S_C(\text{NP})$$

However, it is often possible to use an NP postcopular predicate with a *do* cleft through semantic coercion (e.g., in B “did” can be interpreted as “fixed” or “was responsible for”), so we felt that this latter criterion might be too stringent.

References Higgins (1973)

A.1.4 Filler–gap dependencies

Consider the following sentence, in which all arguments and adjuncts appear “in situ” (in the syntactic position at which they are normally interpreted semantically):

I know that our uncle grabbed the food in front of the guests at the holiday party.

A FILLER–GAP DEPENDENCY can be created by EXTRACTING any of a number of elements from the subordinate clause, including *our uncle* (subject extraction), *the food* (object extraction) or *the guests* (extraction from a prepositional phrase). These possibilities serve as the basis for several test suites on filler–gap dependencies.

References Ross (1967); Crain and Fodor (1985); Stowe (1986); Wilcox et al. (2018); Chowdhury and Zamparelli (2018); Chaves (2020)

Subject extractions

- (A) I know that $\overbrace{\text{our uncle}}^{\alpha}$ grabbed the food in front of the guests at the holiday party. [THAT, NO GAP]
- (B) *I know who $\overbrace{\text{our uncle}}^{\alpha}$ grabbed the food in front of the guests at the holiday party. [WH, NO GAP]
- (C) *I know that $\overbrace{\text{grabbed the food in front of the guests at the holiday party}}^{\beta}$. [THAT, GAP]
- (D) I know who $\overbrace{\text{grabbed the food in front of the guests at the holiday party}}^{\beta}$. [WH, GAP]

Criterion We require that a model successfully pass a two-part criterion for each item: the *wh*-filler should make the unextracted subject α more surprising in the NO-GAP conditions and should make the post-gap material β less surprising in the GAP conditions:

$$S_B(\alpha) > S_A(\alpha) \wedge S_C(\beta) > S_D(\beta)$$

Chance is 25%.

Object extractions

The logic of this test suite is the same as that for subject extraction above. Note that we use obligatorily transitive embedded verbs, so that omitting a direct object should be highly surprising when there is no filler, as in C.

- (A) I know that our uncle grabbed $\overbrace{\text{the food}}^{\alpha}$ in front of the guests at the holiday party.
[THAT, NO GAP]
- (B) *I know what our uncle grabbed $\overbrace{\text{the food}}^{\alpha}$ in front of the guests at the holiday party.
[WH, NO GAP]
- (C) ??I know that our uncle grabbed $\overbrace{\text{in front of}}^{\beta}$ the guests at the holiday party. [THAT, GAP]
- (D) I know what our uncle grabbed $\overbrace{\text{in front of}}^{\beta}$ in front of the guests at the holiday party.
[WH, GAP]

Criterion

$$S_B(\alpha) > S_A(\alpha) \wedge S_C(\beta) > S_D(\beta)$$

Extraction from prepositional phrases

The logic of this test suite is the same as that for subject and object extractions above.

- (A) I know that our uncle grabbed the food in front of $\overbrace{\text{the guests}}^{\alpha}$ at the holiday party.
[THAT, NO GAP]
- (B) *I know who our uncle grabbed the food in front of $\overbrace{\text{the guests}}^{\alpha}$ at the holiday party.
[WH, NO GAP]
- (C) *I know that our uncle grabbed the food in front of $\overbrace{\text{at the holiday party}}^{\beta}$. [THAT, GAP]

- (D) I know who our uncle grabbed the food in front of $\overbrace{\text{at the holiday party}}^{\beta}$. [WH, GAP]

Criterion

$$S_B(\alpha) > S_A(\alpha) \wedge S_C(\beta) > S_D(\beta)$$

Tests for unboundedness

Filler–gap dependencies are “unbounded” in the sense that there is no limit to how many clausal levels above the gap the filler can be extracted. This serves as the basis for harder versions of the object-extracted test suites, involving three or four levels of clausal embedding.

Example [THAT, NO GAP] sentences are given below:

I know that our mother said her friend remarked that the park attendant reported your friend threw the plastic into the trash can. [3 levels of embedding]

I know that our mother said her friend remarked that the park attendant reported the cop thinks your friend threw the plastic into the trash can. [4 levels of embedding]

These base sentences give rise to 4-condition test suites using the same manipulations as for the basic object-extraction test suite (Section D), and the criterion for success is the same.

A.1.5 Main-verb/reduced-relative garden-path disambiguation

This is one of the best-studied instances of syntactic garden-pathing in the psycholinguistics literature. An example 4-condition item is given below:

- (A) !The child kicked in the chaos $\overbrace{\text{found}}^{V^*}$ her way back home. [REDUCED, AMBIG]
- (B) The child who was kicked in the chaos $\overbrace{\text{found}}^{V^*}$ her way back home.
- (C) The child forgotten in the chaos $\overbrace{\text{found}}^{V^*}$ her way back home.

(D) The child who was forgotten in the chaos $\overbrace{\text{found}}^{V^*}$ her way back home.

Criterion Relative to the [REDUCED, AMBIG] condition, not reducing the relative clause should make V^* less surprising, as should changing the participial verb to one that is the same form as a simple past-tense verb. Additionally, the effect of not reducing the relative clause on V^* surprisal should be smaller for unambiguous participial verbs than for participial verbs:

$$S_A(V^*) > S_B(V^*) \wedge S_A(V^*) > S_C(V^*) \wedge \\ S_A(V^*) - S_B(V^*) > S_C(V^*) - S_D(V^*)$$

Chance is somewhere below 25%.

References Bever (1970); Ferreira and Clifton (1986); Trueswell et al. (1994); van Schijndel and Linzen (2018); Futrell et al. (2019)

A.1.6 Negative Polarity Licensing

The words *any* and *ever*, in their most common uses, are “negative polarity items” (NPIs): they can only be used in an appropriate syntactic-semantic environment—to a first approximation, in the scope of negation. For example, the determiner *no* can license NPIs, but its NP has to structurally command the NPI. Below, A and D are acceptable, because *no* is the determiner for the subject noun *managers*. There is no negation in C so the NPI is unlicensed and the sentence is unacceptable; crucially, however, B is unacceptable despite the presence of *no* earlier in the sentence, because *no* is embedded inside a modifier of the main-clause subject and thus does not command the NPI.

- (A) No managers that respected the guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [+NEG, −DISTRACTOR]
 (B) *The managers that respected no guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [−NEG, +DISTRACTOR]
 (C) *The managers that respected the guard have had $\overbrace{\text{any}}^{\text{NPI}}$ luck. [−NEG, −DISTRACTOR]

- (D) No managers that respected no guard have had ^{NPI}any luck. [+NEG,+DISTRACTOR]

In the above test suite, the “distractor” position for *no* is inside a subject-extracted relative clause modifying the main-clause subject. We also used a variant test suite in which these relative clauses are object-extracted:

- (A) No managers that the guard respected have had ^{NPI}any luck. [+NEG,-DISTRACTOR]

- (B) *The managers that no guard respected have had ^{NPI}any luck. [-NEG,+DISTRACTOR]

- (C) *The managers that the guard respected have had ^{NPI}any luck. [-NEG,-DISTRACTOR]

- (D) No managers that no guard respected have had ^{NPI}any luck. [+NEG,+DISTRACTOR]

The above two test suites use *any* as the NPI; we also use test suites with *ever* as the NPI. Subject-extracted relative clause example:

- (A) No managers that respected the guard have ^{NPI}ever gotten old. [+NEG,-DISTRACTOR]

- (B) *The managers that respected no guard have ^{NPI}ever gotten old. [-NEG,+DISTRACTOR]

- (C) *The managers that respected the guard have ^{NPI}ever gotten old. [-NEG,-DISTRACTOR]

- (D) No managers that respected no guard have ^{NPI}ever gotten old. [+NEG,+DISTRACTOR]

Object-extracted relative clause example:

- (A) No managers that the guard respected have ^{NPI}ever gotten old. [+NEG,-DISTRACTOR]

- (B) *The managers that no guard respected have ^{NPI}ever gotten old. [-NEG,+DISTRACTOR]

- (C) *The managers that the guard respected have ^{NPI}ever gotten old. [-NEG,-DISTRACTOR]

- (D) No managers that no guard respected have ^{NPI}ever gotten old. [+NEG,+DISTRACTOR]

Criterion Changing the main-clause subject’s determiner from *The* to *No* should increase the probability of the NPI where it appears, regardless of whether there is a distractor *no* in the subject-modifying relative clause. Furthermore, when there is exactly one *no* in the sentence, the NPI should be higher-probability when it is in a licensing position rather than

in a distractor position:

$$P_A(\text{NPI}) > P_C(\text{NPI}) \wedge P_D(\text{NPI}) > P_B(\text{NPI}) \wedge \\ P_A(\text{NPI}) > P_B(\text{NPI})$$

Chance is $\frac{5}{32}$.

References Ladusaw (1979); Vasishth et al. (2008); Giannakidou (2011); Marvin and Linzen (2018); Futrell et al. (2018)

A.1.7 NP/Z garden-path ambiguity

This is another well-studied syntactic garden-pathing configuration. In A below, the NP *the waters* introduces a local syntactic ambiguity: it could be (1) the direct object of *crossed*, in which case the sentence-initial subordinate clause has not yet ended, or (2) the subject of the main clause, in which case *crossed* is used intransitively and is the last word of the sentence-initial subordinate clause. (This was dubbed “NP/Z” by Sturt et al. (1999) because the subordinate-clause verb might have either an NP object or a Z(ero), i.e. null, object.) The next word, *remained*, is only compatible with (2); the ruling out of (1) generally yields increased processing difficulty for human comprehenders. Marking the end of the subordinate clause with a comma, as in B, makes the sentence easier at V^* , as does an obligatorily intransitive subordinate-clause verb, as in C.

- (A) !As the ship crossed the waters $\overbrace{\text{remained}}^{V^*}$ blue and calm. [TRANS,NO COMMA]
- (B) As the ship crossed, the waters $\overbrace{\text{remained}}^{V^*}$ blue and calm. [TRANS,COMMA]
- (C) As the ship drifted the waters $\overbrace{\text{remained}}^{V^*}$ blue and calm. [INTRANS,NO COMMA]
- (D) As the ship drifted, the waters $\overbrace{\text{remained}}^{V^*}$ blue and calm. [INTRANS,COMMA]

Criterion Similar to the main-verb/reduced-relative garden-pathing ambiguity, a model must pass a three-part criterion. Relative to A, either marking the subordinate-clause end

with a comma or using an obligatorily intransitive verb in the subordinate clause should reduce the surprisal of V^* . Furthermore, the surprisal-reduction effect of the comma should be smaller when the subordinate-clause verb is intransitive than when it is transitive:

$$S_A(V^*) > S_B(V^*) \wedge S_A(V^*) > S_C(V^*) \wedge \\ S_A(V^*) - S_B(V^*) > S_C(V^*) - S_D(V^*)$$

We also use an NP/Z test suite where the second means of disambiguation is not changing the subordinate-clause verb to an intransitive, but rather giving the transitive subordinate-clause verb an overt direct object. For the above example item, the first two conditions are the same and the other two conditions would be:

(C) As the ship crossed the sea the waters $\overbrace{\text{remained}}^{V^*}$ blue and calm.

