HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Linguistics

have examined a dissertation entitled

The Rational Processing of Language Illusions

presented by Yuhan Zhang

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Kathryn Davidson, co-chair, Harvard University

Signature _____

Typed name: Prof. Edward Gibson, co-chair, MIT, Brain & Cognitive Science

Signature _____

Typed name: Prof. Kevin Ryan, Harvard University

Signature _____

Typed name: Prof. Roger Levy, MIT

Date: May 20, 2024

The Rational Processing of Language Illusions

A DISSERTATION PRESENTED
BY
YUHAN ZHANG
TO
THE DEPARTMENT OF LINGUISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
LINGUISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2024

Dissertation Advisor: Kathryn Davidson, Edward Gibson Yuhan Zhang

The Rational Processing of Language Illusions

Abstract

This thesis presents a unified rational account for three so-called language illusions in English. Language illusions refer to sentences that are ungrammatical, semantically ill-formed, or implausible but are still considered acceptable by native speakers of that language. They pose a great challenge to the scientific inquiry into the nature of language processing, especially in the understanding of why there is a mismatch between the actual comprehension and the literal form and meaning of the sentence. My thesis shows that this pattern of language comprehension manifests a unique trait of human beings when they deal with uncertainty.

The three language illusions are the depth-charge illusion (e.g., *No head injury is too trivial to be ignored*.), the comparative illusion (e.g., *More people have been to Russia than I have*.), and the negative polarity illusion (e.g., *Many authors that few critics recommended have ever received acknowledgment for a best-selling novel*). The problems for each of the language illusions can be traced back to meaning implausibility, semantic ill-formedness, and ungrammaticality, respectively. Numerous studies have investigated the effect of different linguistic factors on the degree of the illusion but none has provided a satisfactory explanation, let alone a unified one.

The unified account I attempt here argues that humans rationally infer the intended message behind the perceived linguistic signal, by weighing the prior probability of that intended message and how likely the message is to be encapsulated in the perceived signal. The

inference could cover the entire sentence or only part of it. This theory originates from the noisy-channel theory of language processing (e.g., Futrell et al., 2020; Gibson et al., 2013a; Levy, 2008) under the information theory (Shannon, 1948). Ultimately, this account exemplifies rationality in language comprehension. By synthesizing the literature on rationality, I summarize the rational behavior of language processing as a set of behaviors that optimize the information gain from perceived sentences by applying the Bayes rule to deal with uncertainty and making trade-offs between information accuracy and cognitive load. I provide empirical data from human experiments as support.

I also investigate whether large language models can be "tricked" by these illusions, as a starting point to investigate whether their processing patterns exhibit traits of rationality.

Contents

Acknowledgments

Throughout graduate school, I have been inspired by many great minds that shaped who I am at the end of the journey. With respect to this dissertation, I want to thank my wonderful dissertation committee members: Kathryn Davidson, Edward Gibson, Kevin Ryan, and Roger Levy. Special thanks to Kate, who has unwaveringly encouraged me to take on an experimental and psycholinguistic venture and always trusted my natural speed of maturation in academic rigor. She has been an amazing role model both in work and in life. To Ted, who generously welcomed me as part of his lab, taught me a tremendous amount of how to do good science in the intersection of linguistics and cognitive science, and cultivated my optimism and resilience towards intellectual challenges and opportunities. The rest of the committee have been equally important: thanks to Kevin for the pinpointed comments on the dissertation and for being open-minded to lend expertise on a topic outside phonology, and to Roger for teaching me to hold the highest standard and always push for a deeper understanding of the bigger picture. I feel extremely fortunate to be influenced by these outstanding scholars I admire and aspire to emulate in the future.

Several individuals have made crucial contributions to the dissertation that I like to call out. For Chapter 2, I thank Rachel Ryskin for teaching me how to rigorously design well-controlled experiments, do statistics, and reason with data. For Chapter 3, I am grateful to

Carina Kauf for sharing exciting intuitions and annotation results and to Roger Levy for helping me advance the argument to the next level. For Chapter 5, I am indebted to Forrest Davis for introducing me to the fascinating field of evaluating large language models with linguistic knowledge.

Among the academic communities that I have the privilege to be part of in graduate school, I am thankful to the faculty members in the Department of Linguistics at Harvard University. Thanks to Gennaro Chierchia for being on my first generals paper committee and teaching me the beauty and elegance of formal semantics. Thanks to Jonathan Bobaljik for planting in me the seed of independent spirit of scholarship. Thanks to *Huáng lǎoshī* ("Huang teacher", Professor Huang, i.e., C.-T.James Huang) for guiding me to lay the foundation of syntax. Thanks to Jesse Snedeker for the intellectually stimulating conversations and to Stuart Shieber for the chance to study computational linguistics. Thanks to Helen Lewis for her generous administrative support throughout the six years. I have had great pleasure to grow with my fellow graduate colleagues. Special thanks go to Yujing Huang, Dorothy Ahn, Yingtong Liu, and Giuseppe Ricciardi for being my role model and my go-to place for encouragement at the early stage of my graduate career; and to Hayley Ross, Ankana Saha, Hande Sevgi, Jack Rabinovitch, and Natasha Thalluri for the precious camaraderie; and to Anthony Yacovone, Maggie Kandel, Simge Topaloglu, and Briony Waite for sharing interesting insights after attending the LangCog workshop at William James Hall.

Then I can take the T to Kendall Square to thank my friends, colleagues, and mentors at MIT. In addition to Ted, Roger and Ev, I am grateful for every interaction I have had with scientists who study language from a wide range of perspectives. Each interaction gets me closer to becoming a versatile language scientist. I am indebted to the members at

TEvLab: Mika Braginsky, Sammy Floyd, Aixiu An, Saima Malik-Moraleda, Eric Martinez, Sihan Chen, Thomas Hikaru Clark, Moshe Poliak, Cory Shain, Anna Ivanova, Chengxu Zhuang, Maria Ryskina, Greta Tuckute, Hope Kean, Carina Kauf, Ben Lipkin, Colton Casto, Niharika Jhingan, Aalok Sathe, and Sara Swords. I am also grateful for the communications and inspiration from Jennifer Hu, Peng Qian, Ethan Wilcox, Jiayuan Mao, and Freda Haoyue Shi. The breadth and depth of research involving language are so eye-opening that I have no doubt that language will be the most mesmerizing subject of study in the 21st century.

Next, I want to thank my mentors at Apple during my internship in 2022. Thank you to Hadas Kotek, Zidi Xiu, and David Q. Sun for their guidance and expertise in language technology. Then I can take the time machine to thank my mentors without whom I cannot make it to this point in my education. They are Roger L. Oleson, my college writing teacher, Xiaolu Yang and Peng Zhou, my psycholinguistics professors at Tsinghua University, Yihong Gao, my sociolinguistics professor at Peking University, Dan Parker, my syntax professor at College of William & Mary, Rose Hendricks, my research mentor at University of California, San Diego, and Nick Huang and Chia-Hsuan Liao, my research mentors at the University of Maryland, College Park. I show my sincere gratitude for their generous support and encouragement along my journey. Thank you for seeing the potential in me.

Besides the scholarly communities, I sincerely acknowledge the funding sources of my research. Thanks to the Department of Linguistics, the Institute for Quantitative Social Science, and the Kenneth C. Griffin Graduate School of Arts and Sciences at Harvard University, as well as the Human Sentence Processing Conference 2024 for supporting my research, travel, and living expenses.

I also want to thank my dear friends who are with me through thick and thin. They teach me the essence of life. I cannot enumerate the names here but among them, I owe a debt of gratitude to Yilan, Yulu, Lechen, Shuyu, Chris, Yuhe, Zhe, Xiaorui, Jinlin, Percy, Yu Hao, Yutong, Yaowei, Yijun, Guoli, Yanke, Kuanhao, Dawei, Zhutian, Harry, Beichen, and Ruihan. Thank you all for your friendship. Most importantly, I am immensely grateful to my parents for their unwavering love, encouragement, and sacrifices throughout my doctoral journey. Their support has been my rock. Finally, I would like to thank my boyfriend Pengzhi for his trust and company which I hold dear to my heart. His tenacity and optimism are a consistent long-lasting force that inspires me. I am deeply grateful to have him by my side.

This thesis is dedicated to my parents.

This dissertation has generated the following publications and manuscripts:

Zhang, Y., Ryskin, R., & Gibson, E. (2023). A noisy-channel approach to depth-charge illusions. *Cognition*, 232, 105346. https://doi.org/10.1016/j.cognition.2022.105346

Zhang, Y., Gibson, E., & Davis, F. (2023). Can language models be tricked by language illusions? Easier with syntax, harder with semantics. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 1–14, Singapore. Association for Computational Linguistics.

Zhang, Y., Kauf, C., Levy, R., & Gibson, E. (under review). Comparative illusions are evidence of rational inference in language comprehension.

Zhang, Y. & Gibson, E. (in prep). A memory-based account for a robust negative polarity illusion effect.

Listing of figures

1

Introduction

Language is a unique creation of human beings through which one encodes and decodes information about the world and through which human beings are connected across time and space. Given that language is powerful but ambiguous and ever-changing and that communication takes place in a fleeting nature, why human beings are so proficient at capturing the meaning in complicated and (sometimes) distorted sentences? As a matter of fact, are

human beings always proficient in deriving the information encoded in language?

The goal of this thesis is to introduce a series of very interesting linguistic phenomena that suggest that sometimes during language comprehension, the message humans derive from the perceived linguistic input does not match its literal meaning or form. This work also provides a processing account for these linguistic pieces whose processing mechanisms have been poorly understood. They are called **Language Illusions** because they are ungrammatical, semantically ill-formed, or implausible but are still deemed acceptable by native speakers of the relevant language.* In a world where it is assumed that a human being with typical linguistic competence would have no trouble distinguishing sentences that conform with the grammar of the language from sentences that fail to do so, language illusions question why there is a mismatch between linguistic competence versus performance (Chomsky, 1965). It also enriches the empirical landscape of "shallow processing" (e.g., Sanford & Sturt, 2002) and rational inference during language processing (e.g., Levy, 2008).

The three language illusions investigated in this thesis are shown in (1):

(1) a. No head injury is too trivial to be ignored. (Depth-charge illusion)

b. More people have been to Russia than I have. (Comparative illusion)

c. Many authors that few critics recommended have ever received a best-selling award. (Negative polarity illusion)

The first one is called the depth-charge illusion (Wason & Reich, 1979), where the literal and compositional reading of the canonical sentence (1a) is "no matter how trivial head injuries are, we should **ignore** them." Yet, the message comprehenders take away from this sentence is the opposite, as "we should not ignore the head injury". The depth-charge illusion

*As we see later in this work, the degree of illusion differs by the type of phenomenon.

questions how comprehenders integrate their world knowledge into language processing and what role the individual components – *no*, *too...to*, *trivial*, *ignore* – plays in understanding the sentence.

The second one is called the comparative illusion (Montalbetti, 1984), where the canonical sentence (1b) is deemed acceptable by native speakers of English at first sight even though it is semantically ill-formed (see detailed explanations in Section 3). The comparative illusion asks why comprehenders are blown away by the superficial well-formedness and probes whether the structure and the usage of *more* can modulate the acceptability of the sentence.

The third one is called the negative polarity illusion (Drenhaus et al., 2005; Vasishth et al., 2008), which taps the structural relationship between the negative polarity item (NPI) *ever* and its licensor. For the specific case in (1c), the English grammar dictates that the NPI *ever* should occur in the scope of *few* but it does not. It remains unknown why comprehenders still sometimes accept this sentence and whether cognitive load could be a contributing factor.

In this thesis, I show that there is a unified way to understand why the illusion effect occurs in the first place. I argue that comprehenders make rational inferences over the depth-charge sentence like (1a) and the comparative illusion sentence like (1b) and derive plausible interpretations by considering the most possible message intended by the literal sentence. The assumption I take is that during language comprehension, listeners or readers entertain more than one interpretations or sentence strings, given the perceived sentence. The entertained messages differ from the perceived one by only a few edits in words or phrases. Comprehenders then choose the most probable interpretation by weighing the prior probability of that interpretation against how likely that interpretation is to be encapsulated in the per-

ceived string during information transmission. The unique property of this argument is that it assumes a probabilistic nature of language processing, and that there could be a mismatch between the literal meaning of the perceived sentence and the derived mental representation of the sentence in the comprehender's mind. This mismatch is essentially uncertainty in language comprehension, and could be exemplified as superficial string edits or structural alternations. The assumption is also true for the negative polarity illusion. There, I show that the illusion arises because comprehenders store a distorted representation of the context that precedes the NPI. For example, for the perceived sentence (1c), comprehenders might interpret it as (2) where *few* and *many* are exchanged due to their similarity in meaning and the adjacency in the position. I further assume that it is the limited memory resources comprehenders allocate to language processing that cause such memory distortion. Throughout my argument, the notions of noise and distortion are salient and lead to Section 1.2 where I introduce the major theoretic framework I adopt for this study. Before that, I will introduce what accounts as a good theory of langugae processing in Section 1.1.

(2) Few authors that many critics recommended have ever received a best-selling award.

1.1 A Good Theory of Language Processing

A central question in the intersection of linguistics and cognitive science is understanding how human beings process language and learn about the world through information extracted from language. What cognitive constraints shape the procedure? How do linguistic knowledge and the cognitive components interact with each other in language processing? What computational or algorithmic level of language processing can we adopt to predict the novel performance of human language processing (Marr, 1982)?

The inquiry and challenge of defining a cohesive language processing theory are accompanied by three pieces of fundamental truth. First, language as a symbolic system is highly ambiguous, both in terms of how one symbol refers to objects or events and how symbols combine to carry out more complicated meaning. For example, a lexical ambiguity for the word *bank** in (3a) cannot be resolved until the following context about a financial activity is revealed. A syntactic ambiguity is shown in (3b) where the prepositional phrase *in the jungle* can postpositionally modify both the subject *Penelope* or the object *an elephant* (without changing too much of the meaning, fortunately). Yet another prepositional phrase *in her pajamas* can modify *an elephant* theoretically but is dispreferred by the weird compositional meaning.† World knowledge resolves the ambiguity in this case. A semantic ambiguity is shown in (3c) where the interpretation could be (i) there exists a specific prince that Aurora wants to marry (even though she doesn't know that person is a prince) or (ii) Aurora wants to marry whoever is a prince. The ultimate interpretation is determined by how one deals with the referential property of *a prince*. How do grammatical rules of language, the distributions of subcomponents of sentences, one's linguistic experience, and the broader context work together to compute the interpretation of one ambiguous sentence?

(3) a. I walked past the bank ... but forgot to deposit my salary.

b. Penelope saw an elephant in the jungle/her pajamas.

c. Aurora wants to marry a prince.

Second, it is not always easy to process a sentence due to the constraints of one's limited non-linguistic cognitive resources (e.g., "memory limitations, distractions, shifts of attention

*"Bank" as the financial institution or "bank" as the river bank.

†There is an elephant in Penelope's pajamas and she saw it.

and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance" (Chomsky, 1965, p.1)). People encounter various types of difficulty while processing language. Some might originate from the syntactic complexity, and other might be due to a semantic anomaly. While this deals with the "performance" factor out of the "competence versus performance" debate raised by Chomsky (1965),* a more interesting question is whether we can predict processing difficulty with computational models that contribute to building a cognitively inspired computational program for human language processing (Gibson, 1991). For example, the dependency locality effect deals with the processing difficulty due to the integration cost of words that are far apart in a linear order (Gibson, 1998, 2000). The example in (4) shows the processing difficulty of *out* in (4d) when it needs to be integrated with *throw* that is far away and interrupted by the long object (Lohse et al., 2004).

(4) a. Bob threw out the trash.

b. Bob threw the trash out.

c. Bob threw out the old trash that had been sitting in the kitchen for several days.

d. Bob threw the old trash that had been sitting in the kitchen for several days out.

(Futrell et al., 2020, p.6)

Third, language processing usually takes place not in quiet and well-controlled laboratories but in dynamic, unpredictable, multi-party, and noisy environments (e.g., Shannon, 1948). Under what circumstances people produce speech errors or misunderstand the speaker has brought intellectual amusement and provided naturalist examples to study

*Competence refers to "the speaker-hearer's knowledge of his language" and performance refers to "the actual use of language in concrete situations" (Chomsky, 1965, p. 4).

language processing. In the famous Mondegreen case, Sylvia Wright, an American writer, recalled how she misheard the last line of an excerpt from the seventeenth-century Scottish ballad *The Bonnie Earl o' Moray* (5) that her mother read aloud to her (Wright, 1954). Wright thought the last line was *And the Lady Mondegreen* while the literal line was *And they've laid him on the green*. This case indicates that people sometimes substitute the perceived linguistic input with alternatives that share similar perceptual features and make sense in the context.

(5) Ye Highlands and Ye Lowlands
Oh where hae you been?
They hae slain the Earl of Murray,
And the Lady Mondegreen.
(And they've laid him on the green.)

From the above-mentioned examples, it is not hard to tell that it is not straightforward to come up with a good theory of language processing that predicts every aspect of the behavioral (or even neural) patterns of language processing. On the other hand, a good theory of language processing should at least incorporate considerations about the inherent properties of language, the constraints of the cognitive system, and the dynamics of the environment as well as contexts. The case of language illusion uniquely deals with the second and third pieces of truth: Given limited working memory and other cognitive resources, how do comprehenders process linguistic input that contains errors?

1.2 The Rational Aspect and the Noisy-Channel Theory

In this thesis, I argue that the rational aspect of language processing has the potential to become a good theory that could explain how human beings extract information from language given its intrinsic encoding properties (e.g., ambiguity), the cognitive constraints of the participating agent (e.g., working memory limitation, production and comprehension errors), and the environment conditions (e.g., noises in the environment). The definition of *rational* in this context originates from Herbert A. Simon (1955; 1972) and is developed through later theories in cognitive science such as the rational analysis (Anderson, 1990) and the resource-rational analysis (e.g., Lieder & Griffiths, 2020). This definition has also been widely used in economics (Mankiw, 2018). I sketch critical components of the rational aspect of language processing as follows:

- **Goal**: The agent/listener/comprehender engaged in language processing has the goal to accurately understand the information encoded in sentences and maximize that accuracy.
- **Environmental conditions**: Language processing in the format of communication or reading a passage takes place quickly and often occurs in a noisy environment. The external time pressure and noise can affect the communication/language processing quality.
- **Cognitive constraints**: Any agent involved in language processing is constrained by her cognitive constraints manifested as memory limitations, production errors, or comprehension errors, etc.
- **Nature of the problem**: The problem to solve is to decode a message from a string of

linguistic symbols or encode the message into the symbols. Language is a complicated symbolic system that encodes an infinite number of meanings with finite symbols. The solution to solve the meaning has to conform with the syntactic and combinatorial rules of encoding and decoding, out of many possible solutions that exemplify the ambiguity of language.

- **Rational behavior of language processing**: With the goal and the aforementioned constraints, a rational language processor aims to maximize the information gain with reasonable costs from memory and energy. She would make use of all available information sources to the fullest to process a sentence (Levy, 2008), optimize her own time, spend only as much time as needed on any linguistic task (Gibson, 2024), and make Bayesian statistical inference over uncertain linguistic input (Gibson et al., 2013a). Sometimes, comprehenders have to make trade-offs and settle with "satisficing" solutions * in exchange for less effort (Hahn et al., 2022).

Under this rational framework, the noisy-channel theory of language processing is capable of explaining how humans comprehend linguistic materials that contain errors, and more generally, how comprehenders deal with uncertainty during language processing. The noisy-channel framework of language processing comes from information theory (Shannon, 1948) and captures how communication is conducted under a noisy channel. It regards

*The term *satisficing* comes from Simon (1972) and can be taken as synonymous with *good enough*. In psycholinguistics, there is also a common term called "good enough" which argues that comprehenders adopt *heuristics* to achieve shallow but incorrect understanding of sentences (e.g., Ferreira et al., 2002; Ferreira & Patson, 2007). One example is *While the woman bathed the baby cried*, where comprehenders wrongly believe that the baby was bathed by the woman, while the literal meaning of the sentence does not say so (Ferreira et al., 2002). I leave it to future work to summarize the similarities and differences between the good-enough account of language comprehension and the rational inference approach of language processing. Please see recent discussion in Traxler (2014).

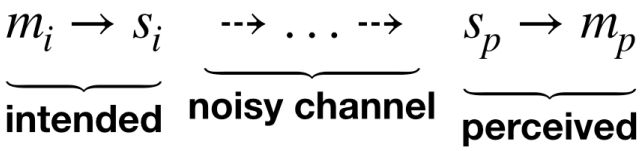


Figure 1.1: Schematic representation of the noisy channel model

language comprehension as a rational Bayesian inference process over noisy input by assuming that during communication, the linguistic input people encounter is not always perfect (Gibson et al., 2013a; Levy, 2008, 2011; Levy et al., 2009; Ryskin et al., 2018, 2021b). Comprehenders assign an interpretation to a perceived sentence by weighing the probabilities of various possible interpretations against the likelihood that the intended interpretation is to be encapsulated into the perceived string coupled with noise corruptions during noisy information transmission. The mechanism is illustrated in Figure 1.1, following Shannon (1948). The speaker initiates an intended message m_i to convey, and m_i is linguistically encoded as the intended sentence s_i, through various modalities such as speaking, writing or signing. During information transmission, the sentence s_i could be corrupted due to producer or comprehender errors, or noises from a noisy environment. Some words could be deleted, inserted, exchanged or substituted by other words. Subsequently, the signal s_p perceived by the listener or the reader could differ from s_i. The meaning m_p derived from s_p could thus deviate from m_i, and result in misunderstanding or misrepresentation of the original message during communication.

Yet, not all language processing sticks to a literal parse of the perceived signal. Comprehenders are shown to leverage Bayesian reasoning and make rational inferences on the possible interpretation of the corrupted sentence. The process can be modeled in Equation 1.1:

$$P(s_i|s_p) \propto P(s_i)P(s_p|s_i) \tag{1.1}$$

In Equation 1.1, s_p represents the signal received by the comprehender; s_i represents both the meaning and the sentence intended by the producer, as well as one of the possible hypotheses of interpretation entertained by the comprehender. This equation is a simplification of Figure 1.1 in that we take s_i in Equation 1.1 to represent both m_i and s_i in Figure 1.1. On the left-hand side, $P(s_i|s_p)$ represents the posterior probability that the comprehender assigns interpretation s_i to the perceived sentence s_p. This is proportional to, on the right-hand side of the equation, the product of prior probability $P(s_i)$ that the producer intends to utter s_i, and the noise likelihood term $P(s_p|s_i)$ which represents how likely the intended sentence s_i is corrupted to s_p during noisy information transmission.

The noisy-channel approach has received extensive empirical support from diverse materials and languages. In Gibson et al. (2013a), a wide variety of English constructions have been tested (e.g., active/passive constructions, double object constructions). They show that participants are likely to interpret an implausible sentence according to its plausible alternative if they can easily infer the plausible intended meaning by postulating a simple edit. For example, when people hear or read *The mother gave the candle the daughter*, they might infer that the intended message should be "The mother gave the candle to the daughter" where there is only one word deletion of *to* from the intended *The mother gave the candle to the daughter* to the perceived *The mother gave the candle the daughter*.

The noisy-channel theory of language processing makes several quantitative predictions which have been supported by empirical evidence. First, the rate of nonliteral inference of implausible sentences, $P(s_i|s_p)$, is proportional to the prior probability of the intended mes-

sage, $P(s_i)$, which can be represented both by the meaning consistency with world knowledge and by the linguistic structural frequency of the intended sentence. As shown in chapter 2, which studies the depth-charge sentence *No head injury is too trivial to be ignored*, when participants held a common attitude toward the topic under discussion (e.g., head injuries should not be ignored in general), they were more likely to assign a high plausibility score to that implausible illusion sentence, compared with when the topic is controversial (e.g., artificial ingredients should, in general, be abandoned in sentence *No artificial ingredients are too harmless to be abandoned*). In terms of the structural prior, it has been found that sentences with a low-frequency structure – e.g., the OSV word order versus the canonical SVO word order in Mandarin (Liu et al., 2020) and the OVS word order versus the canonical SVO word order in Russian (Poliak et al., 2023) – were more likely to induce nonliteral inference because the structure in the nonliteral message has a higher prior probability (e.g., *The boy, the trash threw* interpreted as "the boy threw the trash" in Mandarin). The priors can also be modulated by contextual factors. As shown in Chen et al. (2023), supportive contexts preceding implausible sentences can increase its nonliteral inference rate. For instance, given the implausible sentence *The girl tossed the boy to the apple*, if it appears in the context "The boy and the girl went apple picking together. The girl picked an apple that the boy wanted.", the participants were more likely to nonliterally infer the meaning of this implausible sentence as "The girl tossed the boy the apple". The additional semantic information (e.g., that the boy wanted an apple that the girl had) can increase the prior probability of the intended meaning (i.e., it is highly likely that the girl would toss the boy the apple).

Second, the nonliteral inference rate, $P(s_i|s_p)$, is also proportional to the noise likelihood

term, $P(s_p|s_i)$, which measures how likely the intended sentence is surfaced as the perceived sentence through noise corruption. In daily communication, the errors would be realized as production errors, comprehension errors, or errors caused by noise transmission in the environment. The kinds of noise involve the deletion, insertion, substitution, and exchanges of parts of the strings. Noise edits could happen for both function words and content words (chapter 2); fewer edits are more likely and easier to cause rational inference than more edits (Gibson et al., 2013a); assuming one error, deletion is found to be more likely than those with insertion (Gibson et al., 2013a; Poliak et al., 2023; Zhang et al., 2023c), which also follows the Bayesian size principle (MacKay, 2003; Xu & Tenenbaum, 2007), because a deletion requires a particular word to be randomly chosen from the intended sentence but an insertion requires the selection from the producer's vocabulary. The noise edits could operate on character or subword level (e.g., suffixes for subject-verb agreement in Ryskin et al. (2021a) and Poliak et al. (2023)), word level (e.g., *to* deletion in *The mother gave the daughter to the candle* in Gibson et al. (2013a)), or structural level (chapter 2). Furthermore, comprehenders also actively adapt to the noises in the environment. The rational inference of noisy sentences increases when the environment contains more grammatical errors (Gibson et al., 2013a) and the rational inference of sentences with a specific type of error (e.g., deletion) also tracks the presence of that specific type (Ryskin et al., 2018).