(D) As the ship crossed the sea, the waters $\overbrace{\text{remained}}^{V^*}$ blue and calm.

The success criterion remains the same.

Finally, we create harder versions of both the above test suites by adding a postmodifier to the main-clause subject (in the above example, *the waters* becomes *the waters of the Atlantic Ocean*).

References Frazier and Rayner (1982); Mitchell (1987); Pickering and Traxler (1998); Sturt et al. (1999); Staub (2007)

A.1.8 Subject–verb number agreement

This task tests a language model for how well it predicts the number marking on English finite present-tense verbs (whether it should be the third-person *singular* form, or the non-third-person-singular form, generally referred to as the *plural* form for simplicity, although technically this is the form for first- and second-person singular as well). In controlled, targeted versions of this test, multiple NP precede the verb: the verb’s actual subject, as well as a DISTRACTOR NP with number that is different from that of the subject. A successful language model should place higher probability on the verbform matching that of the subject,

not the distractor. We have three versions of this test suite: one where the distractor is in a prepositional phrase postmodifier of the subject:

- (A) The farmer near the clerks knows_{V_{sg}} many people.
- (B) *The farmer near the clerks know_{V_{pl}} many people.
- (C) The farmers near the clerk know_{V_{pl}} many people.
- (D) *The farmers near the clerk knows_{V_{sg}} many people.

one in which the distractor is in a subject-extracted relative clause postmodifier of the subject:

- (A) The farmer that embarrassed the clerks knows_{V_{sg}} many people.
- (B) *The farmer that embarrassed the clerks know_{V_{pl}} many people.
- (C) The farmers that embarrassed the clerk know_{V_{pl}} many people.
- (D) *The farmers that embarrassed the clerk knows_{V_{sg}} many people.

and one in which the distractor is in an object-extracted relative clause postmodifier of the subject:

- (A) The farmer that the clerks embarrassed knows_{V_{sg}} many people.
- (B) *The farmer that the clerks embarrassed know_{V_{pl}} many people.
- (C) The farmers that the clerk embarrassed know_{V_{pl}} many people.
- (D) *The farmers that the clerk embarrassed knows_{V_{sg}} many people.

Criterion Following Linzen et al. (2016) and Marvin and Linzen (2018), we require successful discrimination of the preferred upcoming verbform of the given lemma (rather than, for example, successful discrimination of the better context given a particular verbform). For success we require that a model successfully predicts the preferred verbform for *both* the singular- and plural-subject versions of an item:

$$P_A(\mathbf{V}_{sg}) > P_B(\mathbf{V}_{pl}) \wedge P_C(\mathbf{V}_{pl}) > P_D(\mathbf{V}_{sg})$$

Chance performance is thus 25%, though a context-insensitive baseline that places different probabilities on V_{sg} and V_{pl} would score 50%.

References Bock and Miller (1991); Linzen et al. (2016); Marvin and Linzen (2018)

A.1.9 Reflexive pronoun licensing

The noun phrase with which a reflexive pronoun (*herself*, *himself*, *themselves*) corefers must command it in a sense similar to that relevant for negative-polarity items (Section A.1.6). In the below example, the reflexive pronoun ending the sentence can only corefer to the subject of the sentence, *author*, with which it must agree in number: a singular subject requires a singular reflexive R_{sg} , and a plural subject requires a plural reflexive R_{pl} .

- (A) The author next to the senators hurt herself $_{R_{sg},fem}$.
- (B) *The authors next to the senator hurt herself $_{R_{sg},fem}$.
- (C) The authors next to the senator hurt themselves $_{R_{pl}}$.
- (D) *The authors next to the senator hurt themselves $_{R_{pl}}$.

We generated a pair of test suites—one in which the singular reflexive is *herself*, and another where the singular reflexive is *himself*, on the template of the above example, where the distractor NP is in a prepositional-phrase postmodifier of the subject NP. We also generated a similar pair of test suites where the distractor NP is inside a subject-extracted relative clause modifying the subject:

- (A) The author that liked the senators hurt herself $_{R_{sg},fem}$.
- (B) *The authors that liked the senator hurt herself $_{R_{sg},fem}$.
- (C) The authors that liked the senator hurt themselves $_{R_{pl}}$.
- (D) *The authors that liked the senator hurt themselves $_{R_{pl}}$.

and a pair of test suites where the distractor NP is inside an object-extracted relative clause modifying the subject:

- (A) The author that the senators liked hurt herself $_{R_{sg},fem}$.

- (B) *The authors that the senator liked hurt herself_{R_{sg.fem}}.
- (C) The authors that the senator liked hurt themselves_{R_{pl}}.
- (D) *The authors that the senator liked hurt themselves_{R_{pl}}.

Criterion For each item in each test suite, we require that for both the singular and the plural versions of the reflexive pronoun the model assign higher conditional probability in the correct licensing context than in the incorrect licensing context:

$$P_A(\mathbf{R}_{sg}) > P_B(\mathbf{R}_{sg}) \wedge P_C(\mathbf{R}_{pl}) > P_D(\mathbf{R}_{pl})$$


Chance is 25%.


References Reinhart (1981); Marvin and Linzen (2018)


A.1.10 Subordination


Beginning a sentence with *As*, *When*, *Before*, *After*, or *Because*, implies that an immediately following clause is not the main clause of the sentence, as would have otherwise been the case, but instead is a SUBORDINATE CLAUSE that must be followed by the main clause. Ending the sentence without a main clause, as in B, is problematic. Conversely, following an initial clause with a second clause MC (without linking it to the initial clause with *and*, *but*, *despite*, or a similar coordinator or subordinator), as in C below, is unexpected and odd.

- (A) The minister praised the building .

END

- (B) *After the minister praised the building .

END

- (C) ??The minister praised the building, it started to rain.

MC

- (D) After the minster praised the building, it started to rain.

MC


In addition to the base test suite exemplified by the item above, we include three versions with longer and more complex initial clauses, which may make the test suite more difficult. In the first of these versions, we postmodify both the subject and object of the initial clauses with prepositional phrases:

the minister praised the building

↓

the minister in the dark suit and white tie praised the new building on the town's main square

In the second of these versions, the postmodifiers are subject-extracted relative clauses:

the minister praised the building

↓

the minister who wore a black suit praised the new building that was built by the square

In the third of these versions, the postmodifiers are object-extracted relative clauses:

the minister praised the building

↓

the minister who the mayor had invited praised the new building that the businessman had built downtown

Criterion Introducing a subordinator at the beginning of the sentence should make an ending without a second clause less probable, and should make a second clause more probable:

$$P_A(\text{END}) > P_B(\text{END}) \wedge P_D(\text{MC}) < P_C(\text{MC})$$

References Futrell et al. (2018)

A.2 Syntactic coverage of test suites

In order to assess the coverage of our syntactic tests, we manually inspected the “Ideas, Rules and Constraints introduced in this Chapter” section for each chapter in Carnie (2012),

CHAPTER 1: GENERATIVE GRAMMAR	Lexical gender Number Person Case	✓
CHAPTER 2: PARTS OF SPEECH	Parts of Speech Plurality Count vs. Mass Nouns Argument Structure of Verbs	✓ ✓ ✓
CHAPTER 3: CONSTITUENCY, TREES, RULES	Constituency Tests Hierarchical Structure	✓
CHAPTER 4: STRUCTURAL RELATIONS	c-command Government	✓
CHAPTER 5: BINDING THEORY	<i>R</i> -expression vs. Pronominals Anaphoric expressions and their antecedents Co-reference and co-indexation Binding Principles (<i>A</i> , <i>B</i> , <i>C</i>) Locality Constraints	✓ ✓ ✓ ✓
CHAPTER 6: X-BAR THEORY	One Replacement Do-so Replacement	
CHAPTER 7: EXTENDING X-BAR THEORY TO FUNCTIONAL CATEGORIES	Fundamental Phrase Types of DP/CP/TP Genitives: of-genitives and 's genitives Subjects and Predicates Clausal Embedding Clausal Tense/Finiteness and its restrictions Yes/No Questions Subject-Auxilliary Inversion	✓
CHAPTER 8: CONSTRAINING X-BAR THEORY: THE LEXICON	Thematic Relations Internal Theta role vs. External Theta Roles Expletive Pronouns and Expletive Insertion Extended Projection Principle	✓
CHAPTER 9: HEAD-TO-HEAD MOVEMENT	V → T Movement T → C movement Do-Support	✓
CHAPTER 10: DP MOVEMENT	Passive Constructions DP-Raising	✓
CHAPTER 11: WH-MOVEMENT	Wh-Movement Structural Constraints on Wh-Movement (Island Constraints) Wh in-Situ and Echo Questions	✓ ✓
CHAPTER 12: A UNIFIED THEORY OF MOVEMENT	Universal Quantifiers vs. Existential Quantifiers Quantificational Scope and Quantifier Raising	
CHAPTER 13: EXTENDED VPS	Light Verbs Object Shift (and end weight) Ellipsis Pseudogapping	
CHAPTER 14: RAISING CONTROL AND EMPTY CATEGORIES	Control, Subject-to-Subject and Subject-to-Object Raising (ECM)	
CHAPTER 15: ADVANCED TOPICS IN BINDING THEORY	Binding Principle <i>A</i> and <i>B</i>	✓

Table A.1: Test suite coverage of syntactic phenomena presented in Carnie (2012).

a standard introductory syntax textbook. We included entries from these sections which are theory-neutral and refer to observable linguistic data. For example, we do not include *affix lowering* (Chapter 7) or *theta criterion* (Chapter 8) because these phenomena presuppose a commitment to one particular syntactic analysis.

We found that our tests covered 16 of the 47 phenomena presented (~34%). Of the 15 chapters surveyed, our tests assessed phenomena in 11 (~73%). We did not assess coverage from the last two chapters of the book, which explore alternative syntactic formalisms. The outcome of our manual inspection is given in Table A.1.

A ✓ indicates that some aspect of that phenomena was tested in one or more of our suites. ✓ does not necessarily mean that the test suite was designed explicitly for the purpose of testing that phenomena, but merely that the phenomena was implicated in model success. For example, we place a ✓ next to *Parts of Speech* because differentiation between verbs and nouns is necessary for models to succeed in the *Cleft Structure* tests.

Appendix B

Supplementary material for Chapter 4

B.1 Example prompts

This section contains example prompts for each task in our experiments. See Section 4.3 and Table 4.1 for details on the materials, and Section 4.4.1 for discussion of how prompts were constructed.

B.1.1 Deceits

Task: You will read short stories that describe two characters interacting with each other. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: Henry is sitting at his desk and watching TV, and reluctantly switches off the TV with the remote control and picks up a textbook. Shortly after, his mother comes in the room and asks, "What have you been doing up here?" Henry responds: "Reading." Why has Henry responded in such a way?

Options:

- 1) He has been reading for some time.
- 2) He does not want to offend his mom by not reading the books that she gave him.
- 3) He does not want to get into trouble for not studying.
- 4) He wants his mom to believe that he has been watching TV.

Answer:

B.1.2 IndirectSpeech

Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer. Your task is to decide what the character in the story is trying to convey. The answer options are 1, 2, 3, or 4.

Scenario: Nate is about to leave the house. His wife points at a full bag of garbage and asks: "Are you going out?" What might she be trying to convey?