The noisy-channel hypothesis of language processing has also accumulated empirical support from a wide range of phenomena. It can be seen to explain reading behaviors in aphasia (Fedorenko et al., 2022; Gibson et al., 2016), cross-linguistic comprehension of implausible sentences (e.g., Poliak et al., 2023; Zhan et al., 2023), and bilingualism (Gibson et al., 2017).

1.3 A Computational Angle

A good theory of language processing not only reveals the fascinating nature of language but also benefits the research of natural language processing in computer science, especially when this more applied endeavor aims to create psychologically plausible algorithms that can tackle various issues of interpreting natural language (e.g., Gibson, 1991; Demberg et al., 2013), such as ambiguity resolution, entity tracking, and finding the most efficient algorithm of computing a sentence meaning. In fact, at the time of the dissertation, no information technology innovation has been more influential and controversial than the current strides of (large) language models (e.g., BERT (Devlin et al., 2019), the GPT series (e.g., Radford et al., 2018, 2019; Brown et al., 2020)) that have demonstrated great capacities to generate coherent linguistic messages. A good theory of language processing from the human side can inform the evaluation of these heavily invested and giant algorithms of natural language processing (Hu et al., 2020; Linzen et al., 2016; McCoy et al., 2019; Kim & Linzen, 2020; Wilcox et al., 2018; Warstadt et al., 2020). The perspectives from cognitive science and linguistics propel the inquiry of whether these computational algorithms are impressive machines to handle complex text, or they are capable of processing information just like a human.

The case of language illusions also opens up an excellent testing ground to investigate the nature and our beliefs about large language models. Specifically, do large language models process illusion sentences, as those in (1), in a human-like manner where the anomalous sentences receive high probability? Or, on the other hand, do large language models "recognize" the anomaly and act like a literal grammarian? If the answer to the first question is yes and human-like, we can further ask whether LLMs exhibit rational behaviors of language pro-

cessing as humans. This answer could add to the discussion of whether we should expect a human-like artificial intelligence system, or a system that surpasses human capacity.

1.4 Dissertation Structure

In this dissertation, Chapters 2 to 4 address each of the three language illusions as introduced in (1) in the framework of the noisy-channel theory of language processing. For the depth-charge illusion (1a) in Chapter 2, I will show, through four experiments, that the probability of an illusive interpretation is modulated by (i) how consistent that illusive interpretation is with commonly-held world knowledge and (ii) how likely that interpretation is encapsulated into the perceived illusion sentence with considerations from structural substitutions. For the comparative illusion (1b) in Chapter 3, the approach is similar to the depth-charge illusion. I claim that given the multiple possible plausible interpretations of the illusion sentence, the more structurally similar the interpretation is to the perceived anomalous string, the easier it is for comprehenders to reach that interpretation.

In Chapter 4, the argument is related to online rational processing, not the same as rational inference in Chapters 2 and 3. For the negative polarity illusion sentence (1c), the memory effect plays a role in comprehenders not recognizing the dependency anomaly between the NPI and the inaccessible NPI licensor. Through four experiments, I will show that the illusion arises from a distorted memory representation of the position of the NPI licensor *few* and the quantifier *many*, due to their similarity in meaning and positional adjacency.

Overall, these human behavioral patterns present a picture of the rational processing of language illusions. Given the limited cognitive resources allocated in processing each sentence and the complex nature of the sentences, comprehenders do not commit themselves to

the meticulous incremental processing endeavor. Instead, they leverage their world knowledge about the topic under discussion, and their linguistic knowledge about the statistical distributions of critical words and constituents in the sentence to achieve an interpretation with a minor cost of processing accuracy. This echoes with the essence of "being rational" – maximizes the gain with restricted efforts in everyday life and during the decision-making process.

In Chapter 5, I introduce the computational evaluation work that investigates whether the processing patterns of large language models exhibit rational characteristics as humans. I present inconclusive remarks generated by inconsistent results, and share perspectives about future work.

2

Depth-charge Illusion

On the surface, sentence (6) seems like a reasonable thing to say:

(6) No head injury is too trivial to be ignored.

This sentence seems to mean something like "no matter how trivial head injuries are, they should not be ignored". However, the literal (compositional) meaning of (6) actually implies the opposite: "all head injuries should be ignored however trivial". The clash be-

tween the literal meaning of the expression and its intended meaning often goes unnoticed. Even if it is pointed out by someone else, it often takes time for the reader to come to the realization that there is a discrepancy between the common takeaway and the literal meaning of the sentence. These sentences have been referred to as "depth-charge" sentences possibly because the experience of processing them is analogous to the explosion of depth-charge bombs. Just as the bombs explode after traveling in the water for a certain amount of time (e.g., Sanford & Emmott, 2012, p.28), it takes time and effort for the comprehender to realize the anomaly underlying sentence (6).

The depth-charge interpretation was observed more than four decades ago (Wason & Reich, 1979) and its source still remains a puzzle. Depth-charge illusions are different from other well-known semantic illusions, such as the Moses Illusion, where people are asked *How many of each type of animal did Moses take on the ark*? (Erickson & Mattson, 1981), or *When an airplane crashes, where should the survivors be buried*? (Barton & Sanford, 1993). To detect the semantic anomaly requires an effective integration of linguistic and world knowledge. These semantic illusions are easier to detect especially when the illusory words appear in the focus position of the sentence (Wang et al., 2009), when the sentence changes from a question to a declarative statement (Büttner, 2007), and when the errors are related to expert knowledge of the participants (Cantor & Marsh, 2017). In contrast, the depth-charge illusion is hard to detect and sometimes the non-literal meaning ends up entrenched in comprehenders' final interpretation, even after explicit instruction of how to interpret structures like *X is too Y to Z* (Giannouli, 2016; Kizach et al., 2016; Natsopoulos, 1985; O'Connor, 2015, 2017; Paape et al., 2020; Paape, 2024). As yet, there has not been a satisfactory explanation of how depth-charge illusions are understood, but researchers have

investigated several potentially relevant factors. In what follows, we first review these factors. We then follow Gibson & Thomas (1999), Vasishth et al. (2010), and Futrell et al. (2020) in hypothesizing that illusions like these might be revealing about how human language is processed, in general. Futrell et al. examined cases of the missing-verb-phrase illusion, as in (7) (example from Frazier (1985), reporting an intuition from Janet Fodor):

(7) a. Missing verb phrase (ungrammatical):
The apartment that the maid who the cleaning service sent over was well-decorated.

b. All three verb phrases (grammatical):
The apartment that the maid who the cleaning service sent over cleaned was well-decorated.

People often find (7a) more acceptable than (7b) (Gibson & Thomas, 1999) in spite of the fact that (7a) is actually ungrammatical: (7a) is missing a verb to connect the second noun phrase "the maid" to the first noun phrase "the apartment". Futrell et al. (2020) suggest that this grammaticality illusion – sentences that fail the syntactic compositional rule – reflects a general property of language processing. Under their proposal, people cannot remember exactly what was produced earlier: they are always forgetting the literal details of what was produced, over time. Futrell et al. propose a forgetting and context-reconstruction model – the lossy-context surprisal proposal – such that people try to reconstruct (7a) as a grammatical variant, as in (8) (see Hahn et al., 2022, for an even further updated approach):

(8) The apartment with the maid that the cleaning service sent over was well-decorated.

Following Futrell et al. (2020) and Hahn et al. (2022), we hypothesize that illusions like

(6) might inform us about how language structure interacts with language processing. In particular, we propose an explanation for depth-charge illusions within the noisy-channel framework for language processing (Gibson et al., 2013a; Levy, 2008; Shannon, 1948). We provide support for this account in four experiments. Experiment 1 replicates the depth-charge effect in Paape et al. (2020). Experiment 2 shows that the more consistent the intended meaning of the depth-charge material is with our commonly held world knowledge, the more likely that misinterpretation will occur. Experiment 3 suggests that depth-charge sentences may be viewed as the result of language production errors where the intended utterance is close to the depth-charge sentence. Experiment 4 shows that the probability of interpreting depth-charge sentences in terms of their non-literal meaning is predicted by the likelihood of the production error which could have corrupted a plausible sentence into the implausible depth-charge sentence. Overall, these findings suggest a noisy-channel explanation for the depth-charge illusion: the probability of reaching a non-literal interpretation of depth-charge sentences is correlated with (i) the prior probability of this non-literal meaning and (ii) the likelihood that captures how likely depth-charge sentences are to result from "noise" during information transmission.

2.1 Current Landscape

2.1.1 World Knowledge

Several researchers have suggested that the most available interpretation of depth-charge sentences is often consistent with world knowledge but opposite to the literal interpretation (Paape et al., 2020; Sanford & Sturt, 2002; Wason & Reich, 1979). Despite the intuitive appeal of this claim, the experimental findings have been mixed. In Wason and Reich's view,

the more implausible the verb-noun phrase combination in depth-charge sentences is, the more likely that misinterpretation will occur. For example, *No head injury is too trivial to be ignored* includes the implausible verb-noun-phrase "ignore head injuries" and often results in an illusion, but *No missile is too small to be banned* includes the plausible "ban the missile" and seldom triggers the illusion. Their empirical test supports this view but provides little information about the relationship between world knowledge and misinterpretation rate due to a confound: items with a plausible verb-noun phrase were also always plausible sentences whereas items with an implausible verb-noun phrase were all implausible.

In a later study, Natsopoulos (1985) measured whether participants "hold beliefs or attitudes toward the topic expressed" (p.388) by the target depth-charge sentence on a 6-point Likert scale. He later asked a different group of participants to paraphrase the same target sentences. He found no correlation between the world knowledge rating score and the paraphrase accuracy rate, but it is worth noting that there were only eight items, so there were perhaps not enough items to find an effect if there was one.

More recently, O'Connor (2015; 2017) measured the consistency between paraphrases related to depth-charge sentences and world knowledge. The paraphrases varied in their sentential quantifiers and in the polarity of their final verbs. For example, given the depth-charge sentence *No social program is too wasteful to oppose*, the paraphrases were (i) "all social programs should be opposed", (ii) "no social programs should be opposed", and (iii) "all social programs should be supported". While O'Connor found that the consistency score of the second type of paraphrase positively correlated with the misinterpretation rate, the other two were not when all three metrics were entered as fixed effects in a single regression model. It is possible that the lack of significant effects of paraphrases (i) and (iii) could have resulted

from collinear relationships among the three paraphrases (e.g., Allen, 2004).

Lastly, Paape and colleagues (2020, Experiment 2B) provide the most convincing evidence to date for a relationship between world knowledge and the misinterpretation pattern. Their world knowledge measure involved rating German equivalents of items like *Some head injuries are too severe to be ignored* for the depth charge item *No head injury is too trivial to be ignored*. They collected norming scores for 32 German items and found that depth-charge items whose world knowledge rating scores were higher received more misinterpretations.

2.1.2 Alternative Degree Quantifier Constructions

The degree quantifier construction *too...to* has also been claimed to trigger the depth charge illusion, especially relative to other degree quantifier constructions such as *enough to* and *so ...that*. O'Connor (2015, Experiment 7b) asked participants to judge whether sentences like *According to the politician, no social program is too wasteful to oppose* make sense compared with *According to the politician, no social program is wasteful enough to oppose*. She found that sentences with *too* were twice as likely to elicit the illusion than those with *enough*. Similarly, in Paape et al. (2020, Experiment 3), German depth-charge sentences which were translated as *No head injury is too innocuous to be ignored* received a significantly higher illusion rate than *No head injury is so innocuous that it should not be ignored*.

The semantics of these degree quantifier constructions might shed light on the processing difficulty posed by *too*. At first sight, all three constructions *too...to*, *enough to*, and *so...that* describe the degree of the subject's property and the associated possibility of the action denoted by the verb phrase (Hacquard, 2005; Heim, 2000; Meier, 2003). Yet looking closer, the implications of the three constructions are different. As shown in (9), *enough to* in

(9a) and *so...that* in (9b) presuppose that the larger the degree of the subject's property, the more probable that the action indicated by the verb will take place. For example, the older Alice is, the more likely that she is capable or allowed to drive. In contrast, *too* in (9c) presupposes that the degree of the adjective surpasses the baseline above which the action would not be allowable or possible. Therefore, sentence (9c) is interpreted as Alice has surpassed the age above which driving is not recommended and the older she is, the less likely she will be able to drive. To sum up, the adjective degree is negatively correlated with the possibility of the action in *too...to* while the correlation is the opposite in *enough to* and *so...that*. This negative correlation embedded in *too...to* might increase the difficulty to process depth-charge sentences compared with parallel sentences with *enough to* and *so...that*.

(9) a. Alice is old enough to drive.

b. Alice is so old that she can drive.

c. Alice is too old to drive.

Relatedly, a heuristic processing strategy has been proposed such that *too* could be mentally transformed to *enough* under global negation during comprehension (O'Connor, 2015, 2017, cf. Christianson et al., 2001; Ferreira, 2003; Karimi & Ferreira, 2016), but the underlying mechanism is still unclear.

The investigation into *too...to* and its potential connection with alternative degree quantifier constructions offer valuable insight for our approach of adopting the noisy-channel theory to explain the depth-charge illusion.

2.1.3 Negation

Wason & Reich (1979) suggested that the final negative verb might pose special diffi-

culty for incremental processing and thus tend to trigger the depth-charge illusion (see also Cook & Stevenson, 2010). But this intuition has not been supported by empirical testing. O'Connor (2015) conducted a series of plausibility judgment experiments where the critical sentences were manipulated to differ from the depth-charge sentences in the sentential quantifier (*no* vs. *all*), the degree quantifier (*too* vs. *enough*), or the final verb polarity (positive vs. negative). The results indicated that while all three changes significantly affected the misinterpretation rate, the final verb manipulation had the smallest effect. In Paape et al. (2020), even when participants were asked to complete a sentence beginning with the preamble *No head injury is too trivial to*, the most typical completion had the negative meaning, as in *ignore*. These findings indicate that the semantic illusion might have already taken place prior to the final verb and the negative verb is not a primary source of the illusion.

Secondly, Paape et al. (2020) framed a memory-based overloading hypothesis based on Wason and Reich's claim that the combination of the global negation *no*, *too...to*, a negative adjective (e.g., *trivial*) and a negative verb (e.g., *ignore*) could overload comprehenders' processing and make depth-charge sentences hard to comprehend, as processing negation incurs extra cost (e.g., Horn, 2009; Just & Carpenter, 1971; Just & Clark, 1973; Sherman, 1976). Paape and colleagues further predicted that the misinterpretation rate of depth-charge sentences should be a function of individual working memory capacity – the higher the working memory, the more likely that individuals can recover the literal meaning. In their Experiment 2A, they measured individuals' working memory based on their performance on the operation span task (Nicenboim et al., 2016; Turner & Engle, 1989) and they found that even though individuals with higher working memory capacity did slow down while reading the negative adjective and the final verb, they were not less likely to arrive at an illusion. A

complicating factor here is that it is challenging to measure individual differences in experimental effects on sentence processing (e.g., James et al., 2018).

Furthermore, Paape et al. (2020) proposed a "negation cancellation" heuristic processing strategy which essentially says two negative meanings in a clause cancel each other out and if the transformed sentence is plausible, the original sentence is plausible, too. For example, the transformed sentence of *It's not like you didn't cheat on me* could be *You cheated on me*. The plausibility of the transformed sentence guarantees that the original double negative is plausible. Analogously, for depth-charge sentences, the sentential negation and the negative adjective could cancel out each other. *No head injury is too trivial to be ignored* is then transformed to the plausible *At least one head injury is too dangerous to be ignored*, which leads to a plausible interpretation of the depth-charge sentence. However, this proposed explanation does not address why the negative meanings in *too* and the final verb escape the cancellation. More importantly, as Horn (2010) pointed out, double negatives do not always make an affirmative in every case in English. Therefore, in its current form, the double negative cancellation strategy does not provide a complete explanation for depth-charge sentences.

2.1.4 A Construction-based Non-illusory Account

Some researchers working within a construction grammar framework (e.g., Goldberg, 1995; Kay & Charles J., 1999) have argued that the common interpretations of depth-charge materials are compositionally derived via a special reading of *too...to* under negation so that there is no "illusion" under this account (Cook & Stevenson, 2010; Fortuin, 2014). For example, in (10a), *too...to* retains its canonical reading as in (9c) but in (10b), *too...to* means that the property of the subject exceeds some degree and the action denoted by the verb is real-

ized as a consequence. Sentence (10b) is thus interpreted as "there is no head injury that is so trivial to the extent that it is ignored" and has a similar meaning to *No head injury is so trivial as to be ignored*.

(10) a. No head injury is too trivial to be treated.

b. (#) No head injury is too trivial to be ignored.

Fortuin (2014) argues that the existence of depth-charge constructions in corpora supports the legitimacy of the two readings for *too...to* but this argument doesn't consider that everyday language production contains errors (e.g., Fromkin, 1980; Dell & Reich, 1981). Another limitation of this non-illusory construction hypothesis is that there is no independent evidence that *too...to* has two interpretations in constructions without negation. There has also been rich discussion in the construction grammar literature of the hypothesis that superficial resemblance between similar constructions can cause interference among their meanings (Pijpops et al., 2018; Pijpops & Van de Velde, 2016). To some extent, this aligns with the structural similarity between *so...as to* and *too...to* in the depth-charge case. In sum, the study of form and meaning together with form frequency, which is the core of construction grammar, likely plays a role in the processing of depth-charge sentences. Yet there has been little direct empirical work linking construction grammar to the depth-charge illusion.

2.2 New Approach

In this thesis, we adopt the noisy-channel framework of language processing (section 1.2, with the Equation 1.1 repeated here as Equation 2.1) to investigate whether the comprehension of depth-charge sentences can be construed as noisy-channel inference. As reviewed

above, readers appear to interpret depth-charge sentences, such as *No head injury is too trivial to be ignored*, according to a more plausible alternative which is inconsistent with the literal meaning of the string. This motivates us to model the depth-charge sentence as the perceived sentence s_p, people's prior understanding of the topic or the intended meaning as the prior $P(s_i)$, and the misinterpretation rate of depth-charge sentences as the posterior $P(s_i|s_p)$.

$$P(s_i|s_p) \propto P(s_i)P(s_p|s_i) \tag{2.1}$$

Based on Equation 2.1, when the intended interpretation of depth-charge sentences is more consistent with world knowledge, readers will be more likely to go with an inferred reading, echoing claims by Wason & Reich (1979) and Paape et al. (2020). As for the noise model $P(s_p|s_i)$, we propose that the intended meaning m_i is "no matter how trivial head injuries are, they should not be ignored" and is intended to be expressed as *No head injury is so trivial as to be ignored* (s_i). We hypothesize that this plausible s_i gets corrupted into the depth-charge s_p, i.e., *No head injury is too trivial to be ignored*, through a production error, specifically a structural substitution. Our proposal is inspired by the structural and semantic similarity between *so...as to* and *too...to* and the existing research on various kinds of degree quantifier constructions. Equation 2.1 predicts that sentences which appear to result from noise corruptions with a higher likelihood of occurring will have a higher rate of inferential (non-literal) reading.

Our noisy-channel model of depth-charge illusions builds on a rich existing literature (reviewed in section 2.1) and provides a crucial advance: rather than isolating each factor to its own effect on the illusion, ours integrates the role of world knowledge and the intuition that

the availability of multiple degree quantifier constructions might motivate comprehenders to switch between a more plausible neighboring sentence and the complicated depth-charge sentence. This framework also makes quantitative/ordinal predictions – the probability of observing the semantic illusion is affected by (i) the comprehender's prior knowledge about probabilities of potential interpretations of the depth-charge, $P(s_i)$, and (ii) how likely a given meaning m_i with a plausible structure s_i is to be produced as the corrupted depth-charge sentence $P(s_p|s_i)$.

2.2.1 Experiment outline

In Experiment 1, we conducted a plausibility rating study in English that replicated the depth-charge illusion in Paape et al. (2020)'s German materials. Experiment 2 was a world knowledge norming study that measured how consistent the intended meaning of depth-charge sentences is with world knowledge, which offered a proxy for the prior $P(s_i)$. The noisy-channel model predicts that the higher the prior $P(s_i)$, the higher the posterior $P(s_i|s_p)$. Indeed, plausibility ratings of depth-charge sentences were positively correlated with the world knowledge consistency score of the alternative meaning.

In Experiment 3, we investigated the noise term $P(s_p|s_i)$ by hypothesizing two types of noise edits inspired by previous findings around multiple negative meanings, the degree quantifier constructions, and theories in speech production: 1) structural substitution where the intended s_i *No head injury is so trivial as to be ignored* is produced as *No head injury is too trivial to be ignored* and 2) antonym substitution where the intended s_i *No head injury is too trivial to be treated* is produced as *No head injury is too trivial to be ignored*. We gathered ratings of how likely each noise type would be to happen during production

as a proxy for the noise likelihood term $P(s_p|s_i)$. Structural substitution had a higher average likelihood rating compared to antonym substitution, suggesting that it is a more likely candidate for the noise corruption. We further found that *so...as to* is more likely to be produced as *too...to* rather than the other way around. In Experiment 4, we found that implausible (depth-charge) sentences with *too...to* were more likely to be interpreted non-literally, based on responses to comprehension questions, than implausible sentences with *so...as to*, as predicted by the noisy-channel theory. In sum, across four experiments, comprehension patterns related to depth-charge sentences are consistent with a noisy-channel explanation.

2.3 Experiment 1: Replication of the Illusion Effect

The goal of Experiment 1 was to replicate the depth-charge illusion in English. The materials consisted of sentences from the plausibility rating task in Paape et al. (2020), translated from German to English. The crucial depth-charge sentences had the sentential negation *no*, the structure *too...to*, negative adjectives like *trivial*, and negative verbs like *ignore*. According to Paape et al. (2020), these sentences should receive higher plausibility ratings compared with the implausible control sentences.

2.3.1 Methods

Participants

64 participants were recruited from Amazon's Mechanical Turk to complete the task. Each participant was paid $3 for their participation. We excluded data from those (a) who did not rate at least 90% of trials; (b) who did not answer at least 75% of the comprehension checks correctly; (c) who gave the same rating across all test trials; and/or (d) who self-

identified as non-native speakers of English. We analyzed the remaining 58 participants' responses.

Materials & Procedure

The materials were translated to English from Paape et al. (2020) Experiment 1 materials. There were 32 target items which appeared in 4 conditions, crossing the sentence initial quantifier and the polarity of the adjective as in (11). The sentence initial quantifier was either *some* or *no* and the adjective was manipulated to be either positive or negative. The polarity of the adjective was determined by the adjective-verb relation. For positive adjectives, higher degree of property leads to lower probability of the action denoted by the verb. For negative adjectives, higher degree of property leads to higher probability of the action. For example, the adjective "severe" in (11) was encoded "positive" because the more severe head injuries are, the less probable we are to ignore them; the adjective "trivial" was "negative" because the more trivial head injuries are, the more likely of the ignorance. The final verb in all conditions was negative (e.g., *ignore*), conveying the meaning that no action or attention would fall upon the target denoted by the sentence subject.

(11) a. quantifier-some, positive-adjective (plausible)
Some head injuries are too severe to be ignored.

b. quantifier-some, negative-adjective (implausible)
Some head injuries are too trivial to be ignored.

c. quantifier-no, positive-adjective (implausible)
No head injury is too severe to be ignored.

d. quantifier-no, negative-adjective (implausible / "depth-charge")

No head injury is too trivial to be ignored.

The materials were designed such that sentences in the quantifier-some, positive-adjective condition (11a) were plausible, whereas those in the other conditions were not. In the quantifier-some, negative-adjective condition (11b), the sentences were implausible because the semantic relation between the adjective and the verb in the construction of *too…to* went against common sense – for example, *too trivial to be ignored* implied the more trivial head injuries are, the more we should not ignore them, which was contrastive with how we view head injuries. The condition with quantifier-no and positive-adjective (11c) was implausible because its literal meaning conveyed "there is no head injury such that it is very severe", which was contrary to common knowledge. The condition of quantifier-no and negative-adjective (11d) featured the critical depth-charge sentence whose literal meaning was "no matter how trivial head injuries are, they should be ignored". Because this meaning also went against the common sense beliefs, it was implausible. Besides the critical trials, there were 60 filler items with varying syntactic structures. Each filler conveyed generic properties of or commonly held attitudes towards certain entities (e.g., *Consuming too much fat increases the risk of heart disease*; *Some earthquakes are very hard to be predicted by current technology*). Within these 60 fillers, 40 were plausible and 20 implausible.

Each participant read a total of 92 randomized trials. The presentation of the trials followed a Latin Square design output by the Turkolizer software (Gibson et al., 2011) and the within-subjects design guaranteed that each participant read the same number of trials under each of the four conditions. After each trial, there was a YES/NO comprehension question asking if a specific word appeared in the target sentence (e.g., Does "head injury" appear in this sentence?). The answers were designed such that half of the trials were "yes"

and the other half "no". The comprehension question served as an attention check. After the comprehension question, there was a rating question that asked participants to indicate whether the sentence made sense and whether it was written well using a 7-point Likert scale (1 = "makes no sense/written poorly", 4 = "intermediate", 7 = "makes perfect sense/written well"), following Paape et al. (2020).

2.3.2 Results

The distributions of plausibility ratings by condition are presented in Figure 2.1*. The sentences in the quantifier-some, positive-adjective condition (11a) were largely rated as plausible (mean = 5.70, confidence interval (CI) = [5.54,5.84]). Sentences from the quantifier-some, negative-adjective condition (11b), were rated less plausible (mean = 3.17, CI = [3.01, 3.35]). Sentences from the quantifier-no, positive-adjective condition (11c), (mean = 4.09, CI = [3.91, 4.25]) were rated in the intermediate range. Crucially, the implausible depth-charge sentences (quantifier-no, negative-adjective condition; 11d) were rated as more plausible than the other two implausible sentence conditions (mean = 4.54, CI = [4.37, 4.71]).

To assess these effects statistically, we fit the data into Bayesian multilevel cumulative ordinal models using the *brms* package (Bürkner, 2017; Bürkner & Vuorre, 2019) in R, following the same analytical strategy in Paape (2020). The raw plausibility ratings per trial were the dependent variable with non-equidistant intervals between levels on the Likert scale. The four condition levels were entered as a dummy-coded fixed effect (reference level = the depth-charge condition). Random intercepts and slopes for condition for both subjects

*The depth-charge sentences were rated as more plausible than the other implausible conditions. (The middle horizontal line in the boxplot represents the median of acceptability rating scores; the black dot represents the mean; the jittered points represent single trial plausibility ratings.)

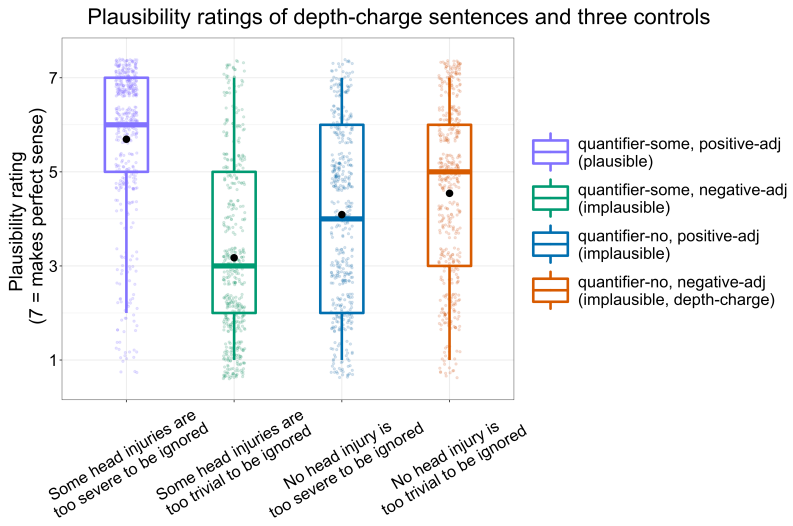


Figure 2.1: Plausibility ratings of depth-charge sentences

and items were included as random effects to obtain the maximal random effect structure for mixed effects models (Barr et al., 2013). Following the prior setting in Paape et al. (2020), the prior distributions for all the intercepts and coefficients of fixed effects were specified as a normal distribution with a mean of 0 and a standard deviation of 2 (i.e., Normal(0,2)); the prior for the correlation matrices of random effects was set to be LKJ(2) – LKJ has been the default weakly informative prior for correlation matrices in *brms* (Lewandowski et al., 2009; Nalborczyk et al., 2019) and 2 was chosen following (Paape et al., 2020). *Brms* default priors were used for all other parameters. These priors mildly restrict the possible coefficient for each parameter but still allow reasonably large variance. The model had four sampling chains each with 4000 iterations. The first 2000 samples were taken as warmup. An $\hat{R}$ close to 1.0 marks the convergence of the sampling chain to the underlying posterior distribution of the target predictor (Gelman & Rubin, 1992). All $\hat{R}$s for the sampling chains for all fixed effects were 1.0, indicating successful convergence. In this paper, we use β to

represent the estimated coefficients for predictors and CrI to represent the credible interval. The analysis code for all experiments and the model output summaries are available at https://osf.io/nhytx/.

Compared with the depth-charge condition, the quantifier-some & positive-adjective condition elicited higher plausibility ratings (β = 0.96, CrI = [0.61, 1.33]); the quantifier-some & negative-adjective condition got lower plausibility rating (β = -0.94, CrI = [-1.21, -0.66]) and so did the condition with quantifier-no & positive-adjective (β = -0.32, CrI = [-0.60, -0.04]).

2.3.3 Discussion

Experiment 1 successfully replicated the findings in Paape et al. (2020). Depth-charge sentences were rated as more plausible than the other two implausible control conditions, suggesting that readers may have inferred a non-literal meaning during comprehension. It is worth noting that the plausibility ratings of sentences like (11c) received intermediate ratings, which resemble the findings in O'Connor (2015) and Paape et al. (2020) and suggest the polarity of the adjective might not be a crucial reason for the depth-charge illusion.

2.4 Experiment 2: World Knowledge Prior

In Experiment 2, we tested the first hypothesis of the noisy-channel explanation by investigating whether the prior probability of the likely intended meanings of depth-charge sentences, (e.g., $P(s_i$:"Head injuries should be treated, no matter how trivial they are")), would predict how plausible they judge the sentence (e.g., *No head injuries are too trivial to be ignored*). On a noisy-channel account, depth-charge sentences may be interpreted rela-

tive to the meaning of an alternative, plausible sentence, rather than their literal meaning. As a proxy for $P(s_i)$, we measured how consistent the plausible alternative meaning (e.g., "Head injuries should be treated, no matter how trivial they are") of each depth-charge sentence was with people's commonly held world knowledge. The prediction is that items whose plausible alternative meanings are more consistent with world knowledge should receive higher plausibility ratings in their implausible, depth-charge form.

2.4.1 Methods

Participants

A total of 35 participants were recruited from Amazon's Mechanical Turk and each was paid $2. In addition to the same screening check as in Experiment 1, participants needed to finish an English sentence completion task to verify their identity as native English speakers. In the end, 31 participants remained for analysis.

Materials & Procedure

We created a sentence to represent the plausible intended meaning of each of the 32 depth-charge items in Experiment 1. All the sentences followed the template of "[TOPIC] can/should be [ANTI VERB-ed], no matter how [NEGATIVE ADJ] they are". For example, the plausible target meaning for (11d) is in (12):

(12) Head injuries should be treated, no matter how trivial they are.

Here "[TOPIC]" was the noun phrase that represents the entity under discussion; the modal verb was selected from can and should to best represent the felicitous modality as-

sociated with the attitude toward the topic and the action; "[ANTI VERB-ed]" was the antonym of the sentence final verb in the original depth-charge item; the "[NEGATIVE ADJ]" was the one from the depth-charge item. Two native speakers of English (Rachel Ryskin and Edward Gibson) verified the grammaticality of the items. There were no filler items in the experiment because we expected the participants to be consciously aware of the contrast between the items so that the variance among scores was larger.

Each participant read 32 sentences in a randomized order and answered the question "According to what you believe about the world, how much do you agree with the sentence?" on a fully labeled 7-point Likert scale (1 = "completely disagree", 4 = "intermediate", 7 = "completely agree"). They were then asked to answer a YES/NO comprehension question (e.g., does "be" appear in this sentence?) that probed their attention. Before the task began, participants were asked to complete five English sentences (e.g., *Lots of people love drinking coffee because* ...). Responses to these catch trials were used to exclude bots from the dataset.

2.4.2 Results

Figure 2.2 displays the distribution of the world knowledge norming scores per critical item. Average world knowledge scores ranged from 3.09 (*Artificial ingredients should be abandoned, no matter how harmless they are*, CI = [2.43, 3.75]) to 6.61 (*Head injuries should be treated, no matter how trivial they are*, CI = [6.40, 6.83]).

Figure 2.3* plots the plausibility ratings from Experiment 1 for the 32 critical items in

*In this figure, lines represent a linear regression line; Shades represent 95% bootstrapped CI; The x-axis was set from 1 to 7 but the result started from above 3; The y-axis was set from 1 to 7 but the result started from above 2.

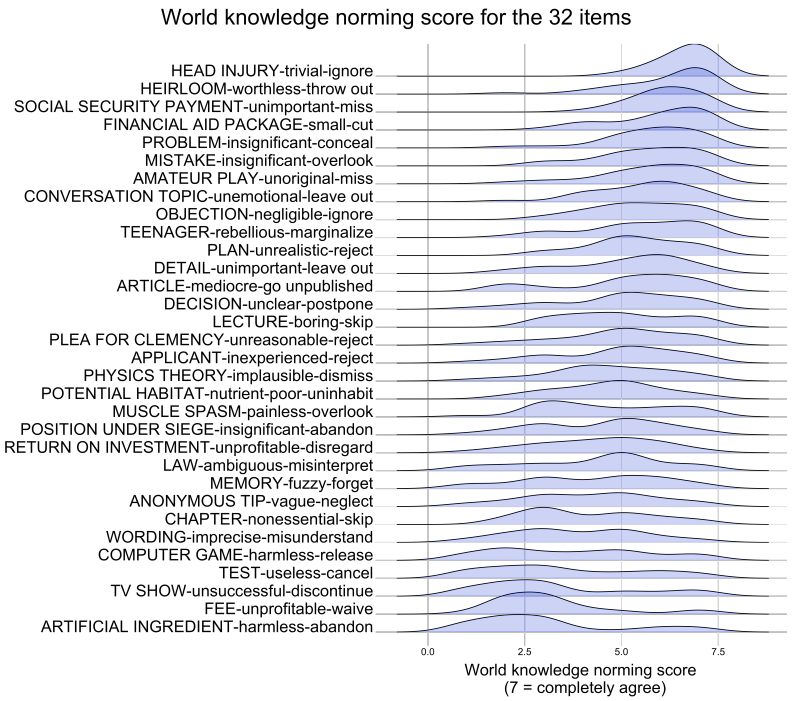


Figure 2.2: The distribution of world knowledge norming score

the four conditions over their average world knowledge norming scores. The depth-charge condition shows a clear positive relationship between the world knowledge score and the plausibility rating. We further analyzed the effect of world knowledge using Bayesian multilevel cumulative ordinal models via the *brms* package. The dependent variable was the raw plausibility score per trial. The fixed effects included the dummy-coded condition variable (reference level = the depth-charge condition), centered world knowledge score, and their interaction terms. The random effects structure contained random intercepts as well as maximal random slopes for both subject and item. The priors, number of sampling chains, number of iterations, and warmup setup were the same as those in Experiment 1. The *emmeans* package (Lenth et al., 2019) was used to estimate the linear trend effect of world knowledge on each condition. For output from this package, we use β to represent the estimated coefficients and HPD to represent the highest posterior density which is the shortest interval with the highest density in the posterior distribution of target coefficient (e.g., Box & Tiao, 2011).

The Bayesian analysis shows that world knowledge score has a positive effect on the plausibility rating in the depth-charge condition (β = 0.27, HPD = [0.10, 0.45]). Yet none of the other three conditions seem to have a clear effect because their HPDs all contain zeros (quantifier-some & positive-adjective: β = -0.14, HPD = [-0.50, 0.22]; quantifier-some & negative-adjective: β = 0.20, HPD = [-0.04, 0.46]; quantifier-no & positive-adjective: β = -0.07, HPD = [-0.28, 0.18]).

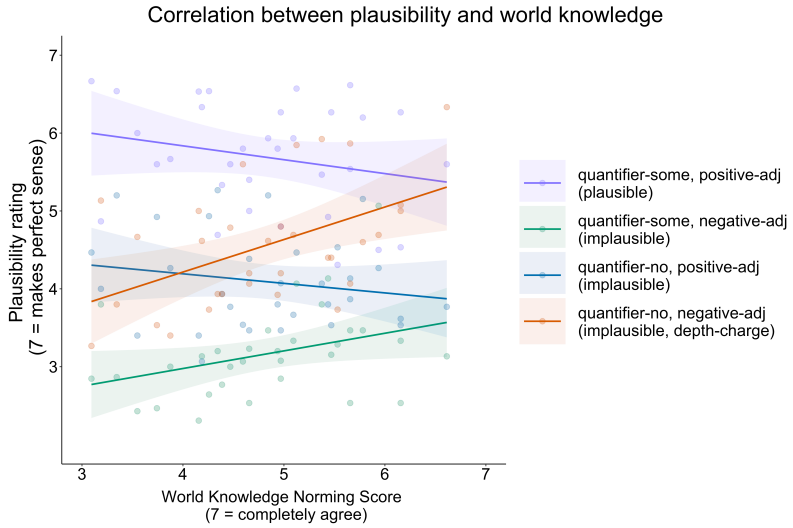


Figure 2.3: The correlation between the world knowledge norming score and the plausibility rating

2.4.3 Discussion

The positive correlation between the world knowledge score and the plausibility rating in the depth-charge condition suggests that sentences which have a plausible alternative that has high (semantic) prior probability elicit higher plausibility ratings. The results echo those in Paape et al. (2020) and are consistent with the noisy-channel account that depth-charge sentences are interpreted non-literally, according to a more plausible alternative.

2.5 Experiment 3: Noise Likelihood Rating

2.5.1 Experiment 3a: Edit likelihood ratings for depth-charge sentences

Experiment 3a focused on alternative plausible formulations of the plausible meaning associated with the depth-charge sentences. We investigated two types of possible noise opera-

tions – structural substitution and antonym substitution – that might lead to the generation of depth-charge sentences. We also measure English speaker's perceptions of the probability of these two operations. These two error types were inspired by the structural similarity between *too...to* and other degree quantifier constructions (see section 2.1.2) and the intuition that the difficulty of computing negations could apply to the selection of the final verb versus its antonym. We refer to the first proposed error type as structural substitution, as in (13) where the intended sentence contains *so...as to* but was produced as *too...to*. The second proposed production error type features antonym substitution as in (14) where the intended plausible sentence has the plausible *too trivial to be treated* but is produced as the implausible *too trivial to be ignored*.

(13) Structural substitution

   a. No head injury is **so** trivial **as to** be ignored. (plausible, intended)

   b. No head injury is **too** trivial **to** be ignored. (implausible, produced, depth-charge)

(14) Antonym substitution

   a. No head injury is too trivial to be **treated**. (plausible, intended)

   b. No head injury is too trivial to be **ignored**. (implausible, produced, depth-charge)

In speech production, substitution errors occur when the intended word and the produced word belong to the same grammatical category and share similar semantic or phonological features (Dell & Reich, 1981; Fay, 1981; Harley, 1984). The constructions *so...as to* and *too...to* share the same syntactic distributions, and the monosyllabic nature of *so* and *too*

makes their phonology similar. They are also semantically related in the sense that both describe the relation between the adjective degree and the probability of occurrence for the action denoted by the sentence final verb, except that the relations of the two constructions are opposite. We further hypothesize that the substitution from *so...as to* to *too...to* is more likely than in the other direction because *too...to* is more frequent than *so...as to* (8828 tokens vs. 500 tokens in the Corpus of Contemporary American English (Davies, 2015)) and it has been shown that words and structures with low frequency are more likely to be substituted by high frequency ones (Harley & MacAndrew, 2001; Kapatsinski, 2010; Kittredge et al., 2008; Stemberger, 1984, a.o.).

Antonym substitution, as in (14), is a well-known type of semantically related word substitution (Hotopf, 1980; Murphy, 2003) (see the example *The authorities had to decide where to bury the survivors* [intended: *the casualties*] in Barton & Sanford (1993)). We further assume that the more negative meanings there are in a sentence, the more likely antonym substitution would occur, because of the increasing demand to track the polarity of the meaning. Therefore, we would be more likely to observe antonym substitution in depth-charge sentences with *too...to* compared with *so...as to* because the former with *too* embeds an implicit negative interpretation of the final verb.

In sum, we make a prediction for each of the two proposed error types. For structural substitution, we predict that it would be more likely for (13a) to be produced as (13b) compared with (15a) being produced as (15b). For the antonym substitution error, we predict it would be more likely for (14a) to be produced as (14b) compared with (16a) being produced as (16b).

(15) a. No head injury is **too** trivial **to** be treated. (plausible, intended)

b. No head injury is **so** trivial **as to** be treated. (implausible, produced)

(16) a. No head injury is so trivial as to be **ignored**. (plausible, intended)

b. No head injury is so trivial as to be **treated**. (implausible, produced)

Experiment 3a used a novel rating task that showed participants pairs of sentences and asked them to judge how likely it is that the first sentence might be intended by a speaker but produced as the second sentence during rapid speech. Here we approximated the error rates in production by relying on native speakers' intuition of the probability of such errors. There are several considerations behind this decision. First, it is difficult to get speakers to produce the kinds of materials that we are interested in here in a naturalistic way. The more naturalistic the method – such that we set up the contexts that are appropriate for depth-charge materials – the smaller the likelihood that participants would produce anything like the target materials that we are interested in (e.g., they would probably not produce sentences with the depth-charge structure in the first place). On the flip side, the less naturalistic the method, the harder it is to use the results as proxies for the likelihood of error in actual production. Second, speech error corpora are not helpful in this case because they tell us little about the complex materials that we are interested in, and suffer from error representative bias and collector biases in error identification methods anyway (e.g., Bock & Levelt, 1994; Pérez et al., 2007; Stemberger, 1992). As a result, we devised this noise likelihood rating paradigm for error estimation, which assumes that readers' responses will reflect their internal noise models. While they may not have direct access to this model, the relative ratings they produce in our task should be monotonically related to – the estimates of noise that live in the minds of readers/listeners. We also validated the noise rating rating task in Experiment 3b by collecting noise ratings for all the noise operations posited in Gibson et al.

(2013a).

Participants

A total of 64 participants were recruited from Amazon's Mechanical Turk and each was paid $3. After applying the same screening check as previous experiments, 43 participants remained for the analysis.

Materials & Procedure

Participants were presented with the following context for the experiment: "People make speech errors all the time when they intend to convey ideas in spoken sentences, especially when they are distracted or speaking fast. These errors include but are not limited to deletions, insertions, exchanges, and substitutions of certain words". Then the participants were presented with both the plausible intended sentence and the implausible perceived sentence and asked "given the intended meaning, how likely is it that someone would say the produced sentence when speaking quickly?". The participants responded on a 7-point scale (1 = "absolutely unlikely", 4 = "intermediate", 7 = "absolutely likely"). Before the rating, there was a YES/NO comprehension question to gather whether the participant understood the intended/literal meaning of the plausible sentence. For example, "According to the intended sentence, should head injuries be ignored?" (correct answer: no). The polarity of the final verb (e.g., *ignore* vs. *treat*) in the comprehension question was counterbalanced so that half of the answers were YES and the other half NO.

We selected 24 items which had the highest world knowledge norming scores in Experiment 2. The material varied by the type of noise operation (structural vs. antonym substi-

Conditions	Intended sentence (plausible)	Produced sentence (implausible)
Structural substitution *so...as to* → *too...to*	No head injury is so trivial as to be ignored.	No head injury is too trivial to be ignored. (depth-charge)
Structural substitution *too...to* → *so...as to*	No head injury is too trivial to be treated.	No head injury is so trivial as to be treated.
Antonym substitution with the intended *so...as to*	No head injury is so trivial as to be ignored.	No head injury is so trivial as to be treated.
Antonym substitution with the intended *too...to*	No head injury is too trivial to be treated.	No head injury is too trivial to be ignored. (depth-charge)

Table 2.1: Noise operation conditions for the depth-charge materials in Experiment 3a

tution) and the degree quantifier construction in the intended sentence (*so...as to* vs. *too to*) (see Table 2.1). The noise type manipulation was within-subjects. Each participant in Experiment 3 read a list of 74 randomized trials – 24 trials from the depth-charge material and 50 trials from Experiment 3b (described in Section 2.5.2). Before the task began, participants were asked to complete five English sentences. Responses to these sentence preambles were used to identify bots.

Results

We analyzed 973 trials out of 1032 (94.3%) whose answer to the comprehension check question was correct. The noise likelihood rating across the four conditions is displayed in Figure 2.4*. Overall, the structural substitution appeared to be a more probable noise op-

*In this figure, the black points are mean and the horizontal lines are median, with dots representing the noise likelihood rating for individual trials.

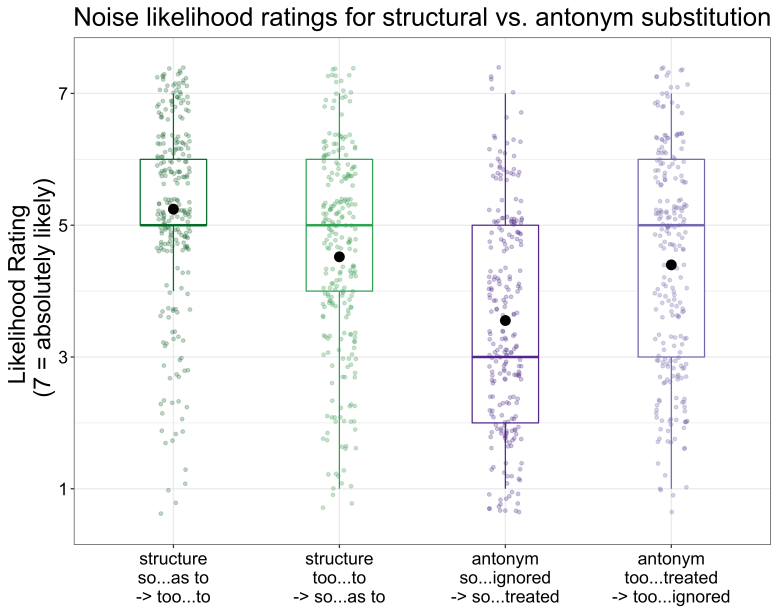


Figure 2.4: Noise likelihood ratings for the depth-charge material

eration (Mean = 4.88, SD = 1.54, CI = [4.75, 5.02]) than the antonym substitution (Mean = 3.97, SD = 1.69, CI = [3.83, 4.13]). Within the structural substitutions, the noise likelihood rating for the *so...as to* → *too to* condition (Mean = 5.24, SD = 1.45, CI = [5.08, 5.24]) was higher than the *too to* → *so...as to* condition (Mean = 4.52, SD = 1.55, CI = [4.32, 4.73]). Within the antonym substitution condition, the noise likelihood rating for sentences with *too...to* was higher (Mean = 4.40, SD = 1.65, CI = [4.18, 4.59]) than sentences with *so...as to* (Mean = 3.35, SD = ±1.63, CI = [3.33, 3.76]).

This pattern is supported by a Bayesian multilevel ordinal model: the dependent variable was the raw noise likelihood rating score from 1 to 7 per trial; the fixed effects included the noise type (structure vs. antonym), the intended sentence structure (*so...as to* vs. *too to*), the interaction term of these two factors; the random effects included random intercepts

and random slopes of the full fixed effects structure for subjects and items. All of the other Bayesian model parameters were the same as previous experiments (see supplemental material for details). All chains converged successfully ($\hat{R}$s = 1.0). We then used *emmeans* for analyzing specific contrasts.

In terms of the type of noise operation, structural substitution had a higher likelihood rating than antonym substitution (β = 0.83, HPD=[0.55, 1.09]). In the structural substitution condition, the intended sentence *No head injury is so trivial as to be ignored* was more likely to be produced as *No head injury is too trivial to be ignored*, compared with *No head injury is too trivial to be treated* being produced as *No head injury is so trivial as to be ignored* (β = 0.72, HPD = [0.39, 1.03]). This is consistent with our hypothesis that structures with a higher frequency could replace ones with a lower frequency in speech errors. On the other hand, in the antonym substitution condition, substitutions of verb antonyms were less likely to happen when the sentence structure contained *so...as to* (e.g., *No head injury is so trivial as to be ignored* → *No head injury is so trivial as to be treated*) compared with *too to* (e.g., *No head injury is too trivial to be treated* → *No head injury is too trivial to be ignored*) (β = -0.71, HPD=[-1.05, -0.38]). This finding is consistent with our hypothesis that the likelihood of antonym substitutions is higher in negative environments, such as with *too to*.

2.5.2 Experiment 3b: Edit likelihood ratings of materials in Gibson et al. (2013)

To contextualize the likelihood ratings for our proposed edits of the depth-charge materials and validate the noise likelihood rating paradigm, we included sentence pairs reflecting edits that have been more commonly assumed in the noisy-channel literature (i.e., word dele-

tions and insertions). Specifically, we used the critical items, across 5 syntactic alternations, from Gibson et al. (2013a) (Table 2.2).

Table 2.2: Critical material design in Gibson et al. (2013)

English alternations	Plausible version (s_i)	Implausible version (s_p)	Noise operation
Active / passive	a. The girl kicked the ball. b. The ball was kicked by the girl.	c. The girl was kicked by the ball. d. The ball kicked the girl.	2 insertions 2 deletions
Object-locative / subject-locative	a. The cat jumped onto a table. b. Onto the table jumped a cat.	c. Onto the cat jumped a table. d. The table jumped onto a cat.	1 insertion 1 deletion 1 deletion 1 insertion
Transitive / intransitive	a. The tax law benefited the businessman. b. The businessman benefited from the tax law.	c. The tax law benefited from the businessman. d. The businessman benefited the tax law.	1 insertion 1 deletion
DO / PO goal	a. The mother gave the daughter the candle. b. The mother gave the candle to the daughter.	c. The mother gave the daughter to the candle. d. The mother gave the candle the daughter.	1 insertion 1 deletion
DO / PO benefactive	a. The cook baked Lucy a cake. b. The cook baked a cake for Lucy.	c. The cook baked Lucy for a cake. d. The cook baked a cake Lucy.	1 insertion 1 deletion

Gibson et al. (2013a) proposed two general patterns of edit likelihood to account for the comprehension question responses that they observed (summarized in Figure 2.5A).* First,

*A. The approximate prediction of people's expectations about the likelihood of a given edit; B. Noise likelihood ratings from native English speakers. The black point indicates the mean of the

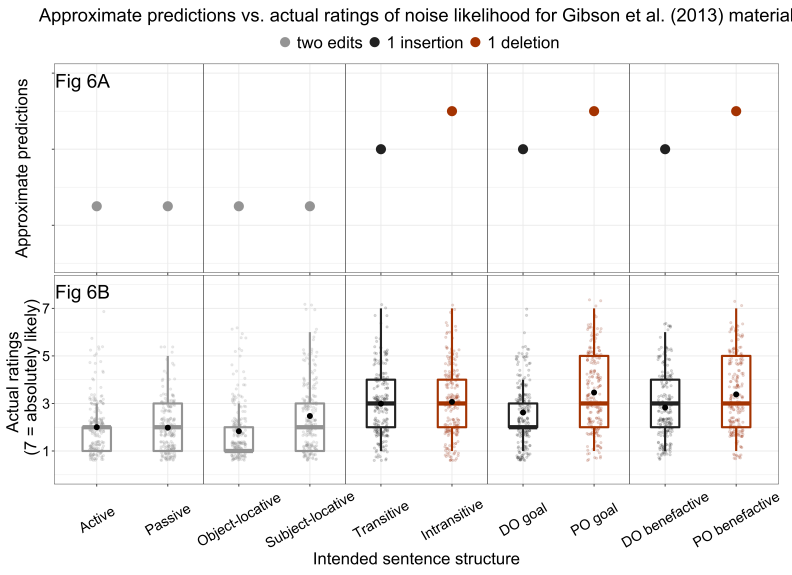


Figure 2.5: Approximate predictions and actual noise likelihood ratings for fillers

they proposed that participants would consider (not necessarily explicitly) a single edit (insertion or deletion) more likely than two or more edits. In other words, the edits that need to be posited to make a noisy-channel inference are less likely for the Active/Passive sentences and Subject-/Object-locatives than the other three alternations (i.e., Transitive/Intransitive, DO/PO goal, DO/PO benefactive). Second, Gibson et al. proposed that participants would consider a single deletion more likely than a single insertion. In other words, the edits that need to be posited to make a noisy-channel inference are less likely for Transitives, DO goal, and DO benefactives – the (a)-(c) pairs in Table 2.2 – than Intransitives, PO goals, and PO benefactives – the (b)-(d) pairs.

distribution and the boxplot indicates the quartiles of the distribution. Each dot represents a single trial noise-likelihood rating in a jittered presentation.