Options:

- 1) She wants Nate to spend more time with the family.
- 2) She wants to know Nate's plans.
- 3) She wants Nate to take the garbage out.
- 4) She wants Nate to bring his friends over.

Answer:

B.1.3 Irony

Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer. Your task is to decide what the character in the story is trying to convey. The answer options are 1, 2, 3, or 4.

Scenario: It is a holiday. Stefan and Kim are sitting in the backseat of the car. They are fighting all the time. Their father says: "Oh, it is so pleasant here." What did the father want to convey?

Options:

- 1) He enjoys listening to his kids fighting.
- 2) He remembers about his wife's birthday.
- 3) He does not want to listen to his kids' arguments.
- 4) AC gives them some needed cool.

Answer:

B.1.4 Maxims

Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: Leslie and Jane are chatting at a coffee shop. Leslie asks, "Who was that man that I saw you with last night?" Jane responds, "The latte is unbelievable here." Why has Jane responded like this?

Options:

- 1) She does not want to discuss the topic that Leslie has raised.

- 2) The man who Leslie saw makes unbelievable lattes.
- 3) She thinks that it is the best latte in the town.
- 4) A coffee break is not a good time to discuss men.

Answer:

B.1.5 Metaphor

Task: You will read short stories that describe everyday situations. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. The answer options are 1, 2, 3, 4, or 5.

Scenario: Andrew and Bob were discussing the investment company where Andrew works. Bob said: "The investors are squirrels collecting nuts." What does Bob mean?

Options:

- 1) The investors dress and eat well.
- 2) Squirrels were hired to work in the company.
- 3) Bob is allergic to nuts.
- 4) They buy stocks hoping for future profit.
- 5) The investors enjoy picking nuts as much as squirrels do.

Answer:

B.1.6 Humor

Task: You will read jokes that are missing their punch lines. A punch line is a funny line that finishes the joke. Each joke will be followed by five possible endings. Please choose the ending that makes the joke funny. The answer options are 1, 2, 3, 4, or 5.

Joke: Martha walked into a pastry shop. After surveying all the pastries, she decided on a chocolate pie. "I'll take that one," Martha said to the attendant, "the whole thing." "Shall I cut it into four or eight pieces?" the attendant asked.

Punchlines:

- 1) Martha said, "My leg is hurting so much."
- 2) Martha said, "Four pieces, please; I'm on a diet."
- 3) Martha said: "Well, there are five people for dessert tonight, so eight pieces will be about right."
- 4) Then the attendant squirted whipped cream in Martha's face.
- 5) Martha said, "You make the most delicious sweet rolls in town."

Answer:

B.1.7 Coherence

Task: You will read pairs of sentences. Reach each pair and decide whether they form a coherent story. The answer options are 1 or 2.

Scenario: Cleo brushed against a table with a vase on it. She decided to study harder to catch up.

Options:

- 1) Incoherent
- 2) Coherent

Answer:

B.2 Timestamps of OpenAI model queries

Table B.1 shows timestamps of requests sent to the OpenAI API.

Model	Phenomenon	Timestamp
text-ada-001	Coherence	2022-10-11 12:28 -0400
text-ada-001	Deceits	2022-10-11 12:28 -0400
text-ada-001	IndirectSpeech	2022-10-11 12:28 -0400
text-ada-001	Irony	2022-10-11 12:28 -0400
text-ada-001	Humor	2022-10-11 12:28 -0400
text-ada-001	Maxims	2022-10-11 12:29 -0400
text-ada-001	Metaphor	2022-10-11 12:29 -0400
text-davinci-002	Coherence	2022-10-11 11:56 -0400
text-davinci-002	Deceits	2022-10-11 11:55 -0400
text-davinci-002	IndirectSpeech	2022-10-11 11:55 -0400
text-davinci-002	Irony	2022-10-11 11:54 -0400
text-davinci-002	Humor	2022-10-11 11:53 -0400
text-davinci-002	Maxims	2022-10-11 11:56 -0400
text-davinci-002	Metaphor	2022-10-11 11:57 -0400

Table B.1: Timestamps of OpenAI API model queries.

B.3 No-context analysis

B.3.1 Details of human experiments

Below, we discuss details of the no-context human experiments described in Section 4.5.3. This study was approved by the Institutional Review Board at the home institution of the authors (protocol 2010000243).

Participants. We collected data from 30 participants using Amazon.com’s Mechanical Turk. All participants were recruited from IP addresses in the US, Canada, and other English-speaking countries and passed a brief English proficiency task to participate. We pre-screened participants using a qualification task in which they were asked to perform 10 simple sentence completions, which were judged for basic levels of coherence and grammaticality. Participants were paid 7 USD for completing the study, which took around 20 minutes to complete. The resulting hourly rate was around 21 USD, which is well above federal minimum wage in the United States.

Procedure. Participants completed these tests during one individual testing session. After giving informed consent, which included assurance of anonymity, participants were shown instructions and a training trial, in which they were told they would be answering questions about a character in a short interaction. They then saw 105 trials (similar to those described in Appendix B.1), without the scenario context. For example:

Bob said: "The investors are squirrels collecting nuts." What does Bob mean?

- 1) The investors dress and eat well.
- 2) Squirrels were hired to work in the company.
- 3) Bob is allergic to nuts.
- 4) They buy stocks hoping for future profit.
- 5) The investors enjoy picking nuts as much as squirrels do.

Items were presented within blocks according to their phenomenon, as in Floyd et al.’s (In prep) original experiments. Blocks and items were presented in a random order.

B.3.2 Raw accuracy scores

Figure B-1 shows accuracy scores achieved by humans and the three best-performing models on the original (shaded bars) and no-context (empty bars) versions of the test items.

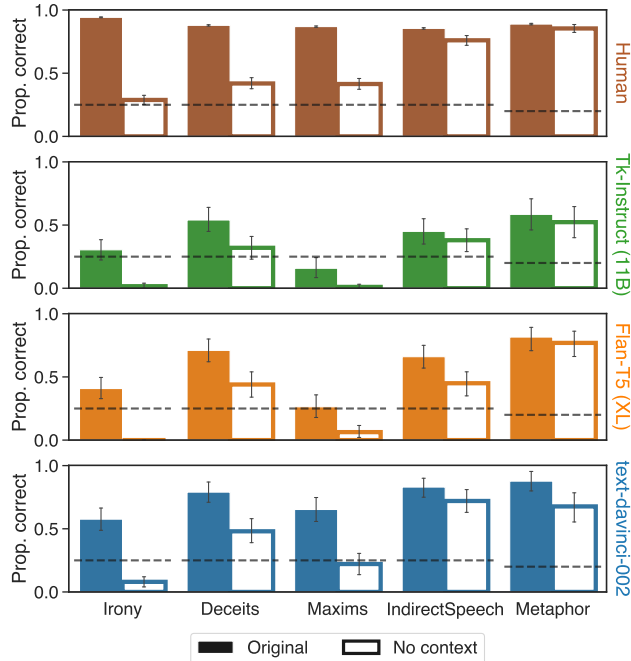


Figure B-1: Proportion of items where humans and models select the correct pragmatic answer, on both original (shaded bars) and no-context (empty bars) versions.

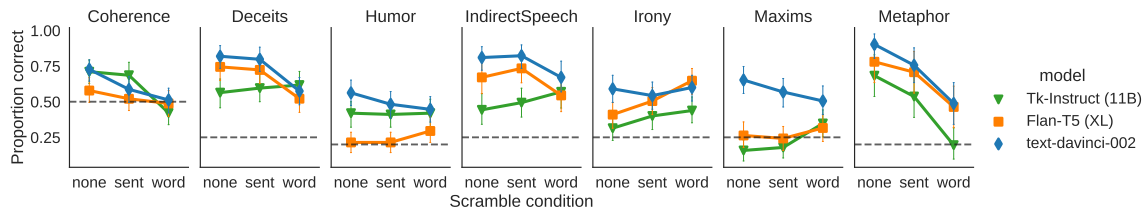


Figure B-2: Model performance across scrambling conditions (none = original, unmodified items). Error bars denote 95% CI. Dashed line indicates random baseline.

B.4 Sentence- and word-level scrambling

Figure B-2 shows accuracy scores achieved by the three best-performing models on each task, across three scrambling conditions: none (original, unmodified items), sentence-level, and word-level. Example prompts are provided below.

B.4.1 Sentence-level scrambled prompt

Task: You will read short stories that describe two characters interacting with each other. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: Dan says, "The dog knocked it over." The vase falls down on the floor and breaks. He brushes against his mother's vase. When Dan's mother comes home, she asks Dan: "What happened to my vase?" Dan is playing in the living room. Why has Dan responded in such a way?

Options:

- 1) Dan does not want his mom to be angry with him for breaking the vase.
- 2) Dan finds this vase ugly and wants to get rid of it.
- 3) Dan wants his mom to know that he knocked it over.
- 4) Dan thinks that the dog has knocked over the vase.

Answer:

B.4.2 Word-level scrambled prompt

Task: You will read short stories that describe two characters interacting with each other. Each story will be followed by a multiple-choice question. Read each story and choose the best answer to each question. Your task is to decide why the character in the story responds in a certain way. The answer options are 1, 2, 3, or 4.

Scenario: to happened Dan "The against in it she comes "What living Dan the vase floor on down The Dan: He dog my brushes vase?" mother When falls breaks. vase. and playing room. his asks knocked says, home, over." the mother's is Dan's Why has Dan responded in such a way?

Options:

- 1) Dan does not want his mom to be angry with him for breaking the vase.
- 2) Dan finds this vase ugly and wants to get rid of it.
- 3) Dan wants his mom to know that he knocked it over.
- 4) Dan thinks that the dog has knocked over the vase.

Answer:

Bibliography

- David Adger. The Autonomy of Syntax. In Norbert Hornstein, Howard Lasnik, Pritty Patel-Grosz, and Charles Yang, editors, *Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics*, pages 153–176. De Gruyter Mouton, Berlin, Boston, 2018. ISBN 978-1-5015-0692-5. URL <https://doi.org/10.1515/9781501506925-157>.
- Gerry T.M. Altmann. Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2(4):146–152, April 1998. ISSN 1364-6613. URL <https://www.sciencedirect.com/science/article/pii/S136466139801153X>.
- John R. Anderson. *The adaptive character of thought*. Erlbaum, Hillsdale, NJ, 1990.
- John R. Anderson and Lael J. Schooler. Reflections of the Environment in Memory. *Psychological Science*, 2(6):396–408, November 1991. ISSN 0956-7976. URL <https://doi.org/10.1111/j.1467-9280.1991.tb00174.x>. Publisher: SAGE Publications Inc.
- Jacob Andreas. Language Models as Agent Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.findings-emnlp.423>.
- Jacob Andreas and Dan Klein. Reasoning about Pragmatics with Neural Listeners and Speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas, November 2016. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/D16-1125>.
- Clara Andrés-Roqueta and Napoleon Katsos. The Contribution of Grammar, Vocabulary and Theory of Mind in Pragmatic Language Competence in Children with Autistic Spectrum Disorders. *Frontiers in Psychology*, 8, 2017. ISSN 1664-1078. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2017.00996>.
- Ian Apperly. *Mindreaders: The cognitive basis of "Theory of Mind"*. Psychology Press, New York, 2011. ISBN 978-1-84169-697-3 (Hardcover).
- Ian A. Apperly and Stephen A. Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, 116(4):953–970, October 2009. ISSN 0033-295X. Place: United States.
- Salvatore Attardo. Irony as relevant inappropriateness. *Journal of Pragmatics*, 32(6):793–826, May 2000. ISSN 0378-2166. URL <https://www.sciencedirect.com/science/article/pii/S0378216699000703>.
- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models, 2021. URL <https://arxiv.org/abs/2108.07732>.
- John L. Austin. *How to do things with words*. 1975.