Participants

Participants were the same as in Experiment 3a.

Materials & Procedure

The procedure was identical to Experiment 3a. From the materials in Gibson et al. (2013a), we selected a random set of 5 pairs of items from each of the 10 conditions in Table 2.2. Each participant read the 50 pairs of sentences in a random order.

Results

Figure 2.5B shows the summaries of the noise likelihood rating across the 10 conditions. Overall, the pattern is consistent with the predictions. Sentence pairs with fewer edits, i.e., Transitive/Intransitive, DO/PO goal, and DO/PO benefactive, had higher likelihood ratings compared with conditions with more edits. Among the syntactic constructions with one edit, deletions were rated in general more likely than insertions (though the distinction was clearer in conditions of DO/PO goal and DO/PO benefactive, compared with Transitive/Intransitive).

The pattern is also supported by two Bayesian multilevel ordinal models. The prior selection and other sampling parameters were the same as those in the previous experiments. All chains converged successfully ($\hat{R}$s = 1.0).

The first model included the number of edits (1 or 2) as the independent variable and the raw likelihood rating as the dependent variable. The random effects included a random intercept for items and a random intercept as well as a random slope for edits for participants. Sentence pairs where the implausible version could result from two edits to the plausible

version were rated lower than those that could result from one edit (β = -1.05, CrI = [-1.35, -0.75]).

The second model analyzed the items with only one edit and included the noise type (deletion vs. insertion) as the fixed effect. The random effects included a random intercept for items and a random intercept as well as a random slope for noise type for subjects. Among the syntactic constructions that involve one edit, sentence pairs where the implausible version could result from deletion edits to the plausible version received higher likelihood ratings than those that resulted from insertions (β = 0.51, CrI = [0.21, 0.81]).

2.5.3 Discussion

In Experiment 3, we measured the likelihood of noise operations during information transmission for depth-charge sentences as well as sentences from Gibson et al. (2013a). The noise likelihood rating patterns are consistent with the findings in Gibson et al. (2013a) in terms of the effect of number of edits and noise type. Two edits were less likely than one edit; within one edit, deletions were more likely than insertions. The alignment between the experimentally rated error rates and the corresponding inference rates in Gibson et al. (2013a) partially validates this approach to estimate noise likelihood. For our critical depth-charge materials, both the structural substitution and the antonym substitution were rated as fairly likely. The noise likelihood ratings for depth-charge sentences were higher than those for Gibson et al. (2013a). Since structural substitutions were rated as more probable than antonym substitutions, we further investigated the former in Experiment 4 and treated its noise likelihood rating as a proxy for $P(s_p|s_i)$ for the depth-charge materials.

2.6 Experiment 4: Inferential Reading

Experiment 4 investigated whether the probability of an inferential interpretation of implausible depth-charge sentences could be predicted by the likelihood ratings of the structural substitution error that was tested in Experiment 3. We used a comprehension study to collect the probability of inferential reading, following the paradigm in Gibson et al. (2013a). The noisy-channel proposal predicts that implausible sentences with *too...to* (the canonical depth-charge sentences) should receive more inferential interpretation than the implausible sentences with *so...as to*.

2.6.1 Methods

Participants

A total of 72 participants were recruited from Amazon's Mechanical Turk and each was paid $3. We excluded from the final analysis those who self-identified or were checked by the English sentence completion task as non-native English speakers, who did not answer at least 75% of the filler comprehension questions correctly, and who did not finish at least 90% of all the trial sentences. 47 participants contributed data to the final analysis.

Material & Procedure

The 24 critical items were the same as those in Experiment 3a. The substitution direction and sentence plausibility were crossed in a 2 by 2 within-subjects design (Table 2.3). Participants were asked to answer a comprehension question, e.g., "Does this sentence mean 'head injuries should be ignored/treated, no matter how trivial they are?'". For the implau-

sible materials, the answer to the comprehension question determined whether participants interpreted the sentence literally or inferentially. The polarity of the verb used in the comprehension question was counterbalanced within each item. As a result, answering YES or NO to indicate a literal interpretation was counterbalanced.

Table 2.3: Conditions in Experiment 4

Direction of potential substitution	Plausible (no substitution)	Implausible (with substitution)
so...as to → too...to	No head injury is so trivial as to be ignored.	No head injury is too trivial to be ignored. (the canonical depth-charge sentence)
too...to → so...as to	No head injury is too trivial to be treated.	No head injury is so trivial as to be treated.

Apart from the 24 items, there were also 40 filler items that all described the generic properties of or the common attitudes towards the topic under discussion (e.g., *Cars should slow down when pedestrians walk across the road*). To ensure that filler sentences had structures similar to the critical items, 12 items began with the quantifier *no*, 12 began with *some*, and the rest of the 16 items had bi-clause structures. Similar to the design of the critical items, the YES/NO answer was also counterbalanced within each structure type.

2.6.2 Results

Figure 2.6* displays the literal interpretation rate for sentences in each condition (the inference rate can be computed by subtracting the literal rate from 1). The plausible sen-

*In this figure, the error bars indicate 95% bootstrapped confidence interval over subject means; the points indicate average literal interpretation rate by subject.

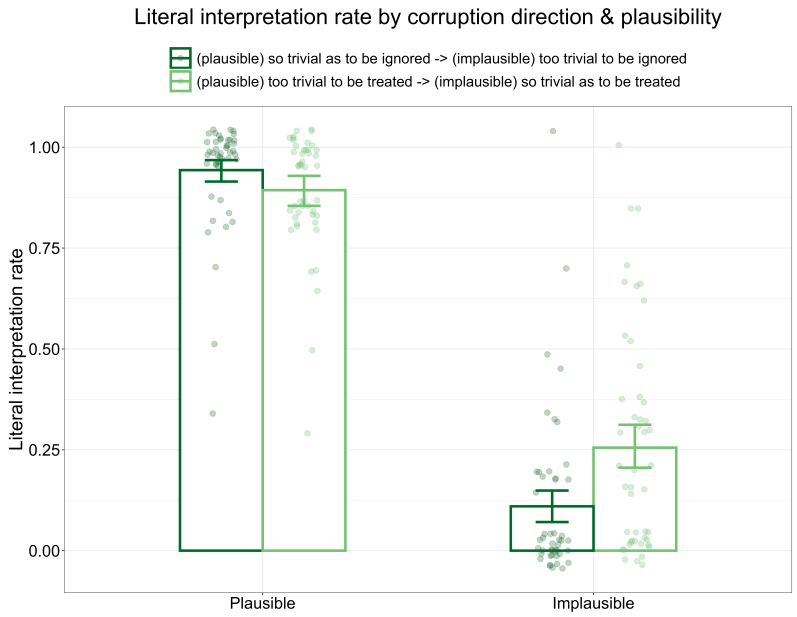


Figure 2.6: The literal interpretation rate of plausible and implausible sentences

tences were overwhelmingly interpreted literally, while the implausible sentences with *too...to* elicited significantly fewer literal interpretations and more inference than the implausible sentences with *so...as to*.

This difference was supported by a Bayesian logistic multilevel model analysis. The dependent variable was coded as 1 for the literal interpretation and 0 for the inference. The fixed effects contained the structure condition (*so...as to* vs. *too...to*, where the reference is *so...as to*), the plausibility of the sentences, and their interaction. The random effects include random intercepts as well as random slopes of all the fixed effects for both subjects and items. We specified moderately regularizing priors, selected via prior predictive simulation (Nicenboim et al., 2021), by setting the distribution of the intercepts to be Normal(0, 1) (i.e., a normal distribution with mean of 0 and standard deviation of 1), the distribution of

the coefficients to be Normal(0, 0.5) (i.e., a normal distribution with mean 0 and the standard deviation 0.5). *Brms* default priors were used for all other parameters.

After 4000 samples for each of the four chains, all chains converged with $\hat{R}$ equal to 1.00, indicating successful convergence. We used *emmeans* to analyze specific contrasts. We found that the literal interpretation rate for implausible sentences was lower than plausible sentences (β = -2.22, CrI = [-3.16, -1.14]). Within the plausible sentences, there was no difference in the literal interpretation rate between the two conditions (β = 0.35, HPD = [-0.31, 0.95]). For the implausible conditions, sentences with *so...as to* (e.g., *No head injury is so trivial as to be treated*) received more literal interpretation and thus less inference than sentences with *too to* (e.g., *No head injury is too trivial to be ignored*, i.e., the depth-charge sentence) (β = 1.36, HPD = [0.43, 2.25]).

In addition, we evaluated the ideal observer model of language comprehension in Equation 2.1 by comparing the rates of non-literal inference for implausible sentences from Experiment 4 with the estimates of normalized posterior probability for an alternative meaning, as shown in Figure 2.7*. The latter was calculated by multiplying the normalized world knowledge rating for each depth-charge item from Experiment 2 and the normalized noise likelihood averaged across all items in each of the two conditions (structural substitution errors from either *so...as to* to *too to*, or from *too to* to *so...as to*) from Experiment 3. The world knowledge rating was taken as the proxy for the prior probability $P(s_i)$ and the noise likelihood rating the proxy for the noise likelihood $P(s_p|s_i)$. Both normalizations were calculated by dividing the raw likert scale rating by the maximum scale value 7. (The normalized ratings are not proper probabilities; here we just assume a monotonic relationship between ratings

*The grey line is a linear fit between the non-literal inference and the estimated posterior probability; the error bars represent standard deviation.

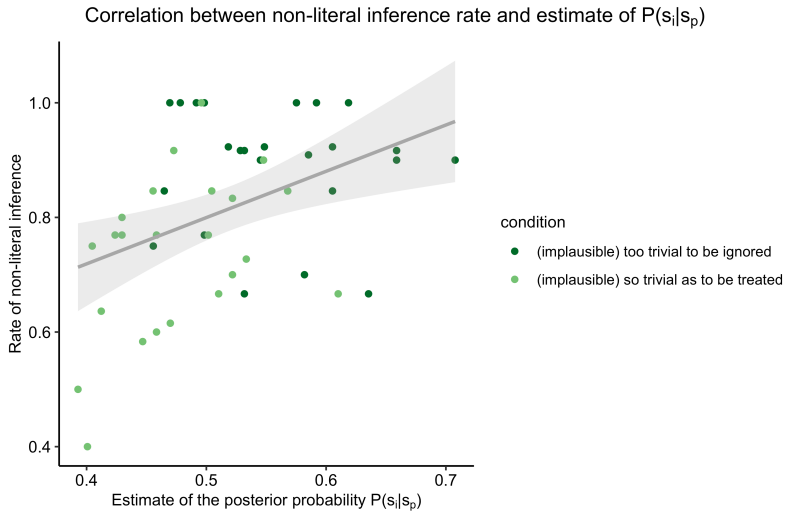


Figure 2.7: Correlations between the non-literal inference rates and estimates of the posterior probability $P(s_i|s_p)$

and probabilities.) Figure 2.7 shows a positive correlation between the two measurements, as predicted (Spearman rank correlation: r = 0.38, p = .008).

2.6.3 Discussion

In Experiment 4, the implausible sentences with *too to*, which are hypothesized to be the corrupted result of the plausible sentence with *so...as to*, were more likely to trigger inference, compared with the implausible *so...as to*. This inference rate pattern is consistent with the noise likelihood rating from Experiment 3 where it was judged that it was more likely for the plausible *so...as to* to be produced as *too to* than the other way around. In addition, we found that the rates of non-literal inference reading in Experiment 4 positively correlates with the rough estimation of the posterior probability which was calculated by multiplying the normalized world knowledge rating in Experiment 2 and the normalized noise likelihood rating

from Experiment 3. The relationship between the depth-charge illusion and the posited noise operations aligns with a noisy-channel explanation.

2.7 Conclusions and Discussions

We conducted four experiments to explain the depth-charge illusion for sentences like *No head injury is too trivial to be ignored* (Wason & Reich, 1979) within the noisy-channel framework of human communication (e.g., Gibson et al., 2013a; Levy, 2008; Shannon, 1948). We hypothesize that depth-charge illusions arise when readers encounter the sentences which are literally implausible and are likely to be the results of production errors. They then interpret the sentence according to how plausible an alternative meaning is according to world knowledge and how likely it is that the intended sentence was corrupted to the perceived form during production.

The findings in the four experiments supported these hypotheses. In Experiments 1 and 2, we showed that the plausibility rating of a depth-charge sentence was modulated by the plausibility of the intended meaning given world knowledge: the more plausible the intended meaning, the higher the plausibility rating for the depth-charge sentence. In Experiment 3, we investigated potential noise corruptions that could result in a depth-charge sentence and found that readers considered the structural substitution of *so...as to* with *too...to* to be the most likely production error. In Experiment 4, we found that, in line with the predictions of the noisy-channel framework, the probability of interpreting depth-charge sentences in terms of their non-literal meaning (i.e., the rate of inference) was predicted by the likelihood of the production error which could have corrupted a plausible sentence into the implausible version. We further show that the rate of inference positively correlates

with the estimated posterior probability (i.e., the product of the proxies of world knowledge prior and the noise likelihood). This synthesis further lends support to the noisy-channel explanation for depth-charge comprehension. Under this framework, the comprehension of depth-charge sentences is considered a rational inference process based on probabilistic reasoning with information from linguistic representations as well as world knowledge, on the assumption that linguistic input is noisy (e.g., Levy, 2008).

One of the major differences between the noisy-channel approach to depth-charge illusions and previous research is that the latter has largely focused on enumerating features of sentences that correlate with the illusion without providing an integrative explanation. Natsopoulos (1985), O'Connor (2015; 2017), and Paape et al. (2020) investigated the role of world knowledge but stopped short of explaining how the world knowledge conveyed by content words in the critical sentences led to an alternative reading. Similarly, previous work showed that degree quantifier constructions such as *so...that* and *enough to* were less likely to elicit the illusion than *too...to* (O'Connor, 2015; Paape et al., 2020), but the reason was not specified. One account posits that depth-charge illusions are caused by a working memory overload (Paape et al., 2020; Wason & Reich, 1979), but the relationship between working memory and interpretation behavior remains underspecified. Similarly, the construction-based non-illusory ambiguity account (Cook & Stevenson, 2010; Fortuin, 2014) claims that the depth-charge sentence is not implausible due to the double meaning of *too...to* but it lacks independent evidence for the two interpretations of *too...to*.

Construing depth-charge illusions through a noisy-channel lens ties together multiple previous research threads in this literature. The prior $P(s_i)$ in Equation 2.1 sheds light on why the interpretation of depth-charge materials is heavily influenced by world knowl-

edge (Paape et al., 2020, cf. Natsopoulos, 1985; O'Connor, 2015, 2017; Wason & Reich, 1979). The noise likelihood term $P(s_p|s_i)$ and the proposed structural substitution error connect observations based on comparisons between *too...to* and other quantifier structures (O'Connor, 2015, 2017; Paape et al., 2020). On the other hand, the depth-charge phenomenon also enriches the noisy-channel theory. While the noise model in the theory does not distinguish corruptions coming from the speaker, the hearer, or the environment (Gibson et al., 2013a; Levy, 2008), we construe the noise of depth-charge illusion as producer errors. Moreover, while previous noisy-channel work has focused on word or character-level edits, here we examine corruptions at the level of the construction.

Since the noise model in this framework deals with substitutions of degree quantifier constructions in English, such as *too...to*, *so...as to*, and *enough to...*, not simple string-level edits as in Levy (2008) or Gibson et al. (2013a), there are several implications. First, it expands the types of uncertainty in different linguistic levels of representations that could take place in language processing. While the original work on the rational approach of language processing (Levy, 2008) mainly introduces word level uncertainty such as word deletion/insertion/substitution (17), the case here illustrates that when producing the depth-charge sentences, the specific uncertainty could rest even deeper in the structure level and the alternatives are bound by their structural and functional similarity. Second, just as Gibson et al. (2013a) makes hypotheses on the cost of noise operations such as deletions are easier than insertions and one operation is easier than two, the general principle of noise operations is probability driven. If a noise operation is more likely to happen during production or information transmission, that operation is more likely to take place in the competition with other alternative operations. That also aligns with predictions and findings in the

depth-charge illusion. That is, out of the available alternatives, high-frequency structures are more likely to be uttered in general production settings and therefore the higher-frequent *too…to* structure substitutes the lower-frequent *so…as to*.

(17) a. The coach smiled at the player tossed a frisbee. (local coherence but difficult sentence)

b. The coach **who** smiled at the player tossed a frisbee. (near neighbor 1)

c. The coach smiled **as** the player tossed a frisbee. (near neighbor 2)

Nonetheless, the noisy-channel explanation for depth-charge illusions, in its current form, has several limitations which will need to be addressed in future work. The method we used for estimating the probability of substitutions in production in Experiment 3 makes several assumptions about the relationship between language use, the comprehender's perception of errors, and how they use that knowledge in the rating task. By asking participants to rate possible noise corruptions during speech production, we assume that the intuition of speech error likelihood correctly reflects the actual production likelihood and that this intuition is applied to the rating judgments. The results of Experiments 3a, 3b, and 4 suggest that it is a reasonable simplification. However, future work should corroborate these results using more naturalistic methods.

One follow-up question regarding the degree quantifier construction is which one out of *too…to* and *enough to…* is more likely to be uttered when the intended plausible sentence for the depth-charge is *No head injury is so trivial as to be ignored*. This question arises because both *too…to* and *enough to…* share similar frequency in a corpus search* and both are

*A search of the Corpus of Contemporary American English shows that *ADJ enough to* has 45,102 occurrences and *too ADJ to* has 48,190 occurrences, compared with *so ADJ as to* and the cor-

much more frequent than the *so...as to*. If the frequency modulates the substitution direction, it would be feasible to predict that the intended *so...as to* construction can be produced as the construction of *enough to*, as shown in (18), compared to the current hypothesized substitution direction (19). I argue that this is likely and this topic ties back to how different linguistic factors and the frequency of the structure interact with speech production in a dynamic way. It would be helpful to collect more empirical evidence, such as the noise likelihood rating, in order to compare the probability of (18) and (19) as alternative hypotheses of substitution.

(18) a. No head injury is **so** trivial **as to** be ignored. (intended, plausible)

b. No head injury is trivial **enough** to be ignored. (produced, plausible)

(19) a. No head injury is **so** trivial **as to** be ignored. (intended, plausible)

b. No head injury is **too** trivial **to** be ignored. (produced, implausible)

Another question is whether the degree quantifier construction *so...as to* could even be the intended plausible sentence at all, given its low frequency. During speech production, could the speaker just utter the depth-charge sentence, without a conscious representation of *so...as to* in their mind? One might argue that the low frequency of *so...as to* would inhibit its activation during speech production, and therefore it is unlikely that this structure is an underlying intended sentence. This points to another limitation of current work: I integrated the intended m_i and the intended s_i as one: the intended *No head injury is so trivial as to be ignored* is regarded both as the meaning that the speaker intends to convey and as the structure that the speaker chooses to linguistically represent the meaning. The bigger

responding 1,799 occurrences. A search of Google Ngram book shows similar results: *ADJ enough to* (0.0038%), *too ADJ to* (0.0040%), and *so ADJ as to* (0.0005%). Search date is June 11, 2024.

question now becomes how to map the hypothesized intended meaning (m_i) and sentence (s_i) in the Bayesian computational model to the algorithmic and the implementational levels of cognition (Marr, 1982). For example, given any hypothesized m_i, s_i, or s_p, when and how does the speaker generate the exact representation in their mind? What is the exact mechanistic realization of such a computational procedure? Due to the challenge to capture the detailed speech production procedure, this study does not provide direct supportive answer. Future studies should definitely try to bring a more mechanistic account to the rational framework.

In terms of achieving a mechanistic explanation of the inference procedure from the comprehender side, this work still faces limitations. So far, there is no concrete consensus about the mechanistic definition in the literature: noisy-channel inference could take place in the moment of processing a sentence or after the fact when the participants explicitly decide that a different sentence was intended. Here we speculate that the noisy-channel comprehension of depth-charge sentences takes place during processing of the sentence. This would be consistent with previous eye-tracking and ERP work demonstrating signatures of noisy-channel inferences taking place in real time when a stimulus is encountered that has low prior probability but can be attributed to a noise process (Levy et al., 2009; Ryskin et al., 2021b). Examining eye-tracking or ERP responses to depth-charge sentences would be a fruitful avenue for future work and may shed light on when in the sentence the noisy-channel inference takes place.

Further, our noisy-channel explanation of the depth-charge phenomenon assumes that the most likely source of noise corruption is the speaker/writer. This doesn't preclude additional corruptions from the reader's side (or from the environment). As the noisy-channel

framework is developed further, we expect that the noise model will be characterized at this finer-grained level.

Most importantly, the current noisy-channel account lacks an explanation of the role of negation. When looking for possible noise types in Experiment 3a, we explored the deletion and insertion of *not* as possible edits in the noisy-channel model. The results suggest that people might not consider deletion of *not* to be a likely noise operation in the context of these sentences and readers interpret sentences with multiple negations non-literally at a higher rate than sentences with fewer negations, even when they are semantically plausible. We also observed that antonym substitution is judged more likely to take place in an environment with multiple negation environments (Experiment 3a). These findings, along with previous reports in the literature (Kizach et al., 2016; O'Connor, 2015), suggest that negation plays an important role in the depth-charge illusion, but the exact mechanism remains an open question.

An alternative account, which could be applied to depth-charge illusions and is related to the noisy-channel approach is the good-enough processing approach (e.g., Ferreira et al., 2002; Ferreira, 2003; Ferreira & Patson, 2007; Goldberg & Ferreira, 2022; Sanford & Sturt, 2002; Traxler, 2014). The good-enough framework posits two routes in language processing: (1) an exact "algorithmic" route and (2) an approximate "heuristic" route. Participants may use one or the other of these routes to process language depending on the goals of the task at hand. Readers may take the heuristic route when processing depth-charge sentences. We leave it to future research to explore potential heuristics that would account for the observed patterns of behavior.

The third alternative account is related to the meaning of *too* and whether it is indeed

ambiguous. As pointed out by our reviewer, *too* might indeed have two interpretations, represented below as (20). In (20a), the extreme degree of being smart is negated. In (20b), the property of being smart is negated. Both are legitimate interpretations of (20). For the depth-charge case, *No head injury is too trivial to be ignored*, the interpretation of *too* could align with (20b) and break the canonical interpretation of *too...to* as "something has an extreme property to the extent that the action cannot be performed". We leave this interesting observation to future studies for more insights.

(20) He is not too smart.

a. He is not extremely smart. For example, he cannot pass the hardest test in the world.

b. He is not smart. For example, he cannot pass a simple test.

3

Comparative Illusion

Sentence (21) may seem acceptable to native speakers of English at first glance, but they often find it difficult to pinpoint the exact meaning when pushed by the experimenter.

(21) More people have been to Russia than I have.

Researchers in psycholinguistics use the term "comparative illusion" (often abbreviated as CI, which we use in the rest of the paper) to refer to the phenomenon that native speak-

ers of English find sentences like (21) acceptable at first while in reality it is semantically ill-formed. The reason for using "comparative" in the terminology is that the semantic ill-formedness of (21) is related to the comparative structure *more...than* in English, which can be manifested as follows. The sentence-final *have* in (21) is an elliptical verb phrase whose antecedent is most naturally understood to be *have been to Russia*, so the sentence should have the same meaning as (22):

(22) More people have been to Russia than I have been to Russia. (semantically ill-formed)

but *I have been to Russia* does not offer the right kind of information – a quantity or degree (Wellwood et al., 2018) – to enter into the comparison. Therefore, the semantic ill-formedness of (22) is inherited by the CI sentence (21).

This CI phenomenon was first mentioned in Montalbetti (1984) and has been studied in various psycholinguistic studies (Christensen, 2010, 2016; Kelley, 2018; Langsford et al., 2019; Leivada & Westergaard, 2020; O'Connor et al., 2013; O'Connor, 2015; Paape, 2024; Pham, 2022; Wellwood et al., 2018; Zhang et al., 2023a). However, a comprehensive explanation is still lacking. In this paper, we propose an account of the theoretically surprising part of the CI effect – that the sentences sound acceptable in the first place – based on the noisy-channel theory of human language comprehension (e.g., Gibson et al., 2013a; Levy, 2008; Shannon, 1948). We hypothesize that during the comprehension of a comparative illusion sentence like (21), readers mentally evaluate not only the literal sentence but also plausible alternative "neighbor" sentences that the speaker might have intended. Since the literal sentence is semantically ill-formed, it may instead be interpreted as one of these alternatives so long as (i) the alternative interpretation is semantically well-formed; and (ii) the alternative interpretation is similar in form to the literal sentence. For sentence (21), for ex-

ample, a relevant alternative sentence might be (23) below:

(23) People have been to Russia more than I have.

which is semantically well-formed and close in form to (21). According to our hypothesis, the CI effect occurs when the availability of well-formed "neighbor" sentences misleads comprehenders into initially finding sentences like (21) acceptable. If this hypothesis is correct, then factors modulating alternative sentences' plausibility and formal proximity to the literal sentence should modulate the magnitude of the CI effect (that is, how likely a native speaker is to find such a sentence acceptable). In this paper, we empirically test this key prediction of our hypothesis, using the mathematical formalization based on noisy-channel principles of Bayesian probabilistic inference.