- Simge Aykan and Erhan Nalçacı. Assessing Theory of Mind by Humor: The Humor Comprehension and Appreciation Test (ToM-HCAT). *Frontiers in Psychology*, 9, 2018. ISSN 1664-1078. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01470>.
- Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. Stop Measuring Calibration When Humans Disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. URL <https://arxiv.org/abs/2210.16133>.
- L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7): 1001–1008, 1989.
- Dare A. Baldwin, Ellen M. Markman, Brigitte Bill, Renee N. Desjardins, Jane M. Irwin, and Glynnis Tidball. Infants’ Reliance on a Social Criterion for Establishing Word-Object Relations. *Child Development*, 67(6):3135–3153, 1996. ISSN 00093920, 14678624. URL <http://www.jstor.org/stable/1131771>. Publisher: [Wiley, Society for Research in Child Development].
- Moshe Bar. Predictions: a universal principle in the operation of the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1181–1182, May 2009. URL <https://doi.org/10.1098/rstb.2008.0321>. Publisher: Royal Society.
- Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a “theory of mind”? *Cognition*, 21(1):37–46, October 1985. ISSN 0010-0277. URL <https://www.sciencedirect.com/science/article/pii/0010027785900228>.
- Marco Baroni. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor & Francis, 2022. URL <https://arxiv.org/abs/2106.08694>.
- Christos Baziotis, Barry Haddow, and Alexandra Birch. Language Model Prior for Low-Resource Neural Machine Translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.615>.
- Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019. URL <https://aclanthology.org/Q19-1004>. Place: Cambridge, MA Publisher: MIT Press.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, 2017.
- Andrea Beltrama and Ming Xiang. Is ‘good’ better than ‘excellent’? An experimental investigation on scalar implicatures and gradable adjectives. *Proceedings of Sinn und Bedeutung*, 17(0), 2013. URL <https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/333>. Section: Articles.
- Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.acl-main.463>.
- Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, March 2003. ISSN 1532-4435. URL <https://dl.acm.org/doi/10.5555/944919.944966>.

- Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL <https://aclanthology.org/D12-1091>.
- Leon Bergen, Roger Levy, and Noah D. Goodman. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9, 2016.
- Tom Bever. The cognitive basis for linguistic structures. In J.R. Hayes, editor, *Cognition and the Development of Language*, pages 279–362. New York: John Wiley & Sons, 1970.
- Anne Beyer, Sharid Loáiciga, and David Schlangen. Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.naacl-main.328>.
- Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P04-3031>.
- Luca Bischetti, Irene Ceccato, Serena Lecce, Elena Cavallini, and Valentina Bambini. Pragmatics and theory of mind in older adults’ humor comprehension. *Current Psychology*, June 2019. ISSN 1936-4733. URL <https://doi.org/10.1007/s12144-019-00295-w>.
- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.703>.
- Kathryn Bock and Carol A. Miller. Broken agreement. *Cognitive Psychology*, 23:45–93, 1991.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avani Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, 2021. URL <https://arxiv.org/abs/2108.07258>.
- W.C. Booth. *A Rhetoric of Irony*. Literature/Criticism. University of Chicago Press, 1974. ISBN 978-0-226-06553-3.

- Emma Borg. On Deflationary Accounts of Human Action Understanding. *Review of Philosophy and Psychology*, 9(3):503–522, September 2018. ISSN 1878-5166. URL <https://doi.org/10.1007/s13164-018-0386-3>.
- Samuel R. Bowman and George Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.naacl-main.385>.
- Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL <https://aclanthology.org/D07-1090>.
- Richard Breheny, Nathan Klinedinst, Jacopo Romoli, and Yasutada Sudo. The symmetry problem: current theories and prospects. *Natural Language Semantics*, 26(2):85–110, June 2018. ISSN 1572-865X. URL <https://doi.org/10.1007/s11050-017-9141-z>.
- Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. Chapter 8 - Two Minds, One Dialog: Coordinating Speaking and Understanding. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, volume 53, pages 301–344. Academic Press, January 2010a. ISBN 0079-7421. URL <https://www.sciencedirect.com/science/article/pii/S0079742110530081>.
- Susan E. Brennan, Alexia Galati, and Anna K. Kuhlen. Two Minds, One Dialog: Coordinating Speaking and Understanding. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, volume 53, pages 301–344. Academic Press, January 2010b. ISBN 0079-7421. URL <https://www.sciencedirect.com/science/article/pii/S0079742110530081>.
- Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- Andreja Bubic, D. Yves Von Cramon, and Ricarda Schubotz. Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, 4, 2010. ISSN 1662-5161. URL <https://www.frontiersin.org/articles/10.3389/fnhum.2010.00025>.
- Brian Buccola, Manuel Križ, and Emmanuel Chemla. Conceptual alternatives: Competition in language and beyond. *Linguistics and Philosophy*, May 2021. ISSN 1573-0549. URL <https://doi.org/10.1007/s10988-021-09327-w>.
- Stephen A. Butterfill and Ian A. Apperly. How to Construct a Minimal Theory of Mind. *Mind & Language*, 28(5):606–637, November 2013. ISSN 0268-1064. URL <https://doi.org/10.1111/mila.12036>. Publisher: John Wiley & Sons, Ltd.
- Joan L. Bybee and Clay Beckner. Usage-based theory. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*. Oxford University Press, 2015.

- Carl Camden, Michael T. Motley, and Ann Wilson. White lies in interpersonal communication: A taxonomy and preliminary investigation of social motivations. *Western Journal of Speech Communication*, 48(4):309–325, December 1984. ISSN 0193-6700. URL <https://doi.org/10.1080/10570318409374167>. Publisher: Routledge.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With Little Power Comes Great Responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.745>.
- Andrew Carnie. *Syntax: A generative introduction*, volume 18. John Wiley & Sons, 2012.
- Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43, 2000. Linguistic Data Consortium.
- Nick Chater and Christopher D. Manning. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344, 2006. ISSN 1364-6613. URL <https://www.sciencedirect.com/science/article/pii/S1364661306001318>.
- Rui P. Chaves. What don't RNN language models learn about filler-gap dependencies? In *Proceedings of the Society for Computation in Linguistics*, 2020.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, 2021. URL <https://arxiv.org/abs/2107.03374>.
- Noam Chomsky. *Syntactic Structures*. Walter de Gruyter, 1957.
- Noam Chomsky. A Review of B.F. Skinner's Verbal Behavior. *Language*, 35(1):26–58, 1959. URL <https://web-archive.southampton.ac.uk/cogprints.org/1148/1/chomsky.htm>.
- Noam Chomsky. *Aspects of the Theory of Syntax*. MIT Press, 1965.
- Noam Chomsky. Quine's Empirical Assumptions. *Synthese*, 19(1):53–68, December 1968. ISSN 1573-0964. URL <https://doi.org/10.1007/BF00568049>.
- Noam Chomsky. On Cognitive Structures and their Development: A reply to Piaget. In Massimo Piattelli-Palmarini, editor, *Language and Learning: The debate between Jean Piaget and Noam Chomsky*. Harvard University Press, 1980.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi

- Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways, 2022. URL <https://arxiv.org/abs/2204.02311>.
- Shammur Absar Chowdhury and Roberto Zamparelli. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA, August 2018. URL <https://www.aclweb.org/anthology/C18-1012>.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, 2022. URL <https://arxiv.org/abs/2210.11416>.
- Herbert H. Clark and Susan E. Brennan. Grounding in communication. In *Perspectives on Socially Shared Cognition.*, pages 127–149. American Psychological Association, 1991. ISBN 1-55798-121-3.
- Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In A.K. Joshi, B. Webber, and I.A. Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, 1981.
- Pablo Contreras Kallens, Ross Deans Kristensen-McLachlan, and Morten H. Christiansen. Large Language Models Demonstrate the Potential of Statistical Learning in Language. *Cognitive Science*, 47(3):e13256, March 2023. ISSN 0364-0213. URL <https://doi.org/10.1111/cogs.13256>. Publisher: John Wiley & Sons, Ltd.
- G. G. Coulton. The Princes of the World. In *From St. Francis to Dante*, Translations from the Chronicle of the Franciscan Salimbene, 1221-1288, pages 239–256. University of Pennsylvania Press, 2 edition, 1972. ISBN 978-0-8122-7672-5. URL <http://www.jstor.org/stable/j.ctv4t8279.25>.
- Stephen Crain and Janet Dean Fodor. How can grammars help parsers? In David Dowty, Lauri Karttunen, and Arnold M. Zwicky, editors, *Natural Language Parsing: Psycholinguistic, Computational, and Theoretical Perspectives*, pages 940–128. Cambridge: Cambridge University Press, 1985.
- Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203):129–132, January 1989. ISSN 1476-4687. URL <https://doi.org/10.1038/337129a0>.
- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. URL <https://www.aclweb.org/anthology/P19-1285>.
- Robert Dale. NLP commercialisation in the last 25 years. *Natural Language Engineering*, 25(3):419–426, 2019. ISSN 1351-3249. URL <https://www.cambridge.org/core/article/nlp-commercialisation-in-the-last-25-years/ECC5887188379FC218E00664767CD2B2>.
- Ishita Dasgupta, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman, and Noah D. Goodman. Evaluating Compositionality in Sentence Embeddings. In *Proceedings of the Cognitive Science Society*, 2018. URL <https://arxiv.org/abs/1802.04302>.
- Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. “Was It Good? It Was Provocative.” Learning the Meaning of Scalar Adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P10-1018>.

- Gerard de Melo and Mohit Bansal. Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290, July 2013. ISSN 2307-387X. URL https://doi.org/10.1162/tacl_a_00227.
- Lambert Deckers and Philip Kizer. Humor and the Incongruity Hypothesis. *The Journal of Psychology*, 90(2): 215–218, 1975. URL <https://doi.org/10.1080/00223980.1975.9915778>.
- Judith Degen. *Alternatives in Pragmatic Reasoning*. Ph.D., University of Rochester, 2013. URL <https://www.proquest.com/dissertations-theses/alternatives-pragmatic-reasoning/docview/1465060224/se-2?accountid=12492>.
- Judith Degen. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8(11):1–55, May 2015.
- Judith Degen. The Rational Speech Act Framework. *Annual Review of Linguistics*, 9(1):519–540, 2023. URL <https://doi.org/10.1146/annurev-linguistics-031220-010811>. [_eprint: https://doi.org/10.1146/annurev-linguistics-031220-010811](https://doi.org/10.1146/annurev-linguistics-031220-010811).
- Judith Degen and Michael K. Tanenhaus. Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science*, 39(4):667–710, 2015. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12171>.
- Judith Degen and Michael K. Tanenhaus. Availability of Alternatives and the Processing of Scalar Implicatures: A Visual World Eye-Tracking Study. *Cognitive Science*, 40(1):172–201, January 2016. ISSN 0364-0213. URL <https://doi.org/10.1111/cogs.12227>. Publisher: John Wiley & Sons, Ltd.
- Judith Degen, Michael Henry Tessler, and Noah D. Goodman. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, 2015. URL <https://cocolab.stanford.edu/papers/DegenEtAl2015-Cogsci.pdf>.
- Marie-Julie Demedardi, Claire Brechet, Edouard Gentaz, and Catherine Monnier. Prosocial lying in children between 4 and 11 years of age: The role of emotional understanding and empathy. *Journal of Experimental Child Psychology*, 203:105045, March 2021. ISSN 0022-0965. URL <https://www.sciencedirect.com/science/article/pii/S0022096520304999>.
- Hanneke Den Ouden, Peter Kok, and Floris De Lange. How Prediction Errors Shape Perception, Attention, and Motivation. *Frontiers in Psychology*, 3, 2012. ISSN 1664-1078. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00548>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/N19-1423>.
- Katharina Dobs, Joanne Yuan, Julio Martinez, and Nancy Kanwisher. Using deep convolutional neural networks to test why human face recognition works the way it does. *bioRxiv*, 2022. URL <https://www.biorxiv.org/content/early/2022/11/24/2022.11.23.517478>. Publisher: Cold Spring Harbor Laboratory [_eprint: https://www.biorxiv.org/content/early/2022/11/24/2022.11.23.517478.full.pdf](https://www.biorxiv.org/content/early/2022/11/24/2022.11.23.517478.full.pdf).
- Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace Lindsay, Konrad Kording, Talia Konkle, Marcel A. J. Van Gerven, Nikolaus Kriegeskorte, and Tim C. Kietzmann. The neuroconnectionist research programme, 2022. URL <https://arxiv.org/abs/2209.03718>.
- Judit Dombi, Tetyana Sydorenko, and Veronika Timpe-Laughlin. Common ground, cooperation, and recipient design in human-computer interactions. *Journal of Pragmatics*, 193:4–20, May 2022. ISSN 0378-2166. URL <https://www.sciencedirect.com/science/article/pii/S0378216622000716>.

- Ryan Doran, Rachel E. Baker, Yaron McNabb, Meredith Larson, and Gregory Ward. On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, 1(2):211–248, January 2009. URL https://brill.com/view/journals/irp/1/2/article-p211_1.xml.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The Hitchhiker’s Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL <https://aclanthology.org/P18-1128>.
- Gabe Dupre. (What) Can Deep Learning Contribute to Theoretical Linguistics? *Minds and Machines*, 31(4):617–635, December 2021. ISSN 1572-8641. URL <https://doi.org/10.1007/s11023-021-09571-w>.
- Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2016.
- Tiwalayo Eisape, Vineet Gangireddy, Roger P. Levy, and Yoon Kim. Probing for Incremental Parse States in Autoregressive Language Models. In *Findings of EMNLP*. Association for Computational Linguistics, 2022. URL <https://arxiv.org/abs/2211.09748>.
- Sarah F. V. Eiteljoerge, Nausicaa Pouscoulous, and Elena V. M. Lieven. Some Pieces Are Missing: Implicature Production in Children. *Frontiers in Psychology*, 9:1928, 2018. ISSN 1664-1078.
- Ivan Enrici, Bruno G. Bara, and Mauro Adenzato. Theory of Mind, pragmatics and the brain: Converging evidence for the role of intention processing as a core feature of human communication. *Pragmatics & Cognition*, 26(1):5–38, 2019. ISSN 0929-0907. URL <https://www.jbe-platform.com/content/journals/10.1075/pc.19010.enr>.
- Fernanda Ferreira and Charles Clifton, Jr. The independence of syntactic processing. *Journal of Memory and Language*, 25:348–368, 1986.
- Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, October 2020. URL <https://doi.org/10.1073/pnas.1905334117>.
- R. Holly Fitch and Paula Tallal. Neural Mechanisms of Language-Based Learning Impairments: Insights from Human Populations and Animal Models. *Behavioral and Cognitive Neuroscience Reviews*, 2(3): 155–178, September 2003. ISSN 1534-5823. URL <https://doi.org/10.1177/1534582303258736>. Publisher: SAGE Publications.
- Sammy Floyd, Olessia Jouravlev, Zachary Mineroff, Leon Bergen, Evelina Fedorenko, and Edward Gibson. Deciphering the structure of pragmatics: A large-scale individual differences investigation. In prep.
- Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975.
- Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28:3–71, 1988.
- Victoria Fossum and Roger P. Levy. Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–69, 2012.
- Danny Fox and Roni Katzir. On the characterization of alternatives. *Natural Language Semantics*, 19(1): 87–107, 2011. URL <https://doi.org/10.1007/s11050-010-9065-3>. ISBN: 1572-865X.

- Michael C. Frank and Noah D. Goodman. Predicting Pragmatic Reasoning in Language Games. *Science*, 336 (6084):998–998, 2012. URL <http://science.sciencemag.org/content/336/6084/998>.
- Michael C Frank, Andrés Gómez Emilsson, Benjamin Peloquin, Noah D. Goodman, and Christopher Potts. Rational speech act models of pragmatic reasoning in reference games, 2018. URL psyarxiv.com/f9y6b.
- Stefan L. Frank and Rens Bod. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological Science*, 22(6):829–834, June 2011. ISSN 0956-7976. URL <https://doi.org/10.1177/0956797611409589>. Publisher: SAGE Publications Inc.
- Michael Franke and Gerhard Jäger. Probabilistic pragmatics, or why Bayes’ rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1):3–44, 2016. URL <https://doi.org/10.1515/zfs-2016-0002>.
- Lyn Frazier and Keith Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14:178–210, 1982.
- Riccardo Fusaroli, Bahador Bahrami, Karsten Olsen, Andreas Roepstorff, Geraint Rees, Chris Frith, and Kristian Tylén. Coming to Terms: Quantifying the Benefits of Linguistic Coordination. *Psychological Science*, 23(8):931–939, August 2012. ISSN 0956-7976. URL <https://doi.org/10.1177/0956797612436816>. Publisher: SAGE Publications Inc.
- Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. RNNs as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*, 2018. URL <http://arxiv.org/abs/1809.01329>.
- Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *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)*, pages 32–42, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL <https://aclanthology.org/N19-1004>.
- S. Gallagher. The practice of mind: Theory, simulation or primary interaction? *Journal of Consciousness Studies*, 8(5-6):83–108, 2001. ISSN 1355-8250. URL <https://www.ingentaconnect.com/content/imp/jcs/2001/00000008/f0030005/1207>.
- M.F. Garrett. The Analysis of Sentence Production. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 9, pages 133–177. Academic Press, January 1975. ISBN 0079-7421. URL <https://www.sciencedirect.com/science/article/pii/S0079742108602704>.
- Aina Garí Soler and Marianna Apidianaki. BERT Knows Punta Cana is not just beautiful, it’s gorgeous: Ranking Scalar Adjectives with Contextualised Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.598>.
- Aina Garí Soler and Marianna Apidianaki. Scalar Adjective Identification and Multilingual Ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.naacl-main.370>.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online, July 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-demos.10>.

- Gerald Gazdar. *Pragmatics: Implicature, presupposition, and logical form*. Academic Press, New York, 1979.
- Dirk Geeraerts. Introduction: A rough guide to Cognitive Linguistics. In Dirk Geeraerts, editor, *Cognitive Linguistics: Basic Readings*, pages 1–28. De Gruyter Mouton, Berlin, New York, 2006. ISBN 978-3-11-019990-1. URL <https://doi.org/10.1515/9783110199901.1>.
- Bart Geurts and Paula Rubio-Fernández. Pragmatics and Processing. *Ratio*, 28(4):446–469, 2015. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/rati.12113>.
- Anastasia Giannakidou. Negative and positive polarity items: Variation, licensing, and compositionality. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An international handbook of natural language meaning*, volume 3, pages 1660–1712. Berlin: Mouton de Gruyter, 2011.
- Raymond W. Gibbs. Contextual effects in understanding indirect requests. *Discourse Processes*, 2(1):1–10, January 1979. ISSN 0163-853X. URL <https://doi.org/10.1080/01638537909544450>. Publisher: Routledge.
- Edward Gibson, Richard Futrell, Steven P. Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences*, 23(5):389–407, May 2019. ISSN 1364-6613. URL <https://www.sciencedirect.com/science/article/pii/S1364661319300580>.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, 2018.
- J.J. Godfrey, E.C. Holliman, and J. McDaniel. Switchboard: A telephone speech corpus for research and development. In *International Conference on Acoustics, Speech and Signal Processing*, pages 517–520, 1992.
- Yoav Goldberg. Assessing BERT’s syntactic abilities. *arXiv preprint arXiv:1901.05287*, 2019. URL <https://arxiv.org/abs/1901.05287>.
- Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah, January 2018. Association for Computational Linguistics. URL <https://aclanthology.org/W18-0102>.
- Noah D. Goodman and Michael C. Frank. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016. URL <https://doi.org/10.1016/j.tics.2016.08.005>.
- Noah D. Goodman and Daniel Lassiter. Probabilistic Semantics and Pragmatics: Uncertainty in Language and Thought. In *The Handbook of Contemporary Semantic Theory*, pages 655–686. John Wiley & Sons, Ltd, 2015. ISBN 978-1-118-88213-9. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118882139.ch21>.
- Noah D. Goodman, Joshua B. Tenenbaum, and Tobias Gerstenberg. Concepts in a probabilistic language of thought. In *The Conceptual Mind: New Directions in the Study of Concepts*, pages 623–653. MIT Press, 2015. URL <http://cicl.stanford.edu/papers/goodman2015concepts.pdf>.
- Nicole Gotzner, Stephanie Solt, and Anton Benz. Scalar Diversity, Negative Strengthening, and Adjectival Semantics. *Frontiers in Psychology*, 9:1659, 2018. ISSN 1664-1078. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2018.01659>.

- Herbert P. Grice. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press, 1975. URL <http://www.ucl.ac.uk/lis/studypacks/Grice-Logic.pdf>.
- Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8): 357–364, August 2010. ISSN 1364-6613. URL <https://doi.org/10.1016/j.tics.2010.05.004>.
- Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.flp-1.12>.
- Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148, September 2017. ISSN 0885-2308. URL <https://www.sciencedirect.com/science/article/pii/S0885230816301395>.
- Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana, June 2018. URL <https://www.aclweb.org/anthology/N18-1108>.
- Michael Hahn, Dan Jurafsky, and Richard Futrell. Universals of word order reflect optimization of grammars for efficient communication. *Proceedings of the National Academy of Sciences*, 117(5):2347–2353, 2020. ISSN 0027-8424. URL <https://www.pnas.org/content/117/5/2347>.
- John Hale. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8, 2001.
- Francesca G.E. Happé. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119, August 1993. ISSN 0010-0277. URL <https://www.sciencedirect.com/science/article/pii/001002779390026R>.
- Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298(5598):1569–1579, November 2002. URL <https://doi.org/10.1126/science.298.5598.1569>. Publisher: American Association for the Advancement of Science.
- Robert D. Hawkins, Michael Franke, Michael C. Frank, Adele E. Goldberg, Kenny Smith, Thomas L. Griffiths, and Noah D. Goodman. From partners to populations: A hierarchical Bayesian account of coordination and convention. *Psychological Review*, 2021. URL <https://arxiv.org/abs/2104.05857>.
- Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992. URL <https://aclanthology.org/C92-2082>.
- Santosh A Helekar. *Animal models of speech and language disorders*. Springer, 2013.
- Daphna Heller, Kristen S. Gorman, and Michael K. Tanenhaus. To Name or to Describe: Shared Knowledge Affects Referential Form. *Topics in Cognitive Science*, 4(2):290–305, 2012. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1756-8765.2012.01182.x>.
- Herodotus. *Histories*. Number II. 440 B.C.