The structure of the chapter is as follows. In section 3.1.1, we review the linguistic factors that have previously been shown to influence the magnitude of comparative illusion in an acceptability judgment task. In section 3.1.2, we review existing accounts that aim to address how the illusion arises. In section 3.2, we introduce the noisy-channel framework and detail how this theory has potential as an explanation, partly because its prediction encompasses all the relevant findings in previous literature in a cohesive way. In sections 3.3 to 3.6, we report four experiments that display supporting evidence for the noisy-channel account of the comparative illusion. In section 3.7, we conclude with discussion and suggestions for future research.

3.1 Current Landscape

3.1.1 Linguistic factors affecting the illusion effect

Several linguistic properties of the CI sentence have been found to affect its acceptability, including the semantic property "repeatability" of the matrix verb phrase (e.g., *...vacationed in Florida*... vs. *...retired to Florida*...), the number feature of the than-clause subject (e.g., *...I have* vs. *...we have*), and the one-to-many mapping between the matrix subject and the matrix object in sentence variations that differ from the canonical CI sentence (e.g., *cats* and *toys* in *More cats have mouse toys than the dog has* vs. *cats* and *tails* in *More cats have striped tails than the dog has*).

First, in English-based experiments, CI sentences with repeatable verb phrases were found to receive higher acceptability than nonrepeatable verb phrases (Kelley, 2018; O'Connor et al., 2013; O'Connor, 2015; Wellwood et al., 2018). Repeatable verb phrases denote the kind of events that one single individual could participate in multiple times in a relevant time scale (Wellwood et al., 2018). Repeatable verb phrases can be modified by the adverb *more*. As in (24), John can take a vacation in Florida multiple times and the comparative *more* can modify the occurrences, duration, or both of John initiating this event in comparison with that of the event by another individual. In contrast, nonrepeatable verb phrases are often telic perfective predicates that map to an event with an emphasis on its endpoint (e.g., Dahl, 1981; Garey, 1957). Usually, the event occurs only once in a relevant time scale. For example, in (25a), the event of retiring to Florida typically occurs once in one's lifetime (except for exceptional cases where one has multiple jobs and retired to Florida multiple times). Using *more* to modify the verb phrase *retired to Florida* generates the infelicitous sentence

in (25b). Therefore, in the context of CI sentences, repeatable verb phrases like (26a) were rated higher than the nonrepeatable counterpart like (26b) in acceptability judgment tasks.

(24) a. John vacationed in Florida.

b. John vacationed in Florida more than Mary.

(25) a. John retired to Florida.

b. John retired to Florida more than Mary. (infelicitous)

(26) a. More judges vacationed in Florida than the lawyer did. (repeatable, CI variant*)

b. More judges retired to Florida than the lawyer did. (nonrepeatable, CI variant)

Second, when the than-clause subject is a noun phrase instead of a pronoun, e.g.,(27) vs. (21), researchers have found that plural than-clause subjects induced higher acceptability score than the singular counterparts (Christensen, 2010; O'Connor, 2015; Wellwood et al., 2018). In (27), sentences like (27b) were rated higher than sentences like (27a); sentences (27d) were rated higher than sentences like (27c). Few studies looked at the canonical CI sentence with a pronoun than-clause subject except for Wellwood et al. (2018) which found that CI sentences with the first-person plural pronoun *we* (e.g., *More girls ate pizza than we did*) did not receive higher acceptability ratings than the counterparts with the first-person singular pronoun *I*.

(27) a. More judges vacationed in Florida than the lawyer did.

*The following paper uses "CI variant" to refer to sentences that are minimally different from the canonical CI sentence (21) but are still semantically ill-formed. The minimal difference lies in the than-clause subject: whether it is a pronoun, as in the canonical example *More people have been in Russia than **I** have*, or whether it is a noun phrase, as in (26a). Another difference is whether the verb phrase features the present perfect tense or the past tense.

b. More judges vacationed in Florida than the lawyers did.

c. More judges retired to Florida than the lawyer did.

d. More judges retired to Florida than the lawyers did.

(The above sentences are all CI variances.)

Third, O'Connor (2015) went beyond the canonical CI case whose matrix verb phrase tends to describe events and configured the matrix clause to describe states or properties, as in (28). They found that when the matrix subject and the object shared a one-to-many relationship (e.g., in (28b), one cat can have many mouse toys), the corresponding sentence was more acceptable than that with a one-to-one relationship (e.g., in (28a), one cat can only have one striped tail if it has a tail at all).

(28) a. More cats have striped tails than the dog does. (one-to-one mapping, CI variance)

b. More cats have mouse toys than the dog does. (one-to-many mapping, CI variance)

These three factors allude to a comprehension mechanism that explains why the illusion arises. It seems that if a certain linguistic component in a CI sentence can be modified by *more*, or if some simple surface edits of the CI sentence can turn it into a plausible sentence, like ~~*more*~~ *judges vacationed in Florida* <u>*more*</u> *than the lawyer did* (26a), *more judges vacationed in Florida than* ~~*the*~~ *lawyers did* (27b), or ~~*more*~~ *cats have* <u>*more*</u> *mouse toys than the dog does* (28b), that CI sentence can be rated higher. In comparison, the counterpart for each pair of the sentences does not lead to a plausible sentence with the same edit. In (26b), ~~*more*~~ *judges retired to Florida* <u>*more*</u> *than the lawyer did* still sounds like a weird statement that leads to

an unlikely scenario. In (27a), *more judges vacationed in Florida than ~~the~~ lawyer did* is ungrammatical because of the bare singular noun *lawyer*. In (28a), *~~more~~ cats have <u>more</u> striped tails than the dog does* is still weird in that before the *than* clause, one would expect the sentence is discussing the likelihood for the cat species to have striped tails (versus other types of tails); with the upcoming *than* clause that presents just a specific dog, not the dog species, the reconstructed sentence does not provide a coherent interpretation.

Out of these edits string, there could be multiple possible derived interpretations. The first choice could be the comparison of event times or frequencies. For example, in (26a), the interpretation of event comparison is salient because the compared entity is the number of times judges vacationed in Florida versus the number of times the specific lawyer vacationed in Florida. The second choice could be a comparison of the cardinality of entities provided by either the matrix subject or the matrix object. For example, in the edited version of (27b), the compared component could be the number of judges versus the number of lawyers who both took the vacation in Florida. Or in (28b), the number of mouse toys cats usually have is compared with the number of mouse toys the specific dog has. This observation indicates that instead of computing the compositional and literal meaning of the perceived comparative illusion, readers might contemplate alternative representations of the perceived string and derive meaning from the alternatives. The alternatives could be based on the versatile usage of *more* – *more* can not only modify nominal expressions as the comparative morpheme of *many*, but also modify verb phrases as an adverb. We will show later how we relate this observation to a noisy-channel explanation.

3.1.2 Existing processing accounts

In this section, we review previous accounts of the comparative illusion. All accounts mention a latent coercive process that forces a plausible interpretation of the comparative illusion sentence given its semantic incoherence. Another commonality is that they also mention some string edits during language processing.

Sentence template blending account

The sentence template blending account claims that the acceptable judgment decision originates from the integration of two sentences similar to the canonical CI sentence, shown in (29). There, the segments that are not bracketed align with the canonical CI sentence and taking the union of all the words not in the brackets of the two sentences results exactly in the CI sentence (21). This account is based on the analysis-by-synthesis model of sentence processing (Townsend & Bever, 2001). This proposal posits that upon hearing a sentence listeners quickly initiate a hypothesis about its meaning by connecting the content words and phrases according to how these content words and phrases usually form a syntactic template with their statistical distributions. Then listeners generate a more complete syntactic parse to check if the initial lexically and statistically motivated hypothesis is consistent with the syntactically well-formed parse. If these two processes do not match in their structure, listeners will be confused by the sentence.

(29) a. More people have been to Russia than I (could believe).

b. People have been to Russia (more) than I have.

Under this proposal, comprehenders of (21) could quickly initiate two sentence tem-

plates as (29a) and (29b). The claim is that people might accept the CI sentence because it could be an integration of parts of the two sentences in (29). Since the two sentence templates are acceptable, the integrated CI sentence might be deemed acceptable, too. Yet since the account also predicts that a more careful syntactic parse would take place, the comprehenders should be able to notice that the canonical CI sentence in (21) in fact doesn't match either of the two sentences in (29). One piece of work that is missing in this account is how to predict when a careful syntactic analysis would take place.

Event interpretation account

The event interpretation account proposed by Wellwood et al. (2018) argues that according to the surface representation of the CI sentence, the comparative morpheme *more* modifies the nominal matrix subject but during processing, participants interpret the sentence as a comparison between numbers of events. Wellwood et al. (2018) hints at a simple syntactic reanalysis process where *more* is interpreted as the adverb and therefore the sentence that participants are entertaining is actually (30).

(30) ~~More~~ people have been to Russia <u>more</u> than I have.

To empirically gather evidence supporting the movement of *more*, they designed a verbatim recall task where participants were asked to recall and retype a CI-type sentence after being exposed to the sentence and finishing a working memory related task. While they predicted recall errors related to moving *more*, as in (31), they claimed that since the frequency of this error was way lower than errors involving changing the number feature of than-clause subjects (32) and those involving changing the verb phrase (33), they "failed to find an error pattern with moving *more*" (p.571).

(31) a. **More** girls ate pizza than I did. (exposure sentence, CI variance)

b. Girls ate pizza **more** than I did. (recall: *more* as the adverb)

c. Girls ate **more** pizza than I did. (recall: *more* modifies the object)

(32) a. More girls ate pizza than **the boy** did. (exposure sentence, CI variance)

b. More girls ate pizza than **boys** did. (recall: change the number of than-clause subject)

(33) a. More girls **began reading the book** than the boy did. (exposure sentence, CI variance)

b. More girls **read the book** than the boy did. (recall: change the verb phrase repeatability. *This is still an illusion.)

Therefore, their evidence on the comprehension process was not conclusive. On the one hand, Wellwood et al. alluded to a syntactic reanalysis process related to the position of *more*; on the other hand, their recall data failed to show substantial evidence. In addition, the event interpretation account did not account for another possible interpretation that compares the cardinality of two sets of individuals. As in (32), the meaning could be "there are more girls than boys who ate pizza". In sum, Wellwood et al. (2018) provided abundant empirical data but its event comparison theory did not fully predict the data pattern.

Repair analysis

Before Wellwood et al. (2018), O'Connor (2015) also proposes a similar repair analysis account. Under this account, comprehenders can flexibly generate various perceptions of the given input by making small changes and turn the input into a plausible sentence. Her

possible edits include (i) moving *more* to achieve an adverbial reading and (ii) repair the form of than-clause noun phrase. In addition to the event comparison interpretation, she claimed that the sentences could also be interpreted as comparing the cardinality of two groups of individuals (34b & c). Sentences in (34) show a mapping between possible edits and the corresponding interpretations.

(34) a. More judges vacationed in Florida than the lawyers did. (CI variance)

b. Judges vacationed in Florida more than the lawyers did. (*more* moved, event comparison interpretation)

c. More judges vacationed in Florida than ~~the~~ lawyers did. (*the* deleted, individual comparison interpretation)

d. More of the judges vacationed in Florida than the lawyers did. (insertion of *of the*, individual comparison interpretation)

The proposal in O'Connor (2015) provides clear hypotheses about the comprehension mechanism regarding the edits and the associated interpretation. Yet the strongest limitation was that she did not provide empirical evidence to support the repair analysis, despite mentioning this in her discussions.

Rational Inference

Paape (2024) proposes that the interpretation of comparative illusion sentences involves rational inferences. In a judgment task, participants were asked to read a sentence and make a choice from five options (a = "Get it, correct", b = "Get it, but incorrect", c = "Don't get it, (probably) correct, d = "Don't get it, incorrect", e = "No idea") ("get it" means that the participant claims that he was able to understand the intended meaning; "correct" means that

the participant thinks that the sentence is formally correct). The results show that around 50% of the responses chose "b", indicating that participants could recognize the errors but might infer a plausible meaning. Around 25% of the responses were "a", which could be taken as direct evidence for an illusion effect. The other 25% chose "d", which indicates difficulty of comprehension. These results were thus taken to support an underlying rational inference procedure.

Even though Paape (2024) does not provide further details about the processing mechanism, it points to a promising direction and could find resonance in both the event interpretation account by Wellwood et al. (2018) and the repair analysis account by O'Connor (2015). That is, even though participants might find the CI sentence formally anomalous, they are able to reconcile with the meaning via possible inference. With these claims foreshadowing a comprehension mechanism involving repairs and inference, we introduce the noisy-channel framework as a computational explanation of the comparative illusion.

3.2 New Approach

Within this noisy-channel framework introduced in section 1.2, we investigate whether the comprehension of the comparative illusion could be modeled as Bayesian inference over a noisy channel. We model the sentence *More people have been to Russia than I have* as s_p because from the review above we know that comprehenders usually arrive at an alternative meaning different from the sentence's literal meaning. We model the comprehender's inferred meaning as s_i. We hypothesize that there could be several plausible alternative messages s_is intended by the producer or inferred by the comprehender given the perceived CI sentence. One could lead to an event comparison interpretation, as pioneered by Well-

wood et al. (2018); another could be an individual comparison interpretation, as discussed in O'Connor (2015). There could be others undiscovered yet (see results in Experiment 2 and suggested by cross-linguistic evidence in Christensen (2016)). Each of the intended s_is would have a different noise corruption profile that transforms them to the corrupted CI sentence s_p. For example, for the event comparison s_i "people have been to Russia more than I have", the noise edit could involve moving *more**. According to Equation 1.1, we predict that the more likely the inferred interpretation is to be encapsulated into the comparative illusion sentence (i.e., with a higher noise likelihood term $P(s_p|s_i)$), the more likely it is to be inferred as the intended message (i.e., with a higher posterior $P(s_i|s_p)$), assuming an equal prior across these messages.

Our noisy-channel explanation of comparative illusion integrates existing proposals and adds a critical step in advance. Similar to the Event Comparison account by Wellwood et al. (2018) and the Repair Analysis account by O'Connor (2015), it acknowledges that the nature of the illusion effect is that comprehenders arrive at a semantically well-formed interpretation that deviates from the literal composition of the CI sentence (21). On the surface level, comprehenders "get" what (21) could mean (Paape, 2024). Since the noisy-channel account does not restrict the number of possible alternatives comprehenders could entertain, this also aligns with the intuition of O'Connor (2015). Crucially, it explicitly connects the multiple possible interpretations with justifiable noise edits or structural changes during language comprehension. Most importantly, this framework also makes testable ordinal predictions of which interpretation is more likely (i.e., $P(s_i|s_p)$) based on the noise edit likelihood (i.e., $P(s_p|s_i)$) and the semantic prior of this interpretation (i.e., $P(s_i)$), based on the

*Or in another conceptualization of the noise model, the noises could be the deletion of the sentence-initial *more* and the insertion of the adverbial *more*.

mathematical formalisms of the rational language processing (Equation 1.1).

We conducted four experiments in our research to support the noisy-channel prediction. Experiment 1 features an acceptability judgment task where we identified the source of illusion by varying critical components of the sentences that have been claimed to affect sentence acceptability. We found that CI-type sentences with a pronoun than-clause subject (e.g., *More students have been to Russia than I have*) received acceptability scores as high as grammatical controls while sentences with the noun phrase than-clause subject experienced a degradation in acceptability (e.g., *More students have been to Russia than the teacher has*). This pattern motivates us to only look at pronoun cases in our following experiments as they exhibit a strong illusion effect.

Experiment 2 was a paraphrasing task that gauged native English speakers' actual interpretation of the comparative illusion sentences with pronoun than-clause subjects. After annotation and calculation, we selected predominant paraphrases as the series of s_is. They featured an event comparison interpretation (e.g., "Students have been to Russia more than I have" for (21)), an individual comparison interpretation (e.g., "There are more students who've been to Russia than just me"), and an event negation interpretation (e.g., "(Many) students have been to Russia but I have not"), respectively.

Experiment 3 featured a forced-choice task that juxtaposed these three most prominent interpretations from Experiment 2 and asked participants to select the most prevalent reading given a certain comparative illusion sentence. The percentage of selection for each interpretation was taken as the proxy for the posterior distribution of $P(s_i|s_p)$. The results show that the event comparison interpretation was the most frequent choice, followed by individual comparison and event negation.

Experiment 4 was a noise likelihood rating task following Zhang et al. (2023c), which aimed to obtain a behavioral proxy for the noise likelihood term $P(s_p|s_i)$. It asked how likely each possible interpretation out of the three was to be produced as the perceived comparative illusion. We found that the noise likelihood rating followed the order of event comparison, individual comparison, and event negation, which is the exact order of the posterior proxy in Experiment 3. Assuming that each intended message has a similar prior probability given each is very probable in the context of, for example, discussing one's trip to Russia, we conclude that the empirical evidence we found in our experiment is consistent with the predictions of a noisy-channel explanation of the comparative illusion.

3.3 Experiment 1: Illusion Effect

One of the limitations of previous research is that the experimental materials did not contain the exact structures in the canonical CI sentence (21) discussed in Montalbetti (1984). While the canonical CI sentence has present perfect tense and aspect (e.g., *have been to Russia*), the tested materials in previous studies (O'Connor, 2015; Paape, 2024; Wellwood et al., 2018) usually had past tense (e.g., *vacationed in Florida*). Given this, it is important to check whether this change affects the illusion pattern observed in previous studies and whether the critical linguistic factors in section 3.1.1 still impact the acceptability score. We conducted two subset experiments to tackle this issue.

The two subset experiments varied by the than-clause subject format: Experiment 1A featured a first-person pronoun than-clause subject and Experiment 1B featured a noun phrase than-clause subject. Both Experiment 1A and 1B manipulated the repeatability of the verb phrase and the number feature of the than-clause subject, as previous studies did.

3.3.1 Participants

In Experiment 1A, 49 participants were recruited from the crowd-sourcing platform Prolific.com. Each participant was paid $4 for their participation. We excluded data from those (a) who did not complete at least 90% of all questions; (b) who did not answer at least 75% of the comprehension checks correctly; (c) who gave the same rating across all test trials; and/or (d) who self-identified as non-native speakers of English or not born in the United States. We analyzed the remaining 45 participants' responses.

In Experiment 1B, another group of 48 participants were recruited from the Prolific. After the same screening procedure, 41 participants remained for the formal analysis.

3.3.2 Materials & Procedure

Both experiments were offline acceptability judgment tasks. The critical materials featured a 2x2 manipulation with two control conditions. The first manipulation was the repeatability of the verb phrase where repeatable verb phrases describe events that could usually take place multiple times given a relevant time frame but nonrepeatable verb phrases describe events that are odd if occurring multiple times. For example, in the context of expressing one's tourist experiences, it is natural to say "I've been to Russia multiple times"; but it is harder to find a context where "I've escaped from Russia multiple times" is natural. The second manipulation concerned the number feature of the than-clause subject: Experiment 1A compared the singular first-person pronoun *I* versus the plural first-person pronoun *we*; Experiment 1B compared singular determiner noun phrases (e.g., *the teacher*) versus the plural counterpart (e.g., *the teachers*).

Table 3.1 shows an example item for Experiment 1A. We constructed the control condi-

tions where more is used as an adverb to modify the times or frequency of the event referred to by the verb phrase. The good control condition featured a plausible sentence where the main verb phrase was a repeatable event and was natural to be modified by more. The bad control condition featured an implausible sentence where it was not natural to combine nonrepeatable events with more. We designed 30 critical items with such design.

Table 3.1: Example material for Experiment 1A

Pronoun as than-clause subject	**Singular subject**	**Plural subject**	**Control**
Repeatable verbs	More students have been to Russia than I have.	More students have been to Russia than we have.	Many students have been to Russia more than I have. (good)
Nonrepeatable verbs	More students have escaped from Russia than I have.	More students have escaped from Russia than we have.	Many students have escaped from Russia more than I have. (bad)

Experiment 1B had a similar experimental design. As shown in Table 3.2, the 2 x 2 manipulation crossed the repeatability of the verb phrase and the plurality of the than-clause subject. Different from Experiment 1A, the control conditions were both featured by plausible sentences because the component being compared was the cardinality of the sets of individuals denoted by the matrix subject and the than-clause subject.

Experiment 1A and 1B shared the same set of filler items. We designed 64 filler items which contained comparative structures or various types of quantifiers. The following constructions appeared four times each: *a few*, *fewer*, *more... than*, *more than*, *many*, *much*, *little*, *less*, *some*, *any*, *a lot of*, *no*, *plenty of*, *enough*, *none*, *all*. All filler items were grammatical and plausible sentences. An example is *Fewer people have visited to Greenland than Canada*.

Table 3.2: Example material for Experiment 1B

NP as than-clause subject	Singular subject	Plural subject	Control
Repeatable verbs	More students have been to Russia than the teacher has.	More students have been to Russia than the teachers have.	More students have been to Russia than teachers have. (good)
Nonrepeatable verbs	More students have escaped from Russia than the teacher has.	More students have escaped from Russia than the teachers have.	More students have escaped from Russia than teachers have.

The procedures for Experiment 1A and 1B were also identical. Each participant read a list of 94 sentences in a randomized presentation. Each trial was followed by a YES/NO comprehension question (e.g., "Does this sentence mention students and Russia?") that encouraged participants to pay more attention to the content of the materials. The answer to the comprehension question was counterbalanced such that half critical trials were "Yes" and half were "No". Then participants were asked to rate "How natural is the sentence?" on a 7-point fully labeled Likert scale (1 = Extremely unnatural, 2 = Unnatural, 3 = Somewhat unnatural, 4 = Neutral, 5 = Somewhat natural, 6 = Natural, 7 = Extremely natural).

3.3.3 Results

Figure 3.1* shows the standardized acceptability rating for sentences with a pronoun than-clause subject in Experiment 1A. It seems that the good control condition and the bad control condition set an upper bound and a lower bound for the acceptability ratings across

*Dots represent the mean; the error bars represent 95% bootstrapped confidence interval; participant number 45

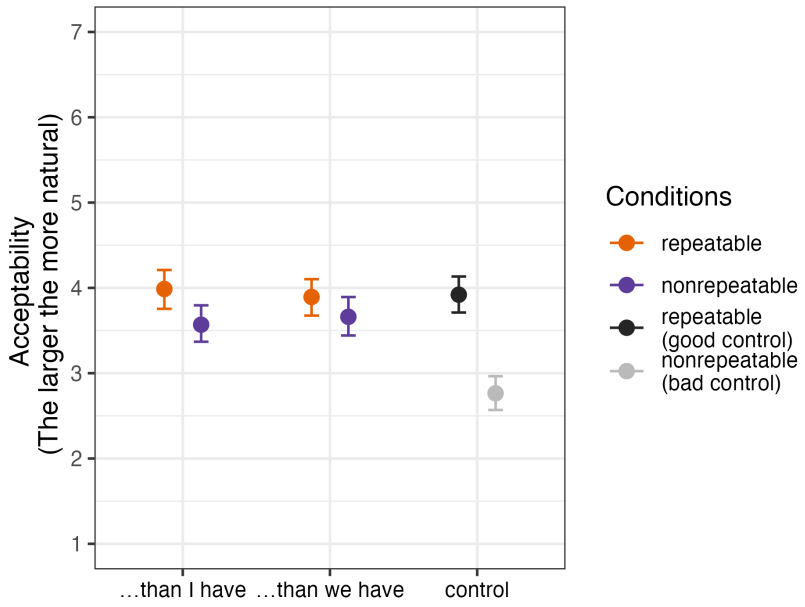


Figure 3.1: Acceptability ratings with pronoun than-clause subjects

the six conditions. A visual inspection shows that the CI sentences in the four critical conditions were rated similarly to the good control condition, with slight degradation in the nonrepeatable verb phrase conditions. It is not easy to tell whether the number feature of than-clause subject plays an important role here.

We ran multiple statistical models using the *brms* package (Bürkner, 2017; Bürkner & Vuorre, 2019) in R to test whether the verb phrase repeatability and the number feature of the than-clause subject affect the acceptability rating. First, we tested whether the mean of the four critical conditions received different ratings than the good control. If the critical materials as a whole set did not differ from the good control, we get closer to claim that the semantically ill-formed sentences in the critical conditions induced a strong illusion effect. All the data points excluding the bad control were included in the model. The dependent variable was the raw trial-level rating score with non-equidistant intervals between levels

on the Likert scale; the fixed effect was a binary-coded categorical variable with two levels (control=1 vs. critical materials = 0). The number of data points belonging to the critical condition was four times the number of data points in the good control condition. Random intercepts and slopes for the categorical variable for both subjects and items were included as random effects to obtain the maximal random effect structure for mixed effects models (Barr et al., 2013). *Brms* default priors were used for all the parameters. The model had four sampling chains each with 4000 iterations. The first 2000 samples were taken as warm-up. An $\hat{R}$ close to 1.0 marks the convergence of the sampling chain to the underlying posterior distribution of the target predictor. In this paper, we use β to represent the posterior mean of the distribution of the estimated coefficients for predictors and CrI to represent the 95% credible interval. The model shows that all $\hat{R}$s for the sampling chains for all fixed effects were 1.0, indicating successful convergence. The point estimate for the critical condition versus the control condition reference shows that on the latent acceptability scale, items in the four critical conditions were 0.15 SD more acceptable than the items in the good control condition. The 95% CrI of this parameter was between -0.15 and 0.44, including zero. We can conclude that the good control sentences didn't have a substantially different acceptability rating from the four critical conditions.

The second model compared the critical items in four conditions with the bad control sentences. Using the same model structure, we took the bad control condition as the reference in a dummy-coded categorical variable and contrasted the reference with the critical conditions. The posterior sampling chains successfully converged. Based on the point estimate of the mean of the posterior distribution of the coefficient and its 95% CrI, we found that the bad control sentences received substantially lower ratings than the critical sentences

(β=-1.04, CrI=[-1.38, -0.73]). The fact that the semantically ill-formed sentences in the four conditions were rated similarly as the good control and higher than the bad control indicates a very strong illusion effect for sentences with a first-person pronoun than-clause subject.