- John Hewitt and Christopher D. Manning. A Structural Probe for Finding Syntax in Word Representations. In *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)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL <https://aclanthology.org/N19-1419>.
- Cecilia Heyes. Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, 9(2):131–143, March 2014. ISSN 1745-6916. URL <https://doi.org/10.1177/1745691613518076>.
- Francis Roger Higgins. *The Pseudo-Cleft Construction in English*. PhD thesis, MIT, September 1973.
- G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. Distributed Representations. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*, pages 77–109. MIT Press, Cambridge, MA, USA, 1986. ISBN 0-262-68053-X.
- Thomas Hobbes. *Leviathan*. 1651.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Laurence R. Horn. *On the semantic properties of logical operators in English*. PhD Thesis, University of California Los Angeles, 1972. URL <https://linguistics.ucla.edu/images/stories/Horn.1972.pdf>.
- Laurence R. Horn. *A Natural History of Negation*. Chicago University Press, 1989.
- Yik Kwan Hsu and Him Cheung. Two mentalizing capacities and the understanding of two types of lie telling in children. *Developmental Psychology*, 49:1650–1659, 2013. URL <https://doi.org/10.1037/a0031128>.
- Jennifer Hu, Sherry Yong Chen, and Roger Levy. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York, January 2020a. Association for Computational Linguistics. URL <https://aclanthology.org/2020.scil-1.39>.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online, July 2020b. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-main.158>.
- Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023a. URL <https://arxiv.org/abs/2212.06801>. To appear.
- Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics*, 2023b. To appear.
- D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of Physiology*, 160(1):106–154, 1962. URL <https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1962.sp006837>.
- Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online, November 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.conll-1.49>.

- James R. Hurford. *The Origins of Meaning: Language in the Light of Evolution*. Studies in the Evolution of Language. Oxford University Press, 2007. ISBN 978-0-19-160723-3.
- James R. Hurford, Michael Studdert-Kennedy, and Chris Knight, editors. *Approaches to the Evolution of Language: Social and Cognitive Biases*. Cambridge University Press, 1998.
- Anna A. Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons, Brain, Data, and Theory*, 2022. URL <https://arxiv.org/abs/2208.10668>.
- Nir Jacoby and Evelina Fedorenko. Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Language, Cognition and Neuroscience*, 35(6):780–796, 2020. URL <https://doi.org/10.1080/23273798.2018.1525494>.
- Shailee Jain, Vy A. Vo, Leila Wehbe, and Alexander G. Huth. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, pages 1–65, January 2023. ISSN 2641-4368. URL https://doi.org/10.1162/nol_a_00101.
- Masoud Jasbi, Brandon Waldon, and Judith Degen. Linking Hypothesis and Number of Response Options Modulate Inferred Scalar Implicature Rate. *Frontiers in Psychology*, 10:189, 2019. ISSN 1664-1078. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2019.00189>.
- F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss. A Dynamic Language Model for Speech Recognition. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991. URL <https://aclanthology.org/H91-1057>.
- Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are Natural Language Inference Models IMPPRESsive? Learning IMPLICature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online, July 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-main.768>.
- Nan-Jiang Jiang and Marie-Catherine de Marneffe. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 2022. URL <https://arxiv.org/abs/2209.03392>.
- Philip E. B. Jourdain. The Logical Work of Leibniz. *The Monist*, 26(4):504–523, 1916. ISSN 00269662. URL <http://www.jstor.org/stable/27900607>. Publisher: Oxford University Press.
- Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016. URL <https://arxiv.org/abs/1602.02410>.
- Dan Jurafsky. Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*. MIT Press, 2002. URL <https://web.stanford.edu/~jurafsky/prob.pdf>.
- Daniel Jurafsky and James H. Martin. N-gram language models. In *Speech and Language Processing*. 2023. URL <https://web.stanford.edu/~jurafsky/slp3/3.pdf>.
- Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. Using artificial neural networks to ask ‘why’ questions of minds and brains. *Trends in Neurosciences*, 2023. ISSN 0166-2236. URL <https://doi.org/10.1016/j.tins.2022.12.008>.
- Justine T. Kao and Noah D. Goodman. Let’s talk (ironically) about the weather: Modeling verbal irony. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2014. URL <https://cocolab.stanford.edu/papers/KaoEtAl2015-Cogsci.pdf>.

- Justine T. Kao, Leon Bergen, and Noah D. Goodman. Formalizing the Pragmatics of Metaphor Understanding. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2014. URL <https://escholarship.org/uc/item/09h3p4cz>.
- Roni Katzir. Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6):669–690, 2007. URL <https://doi.org/10.1007/s10988-008-9029-y>. ISBN: 1573-0549.
- Roni Katzir. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi (2023). 2023. URL <https://ling.auf.net/lingbuzz/007190>.
- Alexander J.E. Kell, Daniel L.K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), May 2018. ISSN 0896-6273. URL <https://doi.org/10.1016/j.neuron.2018.03.044>.
- Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):1–29, November 2014. URL <https://doi.org/10.1371/journal.pcbi.1003915>.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking Benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.naacl-main.324>.
- Joo-Kyung Kim and Marie-Catherine de Marneffe. Deriving Adjectival Scales from Continuous Space Word Representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL <https://aclanthology.org/D13-1169>.
- Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, August 2015. ISSN 0010-0277. URL <https://www.sciencedirect.com/science/article/pii/S0010027715000815>.
- Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/P18-1249>.
- Melissa Kline Struhl, Jeanne Gallée, Zuzanna Balewski, and Evelina Fedorenko. Understanding jokes draws most heavily on the Theory of Mind brain network, 2018. URL <https://psyarxiv.com/h2nyx>.
- Philipp Koehn. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-3250>.
- Genevieve Konopka and Todd F. Roberts. Animal Models of Speech and Vocal Communication Deficits Associated With Psychiatric Disorders. *Animal Models of Psychiatric Disease*, 79(1):53–61, January 2016. ISSN 0006-3223. URL <https://www.sciencedirect.com/science/article/pii/S0006322315005703>.
- Michal Kosinski. Theory of Mind May Have Spontaneously Emerged in Large Language Models, 2023. URL <https://arxiv.org/abs/2302.02083>.

- Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. Concladia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.308>.
- Anthony Kroch. Lexical and inferred meanings for some time adverbs. *Quarterly Progress Reports of the Research Laboratory of Electronics*, 104:260–267, 1972.
- R. Kuhn and R. De Mori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583, 1990. URL <https://ieeexplore.ieee.org/document/56193>.
- William Ladusaw. *Polarity Sensitivity as Inherent Scope Relations*. PhD thesis, University of Texas at Austin, 1979.
- G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, 1980. URL <https://press.uchicago.edu/ucp/books/book/chicago/M/bo3637992.html>.
- Andrew Kyle Lampinen. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans, 2023. URL <https://arxiv.org/abs/2210.15303>.
- Daniel Lassiter. How not to identify a scalar implicature (The importance of priors), 2022. Presentation at Cognitive Semantic and Quantities Workshop, University of Amsterdam.
- Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 5:1202–1247, 2017.
- Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the Evaluation of Theory of Mind through Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-1598>.
- Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan DeVyllder, Michel Walter, Sofian Berrouiguet, and Christophe Lemey. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *J Med Internet Res*, 23(5): e15708, May 2021. ISSN 1438-8871. URL <http://www.ncbi.nlm.nih.gov/pubmed/33944788>.
- Alan M. Leslie, Ori Friedman, and Tim P. German. Core mechanisms in ‘theory of mind’. *Trends in Cognitive Sciences*, 8(12):528–533, December 2004. ISSN 1364-6613. URL <https://www.sciencedirect.com/science/article/pii/S1364661304002608>.
- Willem J.M. Levelt. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, July 1983. ISSN 0010-0277. URL <https://www.sciencedirect.com/science/article/pii/0010027783900264>.
- Stephen Levinson. *Presumptive meaning: The theory of generalized conversational implicature*. MIT Press, 2000.
- Roger Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126 – 1177, 2008. ISSN 0010-0277. URL <http://www.sciencedirect.com/science/article/pii/S0010027707001436>.
- Alexander K. Lew, Michael Henry Tessler, Vikash K. Mansinghka, and Joshua B. Tenenbaum. Leveraging Unstructured Statistical Knowledge in a Probabilistic Language of Thought. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 2020. URL <https://cognitivesciencesociety.org/cogsci20/papers/0520/0520.pdf>.

- Elissa Li, Sebastian Schuster, and Judith Degen. Predicting Scalar Inferences From "Or" to "Not Both" Using Neural Sentence Encoders. In *Proceedings of the Society for Computation in Linguistics*, volume 4, 2021. URL <https://doi.org/10.7275/xr01-a852>.
- Tal Linzen. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online, July 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-main.465>.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. URL <https://aclanthology.org/Q16-1037>. Place: Cambridge, MA Publisher: MIT Press.
- Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016. URL http://www.lrec-conf.org/proceedings/lrec2016/pdf/947_Paper.pdf.
- Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the Ability of Language Models to Interpret Figurative Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States, July 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.naacl-main.330>.
- Nikos K. Logothetis and David L. Sheinberg. Visual Object Recognition. *Annual Review of Neuroscience*, 19(1):577–621, 1996. URL <https://doi.org/10.1146/annurev.ne.19.030196.003045>. [_eprint: https://doi.org/10.1146/annurev.ne.19.030196.003045](https://doi.org/10.1146/annurev.ne.19.030196.003045).
- Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. URL <https://aclanthology.org/W02-0109>.
- Charles Lovering and Ellie Pavlick. Unit Testing for Concepts in Neural Networks. *Transactions of the Association for Computational Linguistics*, 10:1193–1208, November 2022. ISSN 2307-387X. URL https://doi.org/10.1162/tac1_a_00514.
- Eleonore Lumer and Hendrik Buschmeier. Modeling Social Influences on Indirectness in a Rational Speech Act Approach to Politeness. In *Proceedings of the 44th Annual Conference of the Cognitive Science Society*, 2022. URL <https://escholarship.org/uc/item/7qg325fr>.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models: A cognitive perspective, 2023. URL <https://arxiv.org/abs/2301.06627>.
- Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, December 2020. URL <https://doi.org/10.1073/pnas.1907367117>. Publisher: Proceedings of the National Academy of Sciences.
- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330, 1993.
- David Marr. *Vision: A Computational Approach*. Freeman & Co., San Francisco, 1982.
- R.A. Martin and T. Ford. *The Psychology of Humor: An Integrative Approach*. Academic Press, 2018. ISBN 978-0-12-813509-9.

- Rebecca Marvin and Tal Linzen. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL <https://aclanthology.org/D18-1151>.
- James L. McClelland and Timothy T. Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, April 2003. ISSN 1471-0048. URL <https://doi.org/10.1038/nrn1076>.
- Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. URL <https://aclanthology.org/P19-1334>.
- James McGilvray. *Chomsky: Language, Mind, Politics*. 2 edition, 2014.
- Danny Merx and Stefan L. Frank. Human Sentence Processing: Recurrence or Attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.cmcl-1.2>.
- Julian Michael. To Dissect an Octopus: Making Sense of the Form/Meaning Debate, 2020. URL <https://julianmichael.org/blog/2020/07/23/to-dissect-an-octopus.html>.
- Ashley Micklos and Marieke Woensdregt. Cognitive and Interactive Mechanisms for Mutual Understanding in Conversation, August 2022. URL psyarxiv.com/aqtfb.
- G.A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- George A. Miller and Noam Chomsky. Finitary models of language users. In R. Duncan Luce, Robert R. Bush, and Eugene Galanter, editors, *Handbook of Mathematical Psychology*, volume II, pages 419–491. New York: John Wiley & Sons, Inc., 1963.
- Daniel Milway. A Response to Piantadosi (2023). 2023. URL <https://lingbuzz.net/lingbuzz/007264>.
- Lisa Miracchi. A competence framework for artificial intelligence research. *Philosophical Psychology*, 32(5):588–633, July 2019. ISSN 0951-5089. URL <https://doi.org/10.1080/09515089.2019.1607692>.
- Don C. Mitchell. Lexical guidance in human parsing: Locus and processing characteristics. In Max Coltheart, editor, *Attention and Performance XII: The psychology of reading*. London: Erlbaum, 1987.
- Shima Rahimi Moghaddam and Christopher J. Honey. Boosting Theory-of-Mind Performance in Large Language Models via Prompting, 2023. URL <https://arxiv.org/abs/2304.11490>.
- Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in Context: A Pragmatic Neural Model for Grounded Language Understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017. URL <https://www.aclweb.org/anthology/Q17-1023>.
- Jarrold Moss and Christian D. Schunn. Comprehension through explanation as the interaction of the brain’s coherence and cognitive control networks. *Frontiers in Human Neuroscience*, 9, 2015. ISSN 1662-5161. URL <https://www.frontiersin.org/articles/10.3389/fnhum.2015.00562>.
- Mindaugas Mozuraitis, Suzanne Stevenson, and Daphna Heller. Modeling Reference Production as the Probabilistic Combination of Multiple Perspectives. *Cognitive Science*, 42(S4):974–1008, 2018. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12582>.

- Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. Evaluating Theory of Mind in Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL <https://aclanthology.org/D18-1261>.
- Allen Newell and Herbert A. Simon. Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM*, 19(3):113–126, March 1976. ISSN 0001-0782. URL <https://doi.org/10.1145/360018.360022>.
- Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. Pragmatic Issue-Sensitive Image Captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online, November 2020a. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.findings-emnlp.173>.
- Yixin Nie, Yicheng Wang, and Mohit Bansal. Analyzing Compositionality-Sensitivity of NLI Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874, July 2019. URL <https://ojs.aaai.org/index.php/AAAI/article/view/4663>.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online, July 2020b. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-main.441>.
- Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, March 2018. URL <https://doi.org/10.1073/pnas.1708274114>. Publisher: Proceedings of the National Academy of Sciences.
- Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show Your Work: Scratchpads for Intermediate Computation with Language Models, 2021. URL <https://arxiv.org/abs/2112.00114>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.
- Elizabeth Pankratz and Bob van Tiel. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4):562–594, 2021. ISSN 1866-9808. URL <https://www.cambridge.org/core/article/role-of-relevance-for-scalar-diversity-a-usagebased-approach/9DFF372956A0A93ABB4DC6CA8F51DEE7>.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *Neural Information Processing Systems Autodiff Workshop*, 2017.
- Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=gJcEM8sxHK>.
- Ellie Pavlick and Tom Kwiatkowski. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, November 2019. ISSN 2307-387X. URL https://doi.org/10.1162/tac1_a_00293.

- Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. URL <https://aclanthology.org/D14-1162>.
- Slav Petrov and Dan Klein. Improved inference for unlexicalized parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April 2007. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/N07-1051>.
- Steven T. Piantadosi. Modern language models refute Chomsky’s approach to language. 2023. URL <https://lingbuzz.net/lingbuzz/007180>.
- Martin J. Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190, 2004. ISSN 0140-525X. URL <https://www.cambridge.org/core/article/toward-a-mechanistic-psychology-of-dialogue/83442BA495E0D5F81BDB615E4109DBD2>.
- Martin J. Pickering and Matthew J. Traxler. Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 24(4):940–961, 1998.
- Christopher Potts. Is it possible for language models to achieve language understanding?, 2020. URL <https://chrispotts.medium.com/is-it-possible-for-language-models-to-achieve-language-understanding-81df45082ee2>.
- Christopher Potts, Daniel Lassiter, Roger Levy, and Michael C. Frank. Embedded Implicatures as Pragmatic Inferences under Compositional Lexical Uncertainty. *Journal of Semantics*, 33(4):755–802, 2016.
- Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernández Astudillo. Structural guidance for transformer language models. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*. Association for Computational Linguistics, 2021.
- Ciyang Qing, Noah D. Goodman, and Daniel Lassiter. A Rational Speech-Act Model of Projective Content. In *Proceedings of the Cognitive Science Society*, 2016. URL <http://cocolab.stanford.edu/papers/QingGoodmanLassiter2016-Cogsci.pdf>.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners, 2019. URL <https://d4mucfpxsywv.cloudfront.net/better-language-models/language-models.pdf>.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL <http://jmlr.org/papers/v21/20-074.html>.
- Rishi Rajalingham, Kailyn Schmidt, and James J. DiCarlo. Comparison of Object Recognition Behavior in Human and Monkey. *Journal of Neuroscience*, 35(35):12127–12136, 2015. ISSN 0270-6474. URL <https://www.jneurosci.org/content/35/35/12127>.
- Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018. ISSN 0270-6474. URL <https://www.jneurosci.org/content/38/33/7255>.
- Jonathan Rawski and Lucie Baumont. Modern Language Models Refute Nothing. 2023. URL <https://ling.auf.net/lingbuzz/007203>.

Tanya Reinhart. Definite NP anaphora and c-command domains. *Linguistic Inquiry*, 12(4):605–635, 1981.

Sophie Repp and Katharina Spalek. The Role of Alternatives in Language. *Frontiers in Communication*, 6:111, 2021. ISSN 2297-900X. URL <https://www.frontiersin.org/article/10.3389/fcomm.2021.682009>.

Stefan Riezler and John T. Maxwell. On Some Pitfalls in Automatic Evaluation and Significance Testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0908>.

Craige Roberts. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69, December 2012. URL <http://dx.doi.org/10.3765/sp.5.6>.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, January 2021. ISSN 2307-387X. URL https://doi.org/10.1162/tacl_a_00349.

Timothy T. Rogers and James L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach*. The MIT Press, June 2004. ISBN 978-0-262-28250-5. URL <https://doi.org/10.7551/mitpress/6161.001.0001>.

Eszter Ronai and Ming Xiang. Exploring the connection between Question Under Discussion and scalar diversity. In *Proceedings of the Linguistic Society of America*, volume 6, pages 649–662, 2021. URL <https://doi.org/10.3765/plsa.v6i1.5001>.

Eszter Ronai and Ming Xiang. Three factors in explaining scalar diversity. In *Proceedings of Sinn und Bedeutung 26*, 2022. URL https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/c/1271/files/2022/02/RonaiXiang_SuB26_paper.pdf.

Mats E. Rooth. *Association with Focus (Montague Grammar, semantics, only, even)*. PhD Thesis, University of Massachusetts, 1985. URL <https://scholarworks.umass.edu/dissertations/AAI8509599/>.

Richard Rorty. Wittgenstein, Heidegger, and the reification of language. In Charles Guignon, editor, *The Cambridge Companion to Heidegger*, Cambridge Companions to Philosophy, pages 337–357. Cambridge University Press, Cambridge, 1993. ISBN 978-1-139-00051-2. URL <https://www.cambridge.org/core/books/cambridge-companion-to-heidegger/wittgenstein-heidegger-and-the-reification-of-language/9678EE30FC5FD89BD254AAE72730D0BF>.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

John Robert Ross. *Constraints on Variables in Syntax*. PhD thesis, MIT, 1967.

Meredith L. Rowe. Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, 35(1):185–205, 2008. ISSN 0305-0009. URL <https://www.cambridge.org/core/article/childdirected-speech-relation-to-socioeconomic-status-knowledge-of-child-development-and-child-language/9A6AE54D0489EECB7F665B04DC61D365>.

Meredith L. Rowe. A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, 83(5):1762–1774, 2012. URL <https://srcd.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8624.2012.01805.x>.

- Paula Rubio-Fernández. Pragmatic markers: the missing link between language and Theory of Mind. *Synthese*, 199(1):1125–1158, December 2021. ISSN 1573-0964. URL <https://doi.org/10.1007/s11229-020-02768-z>.
- Paula Rubio-Fernández, Francis Mollica, Michelle Oraa Ali, and Edward Gibson. How do you know that? Automatic belief inferences in passing conversation. *Cognition*, 193:104011, December 2019. ISSN 0010-0277. URL <https://www.sciencedirect.com/science/article/pii/S0010027719301842>.
- Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. Large language models are not zero-shot communicators, 2022. URL <https://arxiv.org/abs/2210.14986>.
- David E. Rumelhart and James L. McClelland. On learning the past tenses of English verbs. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, volume 2, pages 216–271. MIT Press, 1986. URL <https://web.stanford.edu/~jlmcc/papers/PDP/Chapter18.pdf>.
- David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. URL <https://doi.org/10.1038/323533a0>.
- David E. Rumelhart, James L. McClelland, and PDP Research Group. *Parallel Distributed Processing*, volume 1. MIT Press, 1987.
- Erfan Sadeqi Azer, Daniel Khashabi, Ashish Sabharwal, and Dan Roth. Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5715–5725, Online, July 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-main.506>.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-1454>.
- Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.248>.
- Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439, December 2022. ISSN 2307-387X. URL https://doi.org/10.1162/tacl_a_00526.
- Ayse Pinar Saygin and Ilyas Cicekli. Pragmatics in human-computer conversations. *Journal of Pragmatics*, 34(3):227–258, March 2002. ISSN 0378-2166. URL <https://www.sciencedirect.com/science/article/pii/S0378216602800017>.
- Philippe Schlenker, Camille Coye, Maël Leroux, and Emmanuel Chemla. The ABC-D of Animal Linguistics: Are Syntax and Compositionality for Real? *Biological Reviews*, 2023. URL <https://ling.auf.net/lingbuzz/006962>.
- Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. URL <https://doi.org/10.1073/pnas.2105646118>.