We further tested whether the repeatability of the verb phrase and the plurality of the than-clause subject affected the acceptability rating. Using the leave-one-out cross-validation method (Bürkner & Vuorre, 2019; Vehtari et al., 2017), we chose the statistical model without the interaction between the two critical manipulations as a better fitted model. It was also a multilevel cumulative ordinal model that took the raw rating score as the dependent variable, the dummy-coded categorical variables of verb repeatability (reference level = nonrepeatable) and than-clause subject plurality (reference level = singular) as the fixed effects, and random intercepts plus slopes for the full fixed structure for subjects and items. Repeatable verb phrases received substantially higher acceptability ratings than the nonrepeatable ones (β=0.29, CrI=[0.11, 0.48]). There was no significant difference between the singular subject I and the plural subject we (β=-0.02, CrI=[-0.17, 0.13]).

Figure 3.2* shows the acceptability rating patterns in Experiment 1B with noun phrases as the than-clause subject. It is obvious that (i) nonrepeatable verb phrases received lower ratings compared with the repeatable verb phrases, (ii) singular than-clause subjects received lower ratings than the plural than-clause subjects, and (iii) sentences in all four critical conditions received lower ratings compared to the control conditions with acceptable sentences.

Experiment 1B utilized a slightly different statistical testing procedure compared to Experiment 1A. We plugged all six conditions in Table 3.2 into a single Bayesian model: the dependent variable was the raw acceptability rating score; the dummy-coded verb repeatabil-

*Dots represent the mean; the error bars represent 95% bootstrapped confidence interval; participant number 41.

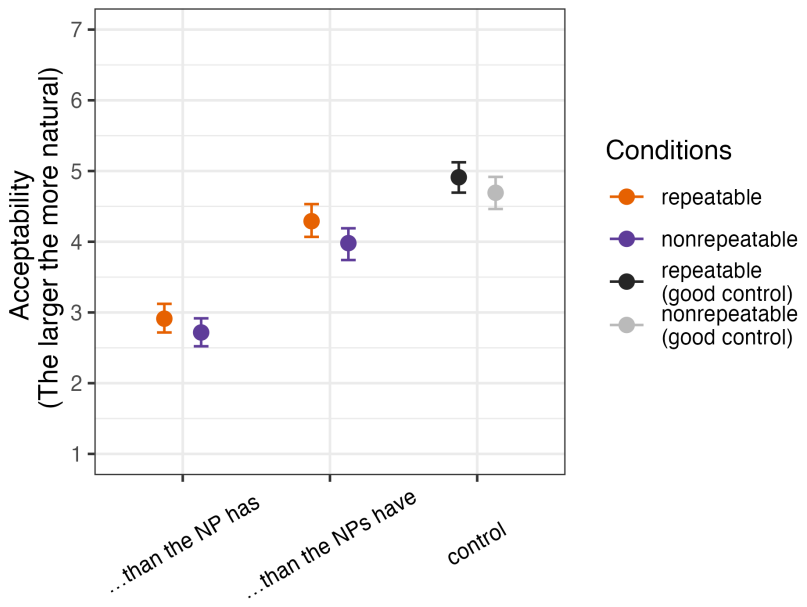


Figure 3.2: Acceptability ratings with noun phrase than-clause subjects

ity (reference level = nonrepeatable) and the number feature (reference level = control) were taken as the fixed effect; the random effects also obtained the maximal structure for both the subjects and the items. Based on the same leave-one-out cross-validation method, we chose the model without the interaction term between the two main effects. Based on post hoc analysis that compares the marginal means across different conditions via the *emmeans* package (Lenth et al., 2019), we found that there was a decreased acceptability along the scale of control – plural – singular condition (control vs. plural, β=0.62, HPD=[0.34, 0.87]; control vs. singular, β=1.84, HPD=[1.47, 2.20]; singular vs. plural, β=-1.22, HPD=[-1.53, -0.93]). Nonrepeatable verb phrases received lower ratings than repeatable verb phrases (β=-0.219, HPD=[-0.412, -0.041]).

3.3.4 Discussion

Experiment 1 provides a picture of the acceptability landscape of CI sentences. We see that CI sentences with a pronoun than-clause subject triggered a stronger illusion effect than those with a noun phrase than-clause subject since the former was rated even similarly to the grammatical and plausible controls but the latter received apparently lower ratings than the good controls. Furthermore, the repeatability of verb phrases played a major role in modulating the acceptability ratings across critical items, replicating the findings in previous studies (e.g., O'Connor, 2015; O'Connor et al., 2013; Wellwood et al., 2018). That the plurality of the than-clause subject was significant in the noun phrase case but not in the pronoun case also replicates the findings in Wellwood et al. (2018). Since the pronoun case—the case for which the CI was originally reported—elicited the strongest illusion effect, the following experiments will focus on providing an explanation for these specific sentence types.

3.4 Experiment 2: Probing Salient Interpretations

It is imperative to identify the range of possible interpretations before we sketch out the inference mechanism. Yet so far, no study has elicited English speakers' actual interpretations of the canonical comparative illusion sentences. For Danish, Christensen (2016) elicited listener interpretations of CI sentences using a paraphrase paradigm. The English equivalents of their prevalent interpretations are in (35), which features an "excluding me" reading (35a), a "more people than just me" reading (35b), and an event comparison reading (35c). The difference between (35a) and (35b) is whether the speaker has been to Russia or not. The salience of these three interpretations follows the order in (35).

(35) a. Some people have been to Russia except me.

b. More people have been to Russia than (just) me.

c. Some people have been to Russia more (often) than I have.

To be noted, the morphosyntactic characteristics of the comparative structure in Danish are different from those in English. In English, *more* could be used as the comparative for the quantifier *many* and as the comparative for the adverb *much*. But in Danish (36), *flere* ("more") is unambiguously used as the comparative for the nominal quantifier *mange* ("many") while *mere* is the adverbial comparative for *meget* ("much").

(36) a. **Flere** folk har været i Rusland end jeg har. (illusion)
More people have been in Russia than I have
"More people have been to Russia than I have."

b. Folk har været i Rusland **mere** end jeg har
People have been in Russia **more** than I have
People have been to Russia more than I have.

The difference in the syntactic distribution of *more* in Danish and English might modulate the distribution and the salience of possible interpretations of the CI sentence. Therefore, it is not feasible to directly transfer the availability of these readings as well as their order from Danish to English. A paraphrase study in English is necessary.

3.4.1 Participants

60 participants who self-identified as native speakers of English were recruited from Prolific. Each received a $2 payment for their participation. Only those who finished at least 90% of all questions and answered at least 90% of the comprehension questions correctly were included in the analysis, resulting in 49 participants all together.

3.4.2 Materials & Procedures

The critical materials were the 30 items that were extracted from Experiment 1 and that featured repeatable verb phrases with a pronoun than-clause subject (e.g., *More students have been to Russia than I/we have*). The only between-subjects condition manipulation was the plurality of the pronoun (*I* vs. *we*). We chose this subset of sentences from Experiment 1 because they elicited the strongest illusion effect and thus form the optimal study materials for us to understand the underlying processing mechanism of CI.

To mitigate the learning and fatigue effect and to give the participants a focused time for paraphrasing, we made two lists, each containing 15 items. Participants only read and paraphrased 15 sentences in the session. For each sentence, the pronoun than-clause subject was randomly chosen between *I* and *we*. No filler items were included in this experiment.

During each trial, the paraphrase instruction was "What does this sentence mean exactly?". After the paraphrase was provided, a YES/NO comprehension question followed to encourage participants to be more attentive (e.g., "Does this sentence mention COVID and vaccines?"). The answers to the comprehension question were counterbalanced.

3.4.3 Results & Analysis

The analysis took place in two steps: paraphrase coding and statistical summarization. The first step involved a manual coding of each paraphrase. The two authors of the paper Yuhan Zhang and Carina Kauf went through 870 paraphrase trials and assigned interpretable labels. They first did a pilot coding session with 100 trials and came up with their own labels. They convened and settled on a coherent coding schema and then they finished the rest of the paraphrases. The inter-rater reliability for all trials was 90.32%.

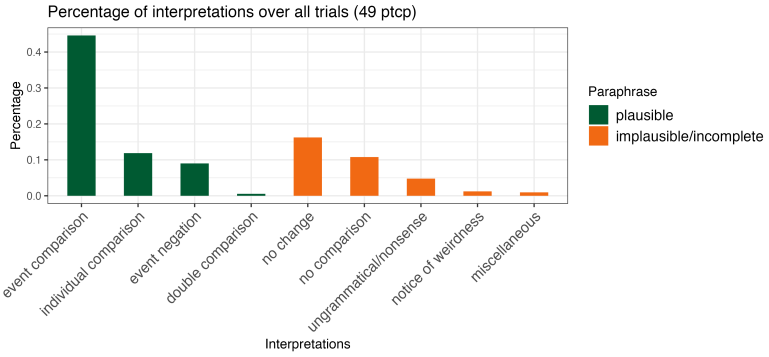


Figure 3.3: Percentage of different categories of interpretations

Figure 3.3 displays the percentage of different categories of interpretations across all trials. The plausible categories are color coded in green and the implausible or incomplete categories are in orange. All together there are nine categories out of which three types emerged as prevalent plausible ones. We labeled them as "event comparison" (41.65%), "individual comparison" (11.34%) and "event negation" (8.35%). It is apparent that the event comparison interpretation was the most prevalent.

For the event comparison interpretation, *more* was used to compare the frequency or times of events occurring for the two groups (e.g., "Students have been to Russia more (times) than I have" for *More students have been to Russia than I have*) or the degree of stative verb phrase (e.g., "We didn't like Harry Potter as much as most middle school students" for *More middle school students have enjoyed Harry Potter than we have*). For the individual comparison interpretation, *more* was used to compare the cardinality of different groups (e.g., "There were more people who have been to Russia than just me"). For the event negation interpretation, participants inserted a contrastive reading by conveying that the than-clause subject hasn't participated in the events that have been experienced by the

matrix subject. For example, "Many people have been to Russia but I haven't" for the canonical Russia example. Or, "the speaker most likely did not talk about the 2022 recession" for *More economists have talked about the recession in 2022 than I have*. The double comparison interpretation constituted a smaller proportion (1.07%) where the corresponding interpretation for the Russia case is "more people have been to Russia more than I have". Or "more Engineers travel to San Francisco more regularly" for *More engineers have travelled to San Francisco than I have* indicating that people might apply *more* twice during the interpretation.

There were five categories of implausible or incomplete paraphrases: "no change" (20.64%), "no comparison" (10.86%), "ungrammatical/nonsense" (3.94%), "notice of awareness" (1.19%) and "miscellaneous" (0.95%). "No change" means the syntactic structure of the paraphrase was almost the same as the original CI sentence. "No comparison" means in the paraphrase the comparative structure with more was dropped. "Ungrammatical/nonsense" means the sentence could not be interpreted either due to a grammatical error or a semantic incoherence. Nor did the sentence preserve the same structure which made it qualified for the category of "no change". "Notice of weirdness" means in the paraphrase the participant mentioned that the original material was weird. "Miscellaneous" encompasses all other kinds of paraphrases.

Figure 3.4 displays the distribution of each paraphrase type for singular versus plural than-clause subjects. For both conditions, the event comparison interpretation was ranked the first, followed by "event negation" and "individual comparison". There might be an interaction between the proportions of "event negation" and "individual comparison" along with the than-clause subject plurality, but a well-controlled experiment is needed to make

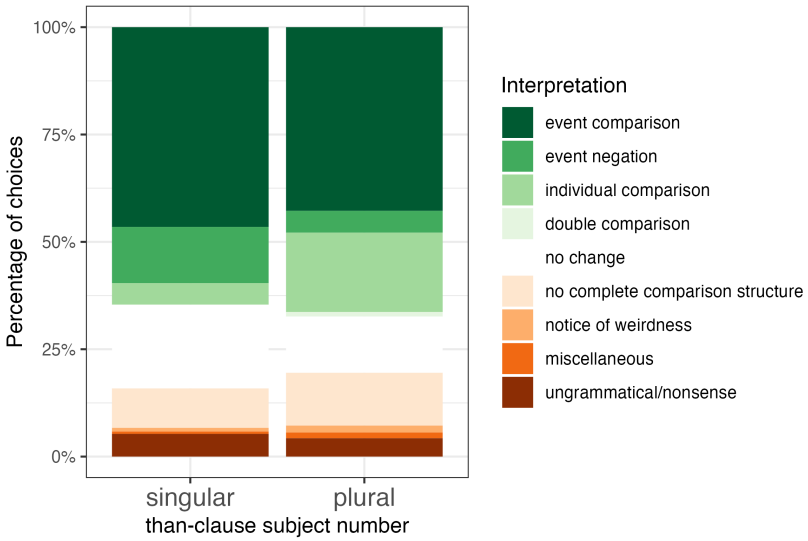


Figure 3.4: Distribution of CI interpretations between singular and plural pronoun than-clause subjects.

sure that this interaction can be replicated. Figure 3.5 represents the distribution of by-item paraphrase categories. The prevalence of the "event comparison" is robust across all items*.

3.4.4 Discussion

Experiment 2 presents a comprehensive paraphrase task that explores native English speakers' interpretation of comparative illusion sentences. The results confirm the existence of multiple plausible interpretations and highlight that the event comparison interpretation is the most prevalent among others. This broad range of interpretations is consistent with

*There are two exceptions – "go scuba diving" with *I* and "visit Universal Studio" with *we*. For the scuba diving case, event negation trumps event comparison, which makes sense because going scuba diving is not a common activity and it is more likely that the participant (the referent of *I*) has not tried it than s/he has tried multiple times. For the activity of "visiting Universal Studio", one possibility is that visiting Universal Studio is a common leisure taken by many people and it is easier to see the emphasis that more people have visited Universal Studio than just us. Still, these are pure *ad hoc* speculations with a possible confirmation bias. More rigorous testing is needed to study the item effect on the prevalence of different interpretations.

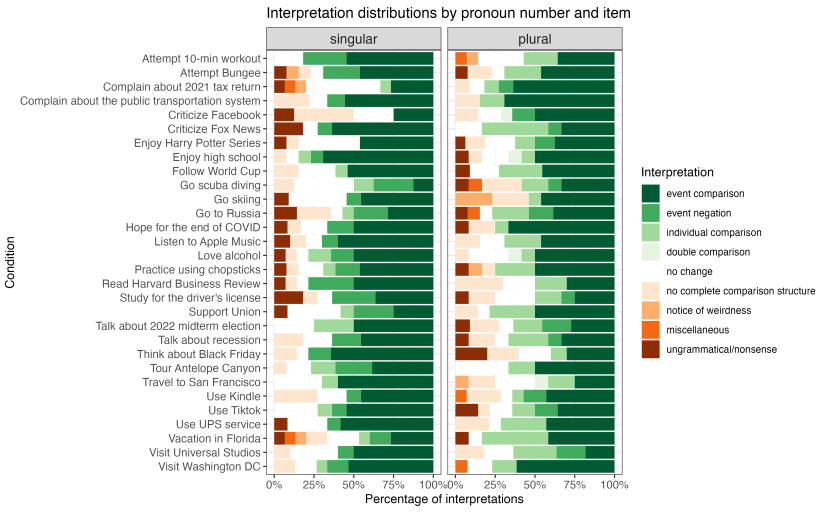


Figure 3.5: Distribution of CI interpretations by than-clause subject plurality and item

O'Connor (2015)'s prediction and pushes against the event comparison proposal by Wellwood et al. (2018).

Furthermore, the fact that more than half of the answers led to plausible interpretations supports the hypothesis that participants were making rational inferences over the comparative illusion sentence, which is consistent with (Paape, 2024).

Because the event comparison interpretation, the individual comparison interpretation, and the event negation interpretation were the emerging representative plausible paraphrases, we take these three as the plausible interpretations, s_is, for Experiments 3 and 4.

Interestingly, the interpretation distribution in English also differs from the distributions in Danish (Christensen, 2016). In English, the event comparison interpretation had a higher percentage than the individual comparison and the event negation interpretations. In Danish, the order was reported to be the opposite (i.e., event negation > individual comparison > event comparision). This difference might be caused by the heterogeneous materials in the

two experiments or an inherent cross-linguistic variance in the comprehension behavior of the comparative illusion. It would be an interesting direction for future investigation.

3.5 Experiment 3: Measuring the Posterior

The noisy-channel theory predicts that the posterior probability of inferring a certain interpretation given the perceived CI sentence is proportional to how likely that interpretation is encapsulated into the perceived sentence. In Experiment 2, three prominent paraphrases of the comparative illusion sentences emerged: "event comparison", "event negation", and "individual comparison" interpretations. In Experiment 3, we constructed a comprehension task, presented each illusion sentence with all three interpretations in a forced-choice multiple-choice presentation, and explicitly asked participants to choose the interpretation that is closest to the illusion sentence. Here, the illusion sentence can be treated as s_p and the three interpretations as s_i. The percentage of each interpretation approximates a distribution of s_is given s_p. We took this distribution as a proxy for $P(s_i|s_p)$.

3.5.1 Participants

60 participants who self-identified as native English speakers finished the task from Prolific. Each was paid $6 for their participation. We analyzed those who have passed all the English sentence completion tasks, finished at least 90% of all trials, correctly answered the comprehension task for at least 75% of the times, and self-identified as from the United States. 36 participants were included in the final analysis*.

*The main reason for such a low admission rate was that a proportion of the "individual comparison" fillers received an "event comparison" interpretation. For example, *More senior citizens than young adults have undergone surgeries* was interpreted as "senior citizens have undergone surg-

3.5.2 Materials & Procedure

To get a behavioral proxy for the posterior distribution, we featured a forced-choice selection task. We presented to native speakers of English the illusion sentence as s_p and the three interpretations as s_i. We asked the participants to read the illusion sentence and choose which interpretation best matches their initial interpretation. There was also a fourth option "None of the three interpretations works. This sentence does not make sense to me."

The 30 critical items were the same as Experiments 1 and 2 where the verb phrase was repeatable and the number of than-clause subject varied between *I* and *we*. Each was paired with three interpretations and one option to indicate "this sentence doesn't make sense to me". Figure 3.6 shows one example trial in the experiment. We made sure that each of the three interpretations had a consistent syntactic structure across all items. The "event comparison" interpretation treated *more* as an adverb (e.g., "Students have been to Russia more than I/we have"). The "individual comparison" interpretation featured a "there be" construction; *more* modified the cardinality of the subject group; the following "than just me/us" highlighted the reading that there are more people who have experienced the event than just the speaker or the group that the speaker belongs to (e.g., "There are more students who have been to Russia than just me/us"). The "event negation" interpretation used the coordinator *but* to connect two clauses and expressed that the speaker (group) hasn't experienced the event at stake (e.g., "Students have been to Russia, but I have not").

There were 60 filler items which were all grammatical and plausible sentences. They also had four corresponding choices three of which featured "event comparison", "individual

eries more than young adults have." The supposed comparison between the cardinalities of the two groups was interpreted as the frequency/times of events, which is not appropriate given the data inclusion standard.

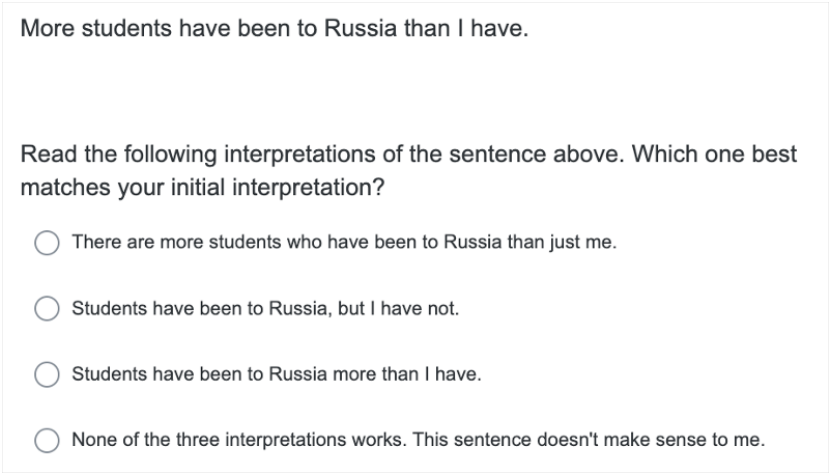


Figure 3.6: An example trial for Experiment 3

comparison" and "event negation" interpretations. Out of the 60 items, 20 had the correct answer as "event comparison" (e.g., College students have been listening to podcasts more than I have); 20 had "event negation" (e.g., Youtubers have earned money with their videos but we have not); 20 had "individual comparison" (e.g., More computer scientists than philosophers have collaborated with colleagues). Within the "event comparison" and "event negation" group, 10 items had the first-person pronoun *I* as the than-clause subject and 10 items had *we*. Within the "individual comparison" group, both the matrix subject and the than-clause subject were plural noun phrases. The syntactic templates for the four options were the same as those for the critical items.

Each participant finished a randomized list of 90 forced-choice tasks. The order of choices was also randomized across all trials. We only analyzed data from participants who picked the correct interpretation for the filler items more than 75% of the time. At the beginning of the task, participants finished five English sentence completion tasks. We relied on their input to filter out participants whose English skills do not reveal native proficiency.

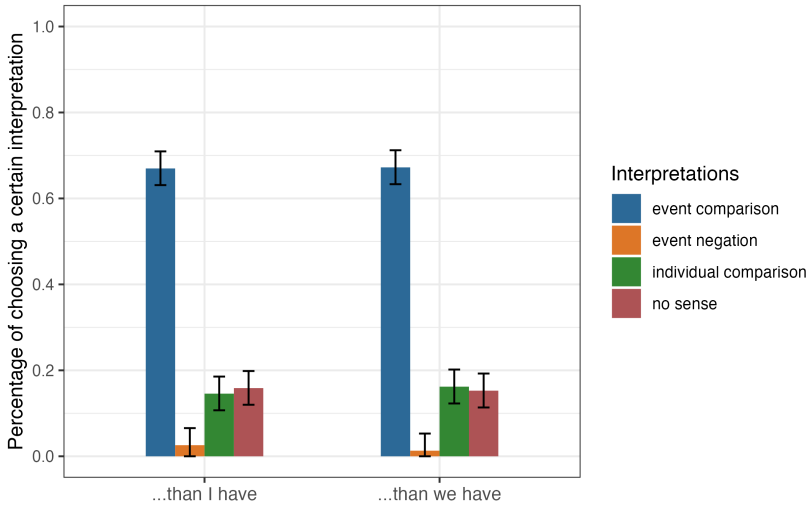


Figure 3.7: The percentage distribution of choosing different interpretations

3.5.3 Results

Figure 3.7 represents the percentage of different interpretations for singular and plural first-person pronoun than-clause subjects. In both conditions, the event comparison interpretation is the most predominant followed by individual comparison and event negation.

To statistically investigate whether the number feature of the than-clause subject played a role in the probability of choosing any of the four interpretations, we fitted the forced-choice data into a Bayesian multilevel categorical regression model. The response variable was the four categories of interpretations ("event comparison", "event negation", "individual comparison", and "no sense"). The independent variable was the dummy-coded categorical variable representing the number feature of the than-clause subject (reference level = "singular"). The random effects included a random intercept and a random slope for the number feature condition for both items and subjects. The model family was set to be categorical which was specifically used when the response variable was categorical with more

than two levels. There were 4 sampling chains each with 2000 iterations to generate the posterior distributions of the estimated parameters. The prior setting was the default. All $\hat{R}$s were 1.0, indicating successful convergence. We relied on Bürkner (2024) to interpret the model output.

First of all, there were substantial differences between the probability of the event comparison interpretation and each of the other three interpretations. The probability of choosing the individual comparison interpretation was 0.08 times the probability of choosing the event comparison interpretation (β = -2.56, CrI = [-3.81, -1.49]). The probability of choosing the event negation interpretation was 0.008 times the probability of choosing the baseline event comparison group (β = -4.87, CrI = [-6.93, -3.43]). The probability of choosing nonsense was 0.02 times the probability of event comparison (β = -3.97, CrI = [-6.26, -2.12]). Second, having a plural than-clause subject did not make any interpretation more or less likely to happen. In other words, there was no substantial effect of the number feature of the than-clause subject on the probability of all the interpretations, as the 95% credible intervals for the estimate coefficients of plural subjects within each interpretation contained zero (plural & individual comparison: β = 0.17, CrI = [-0.43, 0.92], plural & event negation: β = -2.03, CrI = [-6.12, 0.14], plural & no sense: β = -0.51, CrI = [-1.81, 0.50]).

3.5.4 Discussion

Experiment 3 features a forced-choice selection task that explicitly collected native English speakers' choice over the most salient interpretation of the CI sentence out of three plausible alternatives. We found that the plurality of the pronoun than-clause subjects did not affect the distribution of the three interpretations. The "event comparison" interpre-

tation was the most dominant, followed by "individual comparison" and then "event negation". In other words, if we take this distribution as a proxy for the posterior probability $P(s_i|s_p)$, the posterior probabilities follow an ordered rank from "event comparison" to "individual comparison", and to "event negation". What is more, the "nonsense" option received a similar percentage to "individual comparison". The dominance of "event comparison" is consistent with findings in Experiment 2. The existence of other possible interpretations also correspond with claims in O'Connor (2015) that multiple interpretations exist for CI sentences.

3.6 Experiment 4: Measuring the Noise Likelihood

Since the noisy-channel framework predicts that the posterior probability of choosing one interpretation out of many is modulated by how likely that interpretation is produced or encapsulated into the perceived sentence, and since Experiment 3 collected a proxy for the posterior, Experiment 4 was set up as a noise likelihood task to collect a behavioral proxy of $P(s_p|s_i)$. We asked participants how likely it is for a speaker to produce the perceived corrupted comparative illusion sentence like (21) while the speaker intends to utter one of the three plausible intended messages (which we know from Experiment 2).

3.6.1 Participants

63 participants were recruited from Prolific. Each received $6 payment for their participation. We analyzed those who self-identified as native speakers of English from the United States, whose English completion materials were grammatical and plausible, who passed the comprehension check questions more than 75% of time for all relevant trials, and who fin-

ished at least 90% of all trials. 59 participants contributed to the final analysis.