- Sebastian Schuster, Yuxing Chen, and Judith Degen. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online, July 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.acl-main.479>.
- John R. Searle. Indirect Speech Acts. In *Speech Acts*, pages 59–82. Brill, Leiden, The Netherlands, December 1975. ISBN 978-90-04-36881-1. URL <https://brill.com/view/book/edcoll/9789004368811/BP000004.xml>.
- Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger P Levy. Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time, November 2022. URL psyarxiv.com/4hyna.
- Yikang Shen, Shawn Tan, Alessandro Sordoni, and Aaron Courville. Ordered neurons: Integrating tree structures into recurrent neural networks. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B1l6qiR5F7>.
- Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493, Denver, Colorado, May 2015. Association for Computational Linguistics. URL <https://aclanthology.org/N15-1051>.
- Les Sikos, Noortje J. Venhuizen, Heiner Drenhaus, and Matthew W. Crocker. Reevaluating pragmatic reasoning in language games. *PLOS ONE*, 16(3), March 2021. URL <https://doi.org/10.1371/journal.pone.0248388>.
- Herbert A. Simon. Cognitive science: The newest science of the artificial. *Cognitive Science*, 4(1):33–46, 1980. ISSN 0364-0213. URL <https://www.sciencedirect.com/science/article/pii/S0364021381800031>.
- B.F. Skinner. *Verbal Behavior*. Copley Publishing Group, 1957.
- Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302 – 319, 2013. ISSN 0010-0277. URL <http://www.sciencedirect.com/science/article/pii/S0010027713000413>.
- Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Wiley-Blackwell, 1986. ISBN 978-0-631-19878-9. URL https://monoskop.org/images/e/e6/Sperber_Dan_Wilson_Deirdre_Relevance_Communicat_and_Cognition_2nd_edition_1996.pdf.
- Nicola Spotorno, Eric Koun, Jérôme Prado, Jean-Baptiste Van Der Henst, and Ira A. Noveck. Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1):25–39, October 2012. ISSN 1053-8119. URL <https://www.sciencedirect.com/science/article/pii/S1053811912006611>.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci,

Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimeo Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdah Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millièvre, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Mishnerghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny,

- Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, 2022. URL <https://arxiv.org/abs/2206.04615>.
- Adrian Staub. The parser doesn't ignore intransitivity, after all. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 33(3):550–569, 2007.
- Mitchell Stern, Daniel Fried, and Dan Klein. Effective Inference for Generative Neural Parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1695–1700, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL <https://aclanthology.org/D17-1178>.
- Kevin Stowe, Prasetya Utama, and Iryna Gurevych. IMPLI: Investigating NLI Models' Performance on Figurative Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.ac1-long.369>.
- Laurie A Stowe. Parsing wh-constructions: Evidence for on-line gap location. *Language & Cognitive Processes*, 1(3):227–245, 1986.
- Patrick Sturt, Martin J. Pickering, and Matthew W. Crocker. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40:136–150, 1999.
- Chao Sun, Ye Tian, and Richard Breheny. A Link Between Local Enrichment and Scalar Diversity. *Frontiers in Psychology*, 9:2092, 2018. ISSN 1664-1078. URL <https://www.frontiersin.org/article/10.3389/fpsyg.2018.02092>.
- Toshitaka N. Suzuki and Klaus Zuberbühler. Animal syntax. *Current Biology*, 29(14):R669–R671, July 2019. ISSN 0960-9822. URL <https://doi.org/10.1016/j.cub.2019.05.045>. Publisher: Elsevier.
- Toshitaka N. Suzuki, David Wheatcroft, and Michael Griesser. Experimental evidence for compositional syntax in bird calls. *Nature Communications*, 7(1):10986, March 2016. ISSN 2041-1723. URL <https://doi.org/10.1038/ncomms10986>.
- Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan, June 2014. Association for Computational Linguistics. URL <https://aclanthology.org/W14-1601>.
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A Large Language Model for Science, 2022. URL <https://arxiv.org/abs/2211.09085>.
- Joshua B. Tenenbaum, Charles Kemp, Thomas L. Griffiths, and Noah D. Goodman. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, 331(6022):1279–1285, 2011. ISSN 0036-8075. URL <https://science.sciencemag.org/content/331/6022/1279>.
- Michael Henry Tessler and Michael Franke. Not unreasonable: Carving vague dimensions with contraries and contradictions. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2018. URL <https://cogsci.mindmodeling.org/2018/papers/0219/index.html>.

- Michael Tomasello. *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, 2003.
- Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.naacl-main.372>.
- Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu. A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 369–375, 2018.
- Simon W. Townsend, Sabrina Engesser, Sabine Stoll, Klaus Zuberbühler, and Balthasar Bickel. Compositionality in animals and humans. *PLOS Biology*, 16(8):e2006425, August 2018. URL <https://doi.org/10.1371/journal.pbio.2006425>.
- Anna Trosborg, editor. *Pragmatics across Languages and Cultures*. De Gruyter Mouton, 2010. ISBN 978-3-11-021444-4. URL <https://doi.org/10.1515/9783110214444>.
- John C. Trueswell, Michael K. Tanenhaus, and Susan M. Garnsey. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318, 1994.
- Tomer Ullman. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks, 2023. URL <https://arxiv.org/abs/2302.08399>.
- Emiel van Miltenburg. Detecting and ordering adjectival scalemates. In *Proceedings of MAPLEX*, Yamagata, Japan, 2015. URL <https://arxiv.org/abs/1504.08102>.
- Emiel van Miltenburg, Chris van der Lee, and Emiel Kraemer. Preregistering NLP research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 613–623, Online, June 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.naacl-main.51>.
- Marten van Schijndel and Tal Linzen. Modeling garden path effects without explicit hierarchical syntax. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, 2018.
- Marten van Schijndel and Tal Linzen. Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6):e12988, 2021. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12988>.
- Marten van Schijndel, Aaron Mueller, and Tal Linzen. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China, November 2019. Association for Computational Linguistics. URL <https://aclanthology.org/D19-1592>.
- Bob van Tiel, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. Scalar Diversity. *Journal of Semantics*, 33(1):137–175, 2016. ISSN 0167-5133. URL <https://doi.org/10.1093/jos/ffu017>.
- Shravan Vasishth, Sven Brüssow, Richard L Lewis, and Heiner Drenhaus. Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4):685–712, 2008.

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.
- Thomas C. Veatch. A theory of humor. *Humor*, 11(2):161–216, 1998. URL <https://doi.org/10.1515/humr.1998.11.2.161>.
- Corrie Vendetti, Deepthi Kamawar, and Katherine E. Andrews. Theory of mind and preschoolers’ understanding of misdeed and politeness lies. *Developmental Psychology*, 55(4):823–834, April 2019.
- Gabriella Vigliocco, Pamela Permiss, and David Vinson. Language as a multimodal phenomenon: implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130292, September 2014. URL <https://doi.org/10.1098/rstb.2013.0292>. Publisher: Royal Society.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL <https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf>.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019b. URL <https://openreview.net/forum?id=rJ4km2R5t7>.
- Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krma Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. SuperNaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.340>.
- Alex Warstadt and Samuel R. Bowman. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor & Francis, 2022. URL <https://arxiv.org/abs/2208.07998>.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. URL <https://aclanthology.org/Q19-1040>. Place: Cambridge, MA Publisher: MIT Press.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 2020a. URL https://doi.org/10.1162/tac1_a_00321. Publisher: MIT Press.
- Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online, November 2020b. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-main.16>.

- Albert Webson, Alyssa Marie Loo, Qinan Yu, and Ellie Pavlick. Are Language Models Worse than Humans at Following Prompts? It's Complicated, 2023. URL <https://arxiv.org/abs/2301.07085>. arXiv preprint.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*, 2022a. URL <https://openreview.net/forum?id=gEzrGCozdqR>.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2022b. URL <https://openreview.net/forum?id=yzkSU5zdwD>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain of Thought Prompting Elicits Reasoning in Large Language Models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022c. URL https://openreview.net/forum?id=_VjQlMeSB_J.
- Matthijs Westera and Gemma Boleda. A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung*, 24(2):439–454, September 2020. URL <https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/908>.
- Ethan Wilcox, Roger P. Levy, Takashi Morita, and Richard Futrell. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018. URL <https://www.aclweb.org/anthology/W18-5423>.
- Ethan Wilcox, Roger P. Levy, and Richard Futrell. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019a. URL <https://www.aclweb.org/anthology/W19-4819.pdf>.
- Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballestros, and Roger P. Levy. Structural supervision improves learning of non-local grammatical dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3302–3312, Minneapolis, Minnesota, 2019b. URL <https://www.aclweb.org/anthology/N19-1334>.
- Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Cognitive Science Society*, 2020. URL <https://arxiv.org/abs/2006.01912>.
- Ethan Wilcox, Pranali Vani, and Roger Levy. A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. In *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)*, pages 939–952, Online, August 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.acl-long.76>.
- Ethan Wilcox, Richard Futrell, and Roger Levy. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–88, October 2022a. ISSN 0024-3892. URL https://doi.org/10.1162/ling_a_00491.
- Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. Learning syntactic structures from string input. In Shalom Lappin and Jean-Philippe Bernardy, editors, *Algebraic Structures in Natural Language*. Taylor & Francis, 2022b. URL <https://lingbuzz.net/lingbuzz/007271>.

- Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018. URL <http://aclweb.org/anthology/N18-1101>.
- Deirdre Wilson and Dan Sperber. On verbal irony. *Lingua*, 87(1):53–76, June 1992. ISSN 0024-3841. URL <https://www.sciencedirect.com/science/article/pii/002438419290025E>.
- Deirdre Wilson and Dan Sperber. *Meaning and Relevance*. Cambridge University Press, 2012.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.emnlp-demos.6>.
- Daniel L K Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, March 2016. ISSN 1546-1726. URL <https://doi.org/10.1038/nn.4244>.
- Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. URL <https://doi.org/10.1073/pnas.1403112111>. Publisher: Proceedings of the National Academy of Sciences.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf>.
- Alexander Yeh. More accurate tests for the statistical significance of result differences. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, 2000. URL <https://aclanthology.org/C00-2137>.
- Victor H. Yngve. A Model and an Hypothesis for Language Structure. *Proceedings of the American Philosophical Society*, 104(5):444–466, 1960. ISSN 0003049X. URL <http://www.jstor.org/stable/985230>. Publisher: American Philosophical Society.
- Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. Talking with tact: Polite language as a balance between informativity and kindness. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2016. URL <https://cogsci.mindmodeling.org/2016/papers/0477/index.html>.
- Erica J. Yoon, Michael Henry Tessler, Noah D. Goodman, and Michael C. Frank. Polite Speech Emerges From Competing Social Goals. *Open Mind*, 4:71–87, November 2020. ISSN 2470-2986. URL https://doi.org/10.1162/opmi_a_00035.
- George Yule. *Pragmatics*. Oxford Introduction to Language Study. Oxford University Press, 1 edition, 1996. ISBN 978-0-19-437207-7.
- Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A Question Answering Benchmark for Artificial Social Intelligence. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8799–8809, 2019.

- Zheng Zhang, Leon Bergen, Alexander Paunov, Rachel Ryskin, and Edward Gibson. Scalar Implicature is Sensitive to Contextual Alternatives. *Cognitive Science*, 47(2):e13238, 2023. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.13238>.
- Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. GRICE: A Grammar-based Dataset for Recovering Implicature and Conversational Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online, August 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.findings-acl.182>.
- Chengxu Zhuang, Jonas Kubilius, Mitra JZ Hartmann, and Daniel L Yamins. Toward Goal-Driven Neural Network Models for the Rodent Whisker-Trigeminal System. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL <https://proceedings.neurips.cc/paper/2017/file/ab541d874c7bc19ab77642849e02b89f-Paper.pdf>.
- Davide Zoccolan, Nadja Oertelt, James J. DiCarlo, and David D. Cox. A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences*, 106(21):8748–8753, May 2009. URL <https://doi.org/10.1073/pnas.0811583106>. Publisher: Proceedings of the National Academy of Sciences.