3.6.2 Materials & Procedure

The noise likelihood rating task followed the design in Zhang et al. (2023c) where participants were given a pair of sentences (the first sentence was labeled "intended" and the second "produced") and participants were asked how likely it is for a speaker to utter the "produced" sentence while intending to utter the "intended". Zhang et al. (2023c) showed that this methodology generates reliable approximations for the noise likelihood term.

In this study, we varied the intended sentence that paired with the produced sentence in three conditions: "event comparison", "individual comparison", and "event negation". The critical materials also differed in whether the than-clause subject was a singular or plural first-person pronoun. They were structured as (37) and (38). The intended sentences were the same as the possible interpretations in Experiment 3. We chose a joint presentation style following Marty et al. (2020) to increase the sensitivity of this rating task. We presented the three pairs of sentences for the three conditions on one page. The participants gave their rating on a fully-labeled 7-point Likert scale (1 = Extremely unlikely, 2 = Strongly unlikely, 3 = Somewhat unlikely, 4 = intermediate, 5 = Somewhat likely, 6 = Strongly likely, 7 = Extremely likely). Please refer to Figure A.1 for the visualization of the task.

(37) Singular than-clause subjects

Produced: More students have been to Russia than I have.

a. Intended: Students have been to Russia more than I have. (event comparison)

b. Intended: Students have been to Russia but I haven't. (event negation)

c. Intended: More students have been to Russia than just me. (individual comparison)

(38) Plural than-clause subjects

Produced: Produced: More students have been to Russia than we have.

a. Intended: Students have been to Russia more than we have. (event comparison)

b. Intended: Students have been to Russia but we haven't. (event negation)

c. Intended: More students have been to Russia than just us. (individual comparison)

There were 30 critical items and 60 filler items involved in this task. Within the filler items, 20 featured depth-charge illusion sentences extracted from Zhang et al. (2023c), 10 featured simple active voiced sentences, 10 featured passive voiced sentences, and 10 featured double object constructions, and 10 featured ditransitive sentences with a prepositional object. The latter 40 items were extracted from Gibson et al. (2013a). For the depth-charge items in (39), the produced sentence was the canonical illusion sentence, e.g., *No head injury is too trivial to be ignored*. The three intended sentences each involves a possible noise operation that results in a plausible near neighbor of the depth-charge sentence. Respectively, they contain a structural substitution from *so...as to* to *too...to* in (39a), a deletion of *not* in (39b), an antonym substitution in (39c). For the 40 fillers from Gibson et al. (2013a), the three intended sentences contained noise operations that involved word deletions/insertions, an exchange of words in the sentence, or an unrealistic massive change that replaces the entire sentence (see examples (40) – (43)). A careful and responsible participant should assign an "Extremely unlikely" label to the pair with the unrelated intended sentence. We used this tactic as a comprehension check. The other two intended sentences

all implied plausible noise operations (e.g., function word insertions and deletions (Gibson et al., 2013a; Zhang et al., 2023c) and exchanges (Poppels & Levy, 2016)).

(39) Depth-charge sentence

Produced: No head injury is too trivial to be ignored.

a. Intended: No head injury is **so** trivial **as to** be ignored. (structural substitution)

b. Intended: No head injury is too trivial to **not** be ignored. (deletion of *not*)

c. Intended: No head injury is too trivial to be **treated**. (antonym substitution)

(40) Active sentence

Produced: The ball kicked the girl.

a. Intended: The girl kicked the ball. (exchange of subject & object)

b. Intended: The ball was kicked by the girl. (2 deletions)

c. Intended: The man held the woman. (unrelated)

(41) Passive sentence

Produced: The woman was lost by the diamond.

a. Intended: The diamond was lost by the woman. (exchange of subject & object)

b. Intended: The woman lost the diamond. (2 insertions)

c. Intended: The professor criticized the TA. (unrelated)

(42) Double object construction

Produced: The sister mailed the letter the niece.

a. Intended: The sister mailed the letter to the niece. (1 deletion)

b. Intended: The sister mailed the niece the letter. (exchange of the two objects)

c. Intended: The woman was held by the man. (unrelated)

(43) Double object construction with a prepositional object

Produced: The father gave the son to the car.

a. Intended: The father gave the son the car. (1 insertion)

b. Intended: The father gave the car to the son. (exchange of the two objects)

c. Intended: The judge was insulted by the lawyer. (unrelated)

Participants read a randomized list of 90 trials. Within this trial, there were three pairs of intended/produced sentences to judge. The order of three pairs was also randomized. After each trial, they could not return to previous trials. Before the study, they also needed to finish five English sentence completion tasks which helped us filter out non-native speakers of English.

3.6.3 Results

Figure 3.8 displays the distribution of the noise likelihood rating for the three intended sentences by the two conditions. The result shows that the "event comparison" intended sentence was the most likely to be produced as the comparative illusion among the three, followed by the "individual comparison" sentence and the "event negation" sentence. The pattern is the same in both conditions.

We fitted Bayesian multilevel cumulative ordinal models to study the statistical significance of the ratings. The raw rating was taken as the dependent variable and the number feature of the than-clause subject as well as the intended sentence type were taken as fixed effects; no interaction effects were included. The random effects included random intercepts and random slopes of both the number feature and the intended sentence type for

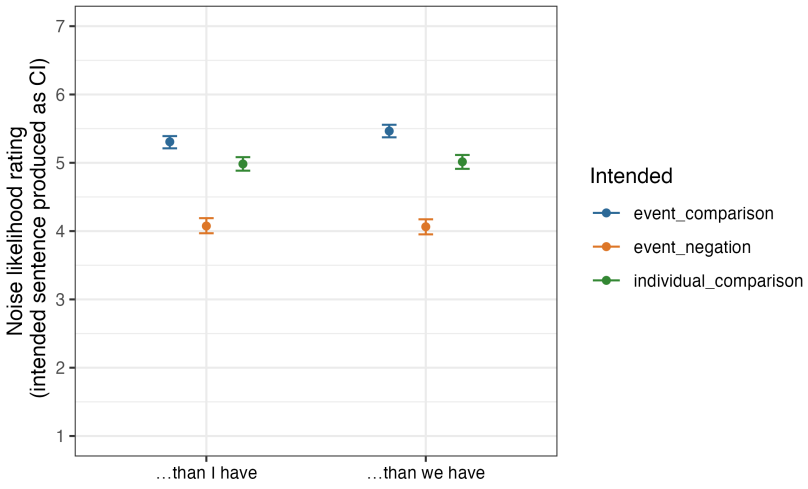


Figure 3.8: Noise likelihood ratings for the three types of intended sentences

subjects and items. The prior parameters were set as the default in R, same as Experiment 1. Through post hoc analysis, we found a main difference of the intended sentence type: the "event comparison" type was rated higher than the "individual comparison" type (β=0.40, HPD=[0.03, 0.74]), and higher than the "event negation" type (β=1.30, HPD=[0.95, 1.65]). The "event negation" type received lower ratings than the "individual comparison" type (β=-0.90, HPD=[-1.27, -0.51]). There was not a difference between the singular pronoun condition and the plural pronoun condition (β=-0.03, HPD=[-0.12, 0.056]).

3.6.4 Discussion

In Experiment 4, we constructed a noise likelihood rating experiment to approximate how likely it is that English speakers would utter the comparative illusion sentence given one of the three intended sentences. The results show that the intended sentence with the event comparison interpretation (e.g., *Students have been to Russia more than I have*) was the most

likely, compared with the individual comparison reading (e.g., *More students have been to Russia than just me*) and the event negation reading (e.g., *Students have been to Russia but I haven't.*). The number feature of the than-clause subject pronoun (i.e., *I* vs. *we*) did not affect this pattern.

3.7 General Discussion

Through four experiments, we provided an explanation for the illusion source of the comparative illusion *More people have been to Russia than I have* (Montalbetti, 1984) within the noisy-channel framework (Gibson et al., 2013a; Levy, 2008; Shannon, 1948). Within this framework, we hypothesized that the comparative illusion sentence is a perceived string with noise corruptions taking place during noisy information transmission. We further hypothesize that during language comprehension, comprehenders postulate multiple interpretations given this perceived string and choose the interpretation that is the most likely to be linguistically encoded in the perceived string, assuming that each of the multiple interpretations has equal prior probability to be conveyed during communication.

Our experimental results provide empirical evidence that support these hypotheses. In Experiment 1, we selected as the study materials the comparative illusion sentences with the canonical structure *More students have been to Russia than I/we have* rather than the variation with the noun-phrase than-clause subject such as *More students have been to Russia than the teacher has*, because of the robust and strong illusion effect induced by the former case. In Experiment 2, we used a paraphrase task to directly map out the underlying interpretations by native speakers of the CI sentences. The manual annotation revealed three prevalent readings that (i) compare the times/frequency of events (e.g., "People have been

to Russia more than I have" which we labeled it as "event comparison"), (ii) compare the cardinality of the two groups of individuals who participated in that event (e.g., "There are more people who have been to Russia than just me", labeled as "individual comparison"), and (iii) deny the participation of the event by the than-clause subject (e.g., "Students have been to Russia but I have not", labeled as "event negation"). We treated them as three possible intended interpretations which can be conceptualized as s_i according to Equation 1.1. In Experiment 3, we adopted a forced-choice selection task where native speakers were asked to choose one interpretation out of the three that best matched their initial understanding of the CI sentence. The "event comparison" interpretation was the most probable one, followed by "individual comparison" and "event negation". We took the percentage distribution among the three interpretations as the behavioral proxy for the posterior probability $P(s_i|s_p)$. In Experiment 4, we set up a noise likelihood rating task that asked participants to rate the likelihood for each of the three intended interpretations to be produced as the anomalous CI sentence when speaking fast. The results show that the "event comparison" sentence was the most likely, followed by the "individual comparison" sentence, and the "event negation" sentence. We took the noise likelihood rating as a behavioral proxy for the noise likelihood term $P(s_p|s_i)$. Assuming that the three interpretations share the same prior probability since every interpretation is quite likely to appear in a conversation that deals with personal experiences and we don't have particular reasons to discourage any interpretation, we found that the behavioral data is consistent with the noisy-channel hypothesis that the posterior probability of $P(s_i|s_p)$ is proportional to the noise likelihood $P(s_p|s_i)$ given equal prior $P(s_i)$. Therefore, these results collectively support the hypothesis that the comprehension procedure of the comparative illusion follows a noisy-channel inference (Gibson

et al., 2013a; Levy, 2008; Ryskin et al., 2018).

The noisy-channel approach also provides an explanation for the judgment patterns shown in previous research. With Experiment 4 showing that people believe that it is very likely to produce *More people have been to Russia than I have* when intending to utter *People have been to Russia more than I have*, we know that the move of *more* at the string level could be a valid operation during comprehension given that comprehension can be affected by the knowledge of production (MacDonald, 2013). This could explain why CI sentences with repeatable verb phrase in (26a), were rated higher than the nonrepeatable counterparts in (26b) – when *more*, as an adverb, can modify repeatable verb phrases (24) but not nonrepeatable verb phrases (25), it is more likely for the CI variation with the repeatable verb phrase to have a plausible near neighbor where *more* modifies the verb phrase. Though we did not directly test CI variations that feature the one-to-many/one relation between the subject and the object (28) in our experiments, we can apply the same logic. Since it makes sense for *more* to modify the object in the one-to-many subject object construction (e.g., *Cats have more mouse toys than the dog does*, (28b)) compared with the one-to-one construction (e.g., #*Cats have more striped tails than the dog does*, (28a)), sentences like (28b) were rated higher than those like (28a). While we leave it to future research to test the noisy-channel approach in explaining judgment patterns in (27) that involve noun-phrase than-clause subjects, a hypothesis is that the noise operations might involve string edits of the than-clause subjects such that they are inferred as bare plurals like *More judges vacationed in Florida than lawyers did* (27a, 27b). In a word, the noisy-channel assumption that comprehenders could infer the meaning of a given CI string by considering other alternative strings via noise edits provide an explanation for the acceptability patterns in O'Connor (2015),

O'Connor et al. (2013) and Wellwood et al. (2018).

The noisy-channel approach also complements previous processing accounts of the comparative illusion and offers a comprehensive and quantitative explanation: First, our finding about the prevalence of the "event comparison" interpretation echoes Wellwood et al. (2018) and O'Connor (2015). In addition, we also show that the event interpretation is not the only possible interpretation by providing native speakers' actual paraphrase patterns. Our empirical data in Experiments 3 and 4 show that the potential repair analysis proposed by O'Connor (2015) is definitely on the right track. The noisy-channel framework also backs up the rational inference insights proposed by Paape (2024). Furthermore, we quantitatively show that what comprehenders infer from the corrupted CI sentence is modulated by their understanding of how likely an inferred interpretation is linguistically encoded into the CI sentence through speed production. In addition to staying with correlational findings such as the effect of linguistic factors on sentence acceptability, we provide a causal explanation inspired by the Bayesian computational model of language processing for the mechanisms behind comparative illusion.

Nonetheless, there is still remaining future work to be addressed. First, within the noisy-channel framework, we assume that the three interpretations have the same prior probability $P(s_i)$. This is because from the perspective of initiating a conversation about, say, personal experiences, it seems equally likely for someone to convey the messages ranging from (i) "there are many students who have been to Russia more times than I have", (ii) "more students have been to Russia than just me", and (iii) "I haven't been to Russia". With this said, we acknowledge that there could be prior differences across different items but we defer this question to further investigation. Second, our focus is mainly on materials with pro-

noun than-clause subjects. It remains to be worked out how the noisy-channel framework can explain the pattern in sentences like *More students have been to Russia than the teacher has*. Third, it is important to map out the cross-linguistic landscape of comparative illusion. It is already a puzzle why the verb repeatability effect was not found in Danish and Swedish data Christensen (2016). It could probably be due to the morphosyntactic distinctiveness of a nominal comparative morpheme and an adverbial comparative morpheme in that language but more rigorous testing across various languages is needed. It is also interesting to use the comparative illusion as a test case for the effect of language experience in language comprehension. For example, could bilingual speakers have different patterns of comprehension which are influenced by either of their spoken language? (Leivada & Westergaard, 2020). Last but not least, the online processing profile of comparative illusion needs to be worked out and the existing data seems conflated with cross-linguistic factors. While O'Connor et al. (2013) and O'Connor (2015) find processing difficulty in the English data which was represented as a slowdown in the reading times of the two-word spillover region, Christensen (2010) provides Danish neuroimaging data showing that the brain seems to process Danish comparative illusion sentences in the same way as normal, unproblematic sentences.

Overall, the noisy-channel approach offers a promising explanation for the comparative illusion. Together with chapter 2 about the depth-charge illusion, it seems that the noisy-channel framework achieves great explanatory power for understanding language comprehension given uncertain linguistic input.

4

Negative Polarity Illusion

In this chapter, I focus on another illusion phenomenon named *the negative polarity illusion* in English and provide an account from the resource-rational approach to language comprehension (Futrell et al., 2020; Hahn et al., 2022) to explain why the illusion occurs at the first place.

This illusion has to do with negative polarity items (NPIs) which refer to words or ex-

pressions that can only appear in a limited set of environments. In English, common NPIs are *any*, *ever*, *yet*, *in years*, and *lift a finger*. The illusion case in this chapter focuses on *ever*. For a sentence that contains *ever* to be grammatical, *ever* has to appear in an environment that has the property of being *downward entailing* (Ladusaw, 1979).* Downward entailing environments are one in which replacing a general term with a more specific term still maintains the truth of the proposition. An example is in (44). If sentence (a) is true, replacing *rain* with a more specific expression *rain heavily*, as sentence (b), still preserves the truth of (b). This is the opposite to the upward entailing environments where replacing a more specific term with a general term still keeps the proposition's truth value. An example is in (45) where if sentence (a), which contains the specific term *rain heavily*, is true, then sentence (b), with a more general term *rain*, is also true.

(44) Downward entailment

a. It did not rain yesterday.

b. It did not rain heavily yesterday.

(45) Upward entailment

a. It rained heavily yesterday.

b. It rained yesterday.

One of the typical downward entailing environments is the scope of negation, such as *no* and *few*, as square-bracketed in (46a-d). The downward entailment can be calculated

*There are other terminologies that are built upon this downward entailing notion to characterize more environments where an NPI can occur, such as Strawson downward entailment (Von Fintel, 1999) and nonveridicality (Giannakidou, 2002). In this chapter, downward entailment sufficiently characterizes the relevant environments. Please refer to the review in Penka (2015) for more discussion.

by comparing (46a) with (46b) or (46c) with (46d). In contrast, the scope of the indefinite quantifier *a* is not downward entailing, as shown in (46), where (46e) does not entail (46f). The focus of this chapter is on *no* and *few*, two negative quantifiers whose scope license the NPI *ever*. We name *no* and *few* as NPI licensors.

(46) a. **No** student [has **ever** been to Antarctica]$_{scope}$.

b. **No** student [has **ever** been to Antarctica this month]$_{scope}$.

c. **Few** students [have **ever** been to Antarctica]$_{scope}$.

d. **Few** students [have **ever** been to Antarctica this month]$_{scope}$.

e. *__A__ student [has **ever** been to Antarctica]$_{scope}$.

f. *__A__ student [has **ever** been to Antarctica this month]$_{scope}$.

Given the setup of what makes sentences containing an NPI acceptable in English, we move on to understand the language comprehension pattern of these sentences. Given a more complicated sentence (47a), one important step in grammaticality/acceptability judgment is to check whether *ever* appears in the scope of *no*. During incremental processing and upon encountering *ever*, comprehenders need to retrieve the NPI licensor *no* from their working memory of the representation of earlier words.

In the language processing literature, negative polarity illusions refer to the phenomenon where readers of the ungrammatical sentence (47b) judge it as more acceptable than an equally ungrammatical sentence (47c). Both sentences are unacceptable because *ever* does not appear in the scope of *no*: in (47b), *ever* appears in the scope of the indefinite article *a*; so does (47c). The illusion effect refers to the pattern where sentence (47b) results in a faster reading time in the NPI and the following region, or decreased grammaticality judgment

accuracy, compared to ungrammatical controls (e.g., Drenhaus et al., 2005; Vasishth et al., 2008).

(47) a. **No** man who had a beard [was **ever** thrifty].*

b. * A man who had **no** beard [was **ever** thrifty].

c. * A man who had a beard [was **ever** thrifty].

While existing literature attempts to account for the processing mechanism that gives rise to such an illusion effect, none provides a satisfactory explanation, especially to account for why the negative polarity illusion only arises in very restricted environments and with time pressure.

In this study, we present a novel observation where a robust illusion effect arises in less constrained processing environments. The canonical example in our case is (48) where we show that the ungrammatical sentence (48b) elicited a strong illusion effect that was easier to detect compared to the case in (49) (Sentences in (49) are the running example in Parker & Phillips (2016)).

(48) a. [**Few** authors [that many critics recommended]] have **ever** ...

b. *[Many authors [that **few** critics recommended]] have **ever** ...

c. *[Many authors [that many critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

(49) a. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

*The square brackets indicate the scope of *no* or *a*.

b. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

c. *no-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

We also put forward a memory-based account inspired by an information-theoretic framework of language processing (Futrell et al., 2020; Hahn et al., 2022; Shannon, 1948). We hypothesize that comprehenders of the illusion sentence have a lossy memory representation of the syntactic position of the NPI licensor *few* and the adjacent determiner/quantifier, in this case, *many*. When they encounter the NPI *ever* and need to retrieve the NPI licensor, they mistakenly represent *few* in a licensing position and therefore deem the sentence acceptable. Our empirical results also show that the illusion effect arises when *few* and another vague quantifier occupy the determiner position of the embedded subject and the matrix subject, respectively. We leverage insights from the semantic similarity of vague quantifiers to explain why this specific configuration, not others, would cause such an illusion effect.

In this Introduction section, we first review the linguistic and experimental restrictions that modulate the appearance of the illusion effect. We then summarize existing language processing theories for the negative polarity illusion. Lastly, we will introduce our theory with predictions waiting to be experimentally tested.

4.1 Current Landscape

Whether researchers can observe an illusion effect for sentences like (47b) depends on multiple factors, including but not limited to the kinds of NPI licensors and NPIs in the

sentence, the types of behavioral measurements, and the methodology to collect such measurements. This complicated empirical landscape poses challenges to develop effective and explanatory theories from the perspective of language processing.

First, not all NPI licensors could give rise to an illusion effect. Negative quantifiers such as *no*, *very few*, and *not a single* were found to cause an illusion effect but sentential negations such as *not* and *n't* did not (de Dios-Flores et al., 2017; Muller et al., 2019; Orth et al., 2019, 2021; Xiang et al., 2009). On the other hand, not all NPIs could cause this effect, either. *Any* was reported to be immune to an illusion effect in sentences like *The authors [that no critics recommended] have received* ***any*** *acknowledgment for a best-selling novel* (Parker & Phillips, 2016, in Exp.1-2).

Second, all the above-mentioned illusions were found in speeded judgment tasks where participants were under time pressure to give their judgment, or the reading paradigm was such that in a sentence one word appeared at a time with a predetermined rate (Orth et al., 2021). No illusion effect has been found in any untimed judgment tasks (Parker & Phillips, 2016, Exp.1).

Third, the reading data and the neural signals seem inconclusive as well, which might be further complicated by cross-linguistic differences. In studies in German with reading measurement, Drenhaus et al. (2005) and Vasishth et al. (2005) found that the reading time for the whole sentence in the illusion condition (50b) was longer than that in the grammatical control condition (50a) and that in the ungrammatical control condition (50c). But Vasishth et al. (2008), in an eye-tracking while reading experiment, found that the reading time of the NPI and the following region in the illusion condition was only longer than that in the grammatical condition, but shorter than that in the ungrammatical condition.

(50) a. **Kein** Pirat, [der einen Braten gegessen hatte,] war **jemals** sparsam
No pirate who a roast eaten had was **ever** thrifty
"No pirate who had eaten roast (meat) was ever thrifty."

b. *Ein Pirat, [der **keinen** Braten gegessen hatte,] war **jemals** sparsam
A pirate who **no** roast eaten had was **ever** thrifty
"A pirate who had eaten no roast (meat) was ever thrifty."

c. *Ein Pirat, [der einen Braten gegessen hatte,] war **jemals** sparsam
A pirate who a roast eaten had was **ever** thrifty
"A pirate who had eaten roast (meat) was ever thrifty."

As for the EEG/ERP data, Xiang et al. (2009) found that the illusion condition had a reduced P600 signal at the NPI *ever* compared with the ungrammatical condition in English materials. But in German, Drenhaus et al. (2005) found a P600 component in both the illusion and the ungrammatical condition with no difference in the magnitude of the P600s. Vasishth et al. (2005) also found an N400 in both conditions for German materials.

Overall, the empirical landscape of the negative polarity illusion points to the heterogeneous nature of processing profiles of the illusion sentence. It is thus encouraging to see work like Orth et al. (2021) that attempts to demarcate the set of linguistic constructions that gives rise to the illusion effect versus the set that does not within the same research paradigm.

4.1.1 Cue-based retrieval account

Given these empirical findings, previous studies have provided several promising accounts to explain the processing mechanism that causes the NPI illusion. This part introduces the cue-based retrieval approach by Lewis & Vasishth (2005), Lewis et al. (2006), and Vasishth et al. (2008). This theory is derived from a more general cognitive architec-

ture named Adaptive Control of Thought-Rational (ACT-R) (Anderson & Lebiere, 1998; Anderson, 2005).

The cue-based retrieval model (Vasishth et al., 2008) assumes that sentence processing involves a series of associative memory retrievals that allow the integration of the upcoming words with the preceding words to form dependency relations. Words or phrases carry *feature value specifications* where a match of these specifications between different words during memory retrieval entails the successful integration of them. In the grammatical sentence (47a), both the NPI *ever* and the NPI licensor *no* carry the features [+negative, +c-commander]. When the incremental parser processes the licensor *no*, the corresponding feature values are activated in working memory. When the parser hits *ever* that looks for a c-commanding licensor ([+Negative, +c-commander]), it initiates a retrieval cue, looks at the matrix subject determiner *no*, and finds a perfect match between the two words with the same feature values. When there is a match, assuming all other relevant dependencies are integrated, the parser judges the sentence as grammatical. In the illusion condition (47b), the matrix subject determiner *A* carries the feature [+c-commander], and the determiner of the embedded NP *no* carries the feature [+negative]. When *ever* starts the retrieval process and can only find a partial match with *a* for the c-command cue or with *no* for the NPI licensor cue, this retrieval process can probabilistically cause errors in comprehenders' grammatical judgments (Drenhaus et al., 2005; Vasishth et al., 2005, 2008).

The cue-based retrieval account in the current specification of the retrieval cues [+negative, +c-commander] (Vasishth et al., 2008) faces several limitations. First, it cannot explain why the illusion effect only appears when the NPI licensor is within the noun phrase but not as a sentential negation (Orth et al., 2021; Parker & Phillips, 2016). Second, it lacks the

explanation for the increased and robust illusion effect exhibited in the *many/few* pair in (48b) compared with the *the/no* pair in (47b). More specification of the features seems necessary than just the [+negative, +c-command] cue.

4.1.2 Pragmatic approach

In addition to the illusion cases where the syntactic structure has a linearly preceding but structurally nonlicensing NPI licensor (49), Mendia et al. (2018) reported novel empirical evidence showing that an illusion effect might also occur in the absence of a legitimate overt licensor. In the example (51), (51a) was rated more acceptable than (51b) while both were unacceptable without an overt licensor of *ever*.

(51) a. # Whenever the summer is really dry, Susy expects **all** of her plants to die. However, a small number of the plants have **ever** died.

b. # Whenever the summer is really dry, Susy expects **none** of her plants to die. However, a small number of the plants have **ever** died.

c. ... However, a small number of the plants, not all, have ever died.

Mendia et al. (2018) adopted the "rescuing NPI method by non-veridicality" (Giannakidou, 2006) for explanation. This approach claims that when an NPI is not licensed in its sentence α but the global context pragmatically enables an alternative sentence β that is an inference of α and that licenses the NPI, the NPI can be deemed as "rescued" and acceptable. While Giannakidou (2006) used this method to explain why *only* and emotive active verbs (e.g., *regret*) license the NPI *ever* in (52), Mendia et al. (2018) also leverage this account to explain (51). Basically, in (51a), the phrase *all of her plants* in the first sentence sets up the comprehender's expectation that the following sentence should be coherent and continue

the discussion of *all of her plants*. The discourse connective *however* contrasts with the original expectation and prompts the comprehender to expect *not all of the plants*. Therefore, a non-veridical inferential continuation of (51a) could be (51c) where the negation in *not all* licenses the NPI *ever*.

(52) a. α: Only John **ever** ate vegetables. $\rightarrow$

β: **Nobody** other than John **ever** ate vegetables.

b. α: John regrets that I **ever** bought a car. $\rightarrow$

β: John would prefer it if I had **not ever** bought a car.

This pragmatic inference of non-veridical negative alternatives was also mentioned in Linebarger (1987) and Xiang et al. (2009). The core idea in Xiang et al. (2009) was that when comprehenders process sentences starting with *The authors that no critics recommended have*..., they could make a negative inference of the contrasting reference group *The authors that **some** critics recommended have **not***... Here, the sentential negation *not* licenses the upcoming NPI *ever* and thus renders the original sentence acceptable. In general, the pragmatic approach argues that comprehenders make pragmatic inferences over the perceived sentence, and if the inferred meaning licenses NPIs, the unlicensed NPI in the original sentence could still be acceptable.

Given the merit of this pragmatic approach, it still faces the challenge of explaining why only NPI licensors occurring in the nominal phrases give rise to an illusion effect. Nor could it predict the robust illusion effect from *few/many*. Similar to the cue-based retrieval account, both claims seem to require further and more specified development.

4.1.3 NPI illusion as a quantification phenomenon

Based on the empirical evidence that an NPI illusion has only occurred when the negative licensor is in the determiner position (e.g., *no*, *very few*, *not a single*), Orth et al. (2021) contends that the NPI illusion is a quantification phenomenon, a product of an application of quantifier raising (QR) (see earlier versions in Orth et al., 2019, 2020).

Quantifier raising has been a standard tool in generative linguistics for analyzing scope in natural language (Chomsky, 1981; Heim & Kratzer, 1998).* It refers to a covert structural movement of a quantifier noun phrase during the semantic interpretation of a sentence, to adjust the scope of quantification and achieve certain semantic interpretations. For example, sentence (53) can be interpreted as "for every door, there is a (different) man sitting in front of it" where the quantifier *every* has a wider scope over the indefinite article *a*. This interpretation can be achieved via quantifier raising during semantic interpretation, shown in (54).

(53) A man sits at every door.

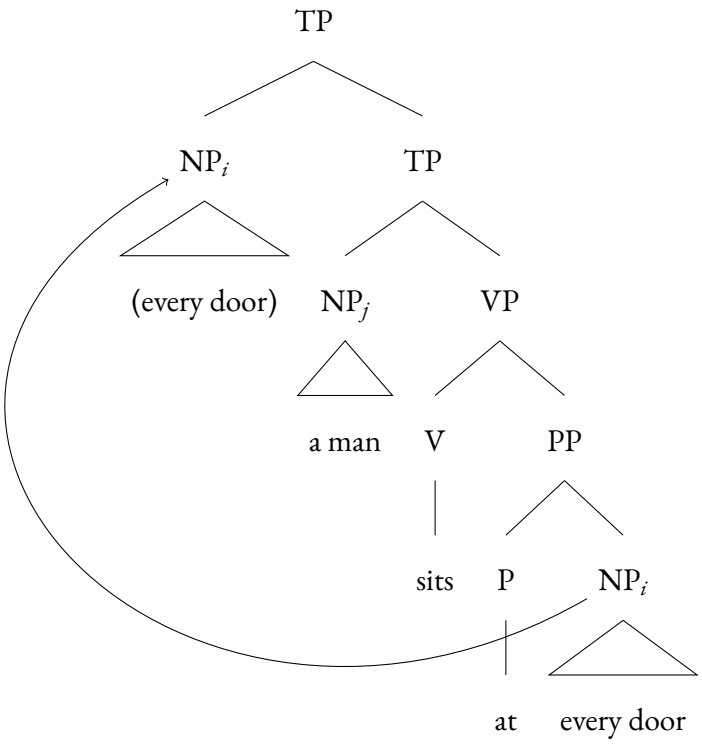
In the case of NPI illusion, Orth et al. (2021) claims that during language processing, quantifier raising is triggered when comprehenders encounter the NPI *ever*. The phrase *no critics*, as in (49b), is raised to a structurally higher position similar to that of the raised noun phrase *every door* in (54). Consequently, this raised negation takes the scope over the entire sentence which encompasses *ever* in its scope. The NPI is thus licensed. As for why comprehenders sometimes do judge the illusion sentences ungrammatical, Orth et al. (2021) explains that covertly raising quantifier noun phrases outside of finite relative clauses could

*Recent advances in generative linguistics see attempts to eliminate QR in the Minimalist Program (e.g., Hornstein, 1999)

violate English grammar in general (e.g., Farkas, 1981; Hornstein, 1995; May, 1985) and this illicit movement might cause the ungrammatical judgment.

This quantification theory would be greatly improved if the following questions were investigated. First, it remains unclear why QR is in general prohibited from relative clauses but is sometimes allowed (see another QR example in (Syrett, 2015)). Without a clear understanding of the environments that enable or disallow QR, it is hard to make quantitative predictions on the probability of the illusion effect. Second, as we will show later, the current quantification theory makes great predictions to capture the NPI illusion distribution with the negative quantifiers and yet it does not make finer-grained predictions as to which negative quantifier under what condition is more likely to give rise to an illusion effect. This is where we hope our investigation could add to the active research community dedicated to solving the NPI illusion problem.

(54)



4.2 New Approach

This paper hopes to contribute two aspects to the investigation of the negative polarity illusion. First, we provide a crucial data point (48) that exhibits a robust illusion effect even in untimed acceptability judgment tasks. Second, we offer a memory-based account inspired by the noisy-channel framework of language processing (e.g., Futrell et al., 2020; Gibson et al., 2013a; Hahn et al., 2022; Levy, 2008; Shannon, 1948) to account for why the illusion arises.

More specifically, through Experiments 1 – 3, we show that given the canonical materials

in previous research, e.g., (49) (de Dios-Flores et al., 2017; Parker & Phillips, 2016; Muller et al., 2019; Xiang et al., 2009), when the NPI licensor is changed to *few* and the determiner of the matrix subject is changed to a vague quantifier (e.g., *many*, *most*), the illusion sentence (48b) was rated higher than the ungrammatical sentence (48c) in untimed judgment tasks.

Based on this observation, we argue that one possible explanation is that during language processing, comprehenders store a lossy mental representation of the position of *few* and *many* (or *most*) which could be facilitated by the semantic and cognitive properties of these two quantifiers. When encountering the NPI *ever*, comprehenders retrieve the possible NPI licensor and the false positional representation of *few* makes them judge the illusion sentence as acceptable. We provide independent evidence supporting the lossy memory hypothesis in Experiment 4.

The uniqueness of quantifier words *many*, *few*, and *most* is that the exact quantity of the sets they quantify is harder to tell than Aristotelian quantifiers such as *no* or numerical phrase such as *more than two*. *Many* and *few* are called vague and context-dependent quantifiers whose meaning could either be a context-dependent cardinal number, (*at least n* for *many*, *at most n* for *few*), or a proportion (*at least k%* for *many*, *at most k%* for *few*) (Partee, 1989; Pezzelle, 2018). *Most* is termed proportional quantifier because its most direct interpretation *more than half* involves a verification procedure that measures the proportion of the target set against its superset (Szymanik, 2016).* In this paper, we use the term "vague quantifier" to refer to the set of quantifiers *many*, *few*, and *most*.

The context-dependency nature of these quantifiers makes it possible to tune the exact

*We only show the most direct interpretation of these quantifiers here and we acknowledge that there exists a rich repertoire of studies on the meaning of *most* and its psychological representation (see Kotek et al., 2015; Pietroski et al., 2009; Solt, 2016, among others).

quantities they refer to. In a study that probed the mental representation of different quantifiers (Pezzelle et al., 2018), *few* was shown to be the polar opposite of *many* and *most* on a proportion scale. While *few* largely refers to proportion 0% – 40%, both *many* and *most* refer to proportion 40% – 100%, although there is also a narrow range of proportions, 40%–50%, that can be referred to by the three quantifiers. Heim et al. (2020) shows that the exact proportion these quantifiers refer to can be adjusted when comprehenders are habituated to a frequently appearing level. From the language processing perspective, proportional quantifiers (e.g., *most, less than half*) were shown to take extra cognitive resources and take longer to process than Aristotelian quantifiers (e.g., *all*, *some*) and cardinal quantification phrases (*more than eight, more than seven*) (Bremnes et al., 2022, 2023; Szymanik & Zajenkowski, 2010; Zajenkowski & Szymanik, 2013; Zajenkowski et al., 2014).

The semantic vagueness and the added cost of processing for the pair of quantifiers *few* and *many** indicate that, should comprehenders spend an equal amount of energy and cognitive resources on processing any linguistic material, they would find it difficult to encode and hold in working memory the exact structural representation of the string *Many authors that few critics recommended ...* (48b) versus *Few authors that many critics recommended ...* (48a). The similarity of *few* and *many* could prompt comprehenders to misrepresent their respective syntactic position. This possible mental operation of transposition of *few* and *many* would obscure comprehenders' memory of the context preceding the NPI. With a noise-corrupted memory of the context where the NPI licensor *few* could appear at the beginning of the sentence (48b), the chance for comprehenders to make a judgment error would just increase.

*We use the *few*/*many* pair to refer to combinations of *few* with other vague quantifiers of interest.

The inaccurate, noise-inflated memory during language processing is an important component of the noisy-channel framework of language processing which posits that language encoding and decoding involve errors and noise (e.g., Futrell et al., 2020; Gibson et al., 2013a; Shannon, 1948). While previous research emphasizes noises during production (Zhang et al., 2023c), in the environment (Ryskin et al., 2018), or at the comprehension stage (Levy et al., 2009), we focus on the noisy representation specifically in the memory component (see Futrell et al., 2020; Hahn et al., 2022, for detailed discussions).

4.3 Experiment 1: Illusion with *Few* and *Many*

Experiment 1 featured an untimed acceptability judgment task that juxtaposed the canonical sentences in Parker & Phillips (2016) and similar sentences that contained *few* and *many* in place of the NPI licensor *no* and the determiner of the adjacent noun phrase *the*. The results showed that illusion sentences with *few* and *many* (e.g., *Many authors that few critics recommended have ever received acknowledgments for a best-selling novel*) elicited a strong illusion effect while the canonical sentence with *no* and *the* did not.

4.3.1 Methods

Participants

52 participants were recruited from Prolific to complete the task. Each participant was paid at an $8/hr rate for their participation; the specific payment depended on the exact time participants spent on the task. We excluded data from those (a) who did not rate at least 90% of trials; (b) who did not answer at least 75% of the comprehension checks correctly; (c) who gave the same rating across all test trials; and/or (d) who self-identified as non-native speakers

of English or from countries other than the United States. 48 participants contributed to our main data analysis.

Materials & Procedure

There were 36 target items that appeared in 6 conditions, crossing the determiner pair and the position of the two determiners within each pair as in (55). The materials that manipulated the position of *no* and *the*, as the examples in (55d) – (55f), were adopted from Parker & Phillips (2016)'s Experiment 1 with the NPI *ever*. The determiner pair was either *few/many* or *no/the* where the negative quantifier *few* or *no* was the NPI licensor and *many* and *the* was the determiner of the noun phrase that is linearly adjacent to the noun phrase that bears the licensor. There were three conditions within each determiner pair: *Matrix licensor*, *Embedded licensor*, and *No licensor*.

(55) a. *few-many, matrix licensor* (grammatical)

[**Few** authors [that **many** critics recommended]] have **ever** ...

b. *few-many, embedded licensor* (target)

∗[**Many** authors [that **few** critics recommended]] have **ever** ...

c. *few-many, no licensor* (ungrammatical)

∗[**Many** authors [that **many** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

d. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

e. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

f. *no-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

In the Matrix licensor condition, the NPI licensor *few/no* appeared as the determiner of the matrix subject (e.g., *few/no authors* in (55a) and (55d)) and the other determiner *many/the* was the determiner of the embedded subject (e.g., *many/the critics*). In this condition, the NPI *ever* was in the domain of the negative licensor and therefore the sentence should be grammatical. In the Embedded licensor condition, the relative position of the NPI licensor and the counterpart determiner was swapped – the NPI licensor was the determiner of the embedded subject (e.g., *few/no* critics in (55b) and (55e)) and the counterpart determiner was in the matrix subject position (e.g., *many/the authors*). In this condition, the NPI *ever* was not in the domain the NPI licensor and therefore should be ungrammatical. Whether participants could recognize this error would indicate whether there would be an illusion effect. In the No licensor condition, both the matrix subject and the embedded subject were *many* (55c) or *the* (55f). Because the grammatical anomaly should be fairly easy to detect, this condition acted as the ungrammatical control and a benchmark to discover the illusion effect in the Embedded licensor condition.

Besides the critical materials, there were 72 filler items with varying syntactic structures but each contained one or two of the words *no*, *few*, *the*, *many*. An example sentence is *No environmentalist's conjecture that the local wildlife would be affected by the oil spill was discussed with the Coast Guard*. Within these fillers, 48 were grammatical and 24 were ungrammatical with errors such as subject-verb agreement mismatch, and functional word deletion.

For each trial, there was a YES/NO comprehension question asking if the trial men-

tioned something (e.g., *Does this sentence mention critics who recommended authors?*). The answers were designed such that half of them were "yes" and half were "no". This comprehension question acted as an attention check. After the comprehension question, there was a sentence rating task asking *how natural the sentence is*. Participants gave their judgment on a 7-point fully labeled Likert scale (1 = "Extremely unnatural", 2 = "Unnatural", 3 = "Somewhat unnatural", 4 = "Neutral", 5 = "Somewhat natural", 6 = "Natural", 7 = "Extremely natural").

Each participant read a randomized list of 108 trials. The presentation of the trials followed a Latin Square design and the within-subjects design made sure that each participant would read the same number of trials across all the six conditions.

4.3.2 Results

The distributions of the naturalness ratings by condition are presented in Figure 4.1.* The means and confidence intervals of the six conditions were few-many, Matrix licensor (55a)(mean = 3.75, CI† = [3.56, 3.94]), few-many, Embedded licensor (55b) (mean = 3.28, CI = [3.12, 3.47]), few-many, No licensor (55c) (mean = 2.94, CI = [2.79, 3.11]); no-the, Matrix licensor (55d) (mean = 4.40, CI = [4.21, 4.56]), no-the, Embedded licensor (55e) (mean = 2.68, CI = [2.53, 2.83]), and no-the, No licensor (55f) (mean = 2.88, CI = [2.71, 3.04]).

The Bayesian result shows that there is only an illusion effect in the *few-many* condition, not the *no-the* condition. More specifically, within *few-many*, the No licensor condition received lower ratings than the Embedded licensor condition (β = -0.39, HPD = [-0.63, -0.15])

*The error bars represent 95% confidence intervals from bootstrapping.

†This 95% confidence interval was calculated via bootstrapping.

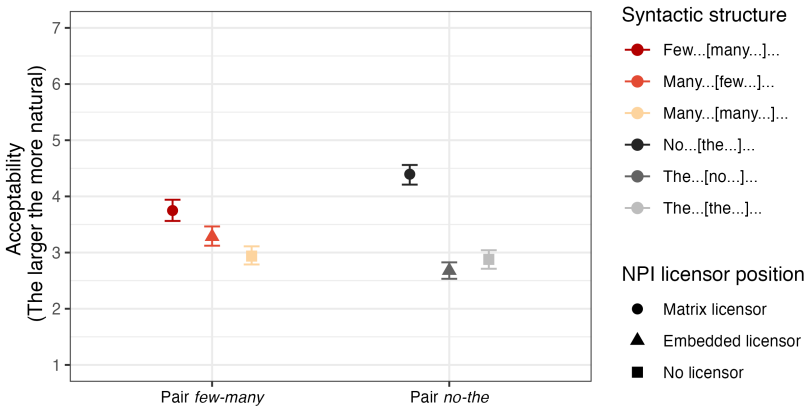


Figure 4.1: Distribution of the acceptability rating for *few* and *many*.

and the Matrix licensor condition (β = -0.87, HPD = [-1.17, -0.59]); the Embedded licensor condition also received lower ratings than the Matrix licensor condition (β = -0.47, HPD = [-0.74, -0.19]). Within the *no-the* condition, the ratings between the No licensor condition and the Embedded licensor condition were not different (β = 0.19, HPD = [-0.03, 0.42]) where the Embedded licensor condition tended to be even worse. Both these conditions received lower ratings than the grammatical condition (compared with the No licensor condition, β = -1.51, HPD = [-1.83, -1.20]; compared with the Embedded licensor condition, β = -1.70, HPD = [-2.02, -1.38]).

4.3.3 Discussion

Experiment 1 investigated whether the newly introduced *few-many* variation of the NPI illusion could appear with an offline grammaticality measure. The results indeed show an illusion effect where the *few-many* Embedded licensor condition received higher ratings than the ungrammatical No licensor control. On the contrary, the *no-the* Embedded licensor condition received similar ratings to its ungrammatical control, replicating the existing findings

that the *no-the* pair has been found to only elicit an illusion effect during speeded judgment tasks and word-by-word reading tasks (Parker & Phillips, 2016). This was the first time since the study of NPI illusion that an illusion was found in offline measures.

4.4 Experiment 2: Illusion with *Few* and *Most*

Experiment 2 adopted a minimal change from Experiment 1 by replacing *many* with *most*. Here, we also found an illusion effect with *few* and *most* but not *no* and *the*.

4.4.1 Methods

Participants

55 participants were recruited from Prolific. Each was paid at an $8.64/hour rate for their participation. We applied the same screening method as Experiment 1 and 51 participants remained for the final data analysis.

Materials & Procedure

The materials and the procedure were almost the same as those in Experiment 1, except that the determiner pairs changed to *few-most* versus *no-the*. An example of the 6 conditions was in (56). The Matrix licensor conditions in (56a) and (56d) were grammatical controls while the No licensor conditions in (56c) and (56f) were ungrammatical controls. The crucial condition was the Embedded licensor condition.

(56) a. *few-most, matrix licensor* (grammatical)

[**Few** authors [that **most** critics recommended]] have **ever** ...

b. *few-most, embedded licensor* (target)

∗[**Most** authors [that **few** critics recommended]] have **ever** ...

c. *few-most, no licensor* (ungrammatical)

∗[**Most** authors [that **most** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

d. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

e. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

f. *no-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

There were 36 critical items and 72 filler items involved in the task. The 72 fillers were the same as those in Experiment 1 except that those with the quantifier *many* in Experiment 1 substituted *many* with *most*. The Latin Square design, randomization, and the within-subjects design were the same as Experiment 1.

4.4.2 Results

Figure 4.2 displays the naturalness ratings across the 6 conditions in Experiment 2. The means and confidence intervals of the six conditions were, respectively, few-most, Matrix licensor (56a)(mean = 3.83, CI = [3.66, 3.99]), few-most, Embedded licensor (56b) (mean = 3.24, CI = [3.08, 3.40]), few-most, No licensor (56c) (mean = 3.05, CI = [2.88, 3.22]); no-the, Matrix licensor (56d) (mean = 4.25, CI = [4.08, 4.42]), no-the, Embedded licensor (56e)

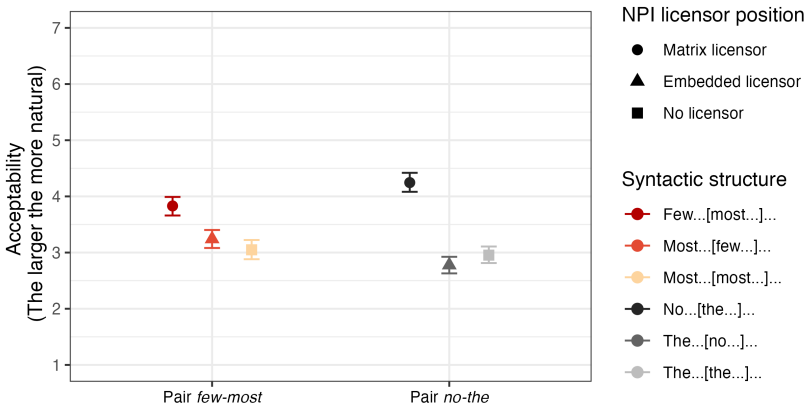


Figure 4.2: Distribution of the acceptability rating for *few* and *most*.

(mean = 2.78, CI = [2.63, 2.92]), and no-the, No licensor (56f) (mean = 2.95, CI = [2.81, 3.11]).

As in Experiment 1, we ran Bayesian multilevel cumulative ordinal models to investigate the effect of the determiner pair and the licensor conditions. The dependent variable was the raw rating of each critical trial and the fixed effects were the dummy-coded determiner pair (reference level = *no-the*) and the dummy-coded licensor position (reference level = No licensor) with their interaction term*. The structure of the random effects, the prior setup, the training parameters, and the follow-up *post hoc* comparison analysis were identical to those in Experiment 1.

The Bayesian model shows that only in the *few-most* condition was there an illusion effect – the No licensor condition received lower ratings than the Embedded licensor condition (β = -0.24, HPD = [-0.47, -0.003]), whereas in the *no-the* condition, these two conditions were quite similar (β = 0.22, HPD = [0.03, 0.41]). The other comparisons yield ex-

*Similar to Experiment 1, we used leave-one-out cross-validation for model comparison. The model with the interaction term was preferred.

pected results: the grammatical Matrix licensor condition received higher ratings than the other two conditions for both determiner pairs (For *few-most*, against the No licensor condition, β = -0.87, HPD = [-1.20, -0.55]; against the Embedded licensor condition, β = -0.63, HPD = [-0.94, -0.33]. For *no-the*, against the No licensor condition, β = -1.37, HPD = [-1.71, -1.05]; against the Embedded licensor condition, β = -1.59, HPD = [-1.94, -1.25]).

4.4.3 Discussion

Experiment 2 adopted the same offline methodology as Experiment 1 and found that when the NPI licensor was *few* and the adjacent determiner was *most*, the target illusion sentence (56b) indeed elicited an illusion effect. The pattern that the Embedded licensor condition in the *no-the* pair was regarded by participants as more unnatural than the ungrammatical control replicated Experiment 1. This suggests that the comprehension behavior of these sentences might indeed be affected by the NPI licensor and the adjacent determiner.

4.5 Experiment 3: Non-illusive Effect in Other Configurations

Experiment 3 consists of four sub-experiments which altogether show that only the combination of *few* and a vague quantifier like *many* or *most* can trigger a robust illusion effect in unspeeded judgment tasks. All four experiments had the *no-the* pair as the comparison condition. The target determiner pair was, respectively, *few-the* in Experiment 3a, *no-many* in Experiment 3b, *few-all* in Experiment 3c, and *no-all* in Experiment 3d.

Experiment 3a and Experiment 3b raised the hypotheses that the illusion effect observed in Experiment 1 was caused by the sole appearance of *few* or *many*, not the co-occurrence of the two. Experiment 3a tested whether sentences with the determiner pair *few-the* gave

rise to the illusion effect and Experiment 3b used the same logic and paradigm to test the determiner pair *no-many*.

Experiment 3c and Experiment 3d were set up to test what linguistic features of the adjacent determiner could play a role in causing the illusion effect, since *many* and *most* could be categorized as vague quantifiers or simply quantifiers. Hence, Experiment 3c and 3d explored whether common quantifiers like *all* can also elicit illusion when combined with *few* or *no*. Experiment 3c can be viewed as an implicit between-subjects comparison with Experiments 1 and 2. These three experiments subsequently tested when *few* paired with *many*, *most*, and *all*, whether an illusion would arise. Experiment 3d explored whether *no-all* could be a possible pair to induce an illusion effect.

4.5.1 Participants

In Experiment 3a, we recruited 52 participants from Prolific who received $4.5 for their participation. After we screened participants for their native language ("English"), country of origin ("United States"), survey completion rate (90%), and comprehension question performance (75%), 49 participants contributed to the final analysis. In Experiment 3b, a separate group of 48 participants were recruited with the same payment. 45 participants remained for the final analysis. In Experiment 3c, 55 participants finished the study and 51 left for further analysis. In Experiment 3d, the ratio of participation versus the inclusion of analysis was 49/46.

4.5.2 Methods & Procedure

The research pipeline of Experiment 3 was the same as that in Experiments 1 and 2, except for the materials under investigation. Experiment 3a examined the *few-the* pair (57). There were 36 critical items with a 2x2 within-subjects design crossing the determiner pair (*few-the* vs. *no-the*) and the NPI licensor position (Matrix licensor vs. Embedded licensor). We included an ungrammatical control condition where the determiner pair was *the-the* which could be the No licensor condition for both of the two pairs of configurations.

(57) **Experiment 3a: few-the vs. no-the**

a. *few-the, matrix licensor* (grammatical)

[**Few** authors [that **the** critics recommended]] have **ever** ...

b. *few-the, embedded licensor* (target)

∗[**The** authors [that **few** critics recommended]] have **ever** ...

c. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

d. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

e. *few-the/the-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

For Experiment 3b, the 36 critical items appeared in 6 conditions, crossing the determiner pair and the relative positions of the two determiners. An example trial was in (58)

where the only difference from Experiments 1 and 2 was that *no-many* was the target determiner pair under investigation.

(58) **Experiment 3b: no-many vs. no-the**

a. *no-many, matrix licensor* (grammatical)

[**No** authors [that **many** critics recommended]] have **ever** ...

b. *no-many, embedded licensor* (target)

∗[**Many** authors [that **no** critics recommended]] have **ever** ...

c. *no-many, no licensor* (ungrammatical)

∗[**Many** authors [that **many** critics recommended]] have **ever** ...

d. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

e. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

f. *no-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

In Experiment 3c, the critical determiner pair was *few* and *all*. The 2 x 3 critical manipulation setup was seen in (59).

(59) **Experiment 3c: few-all vs. no-the**

a. *few-all, matrix licensor* (grammatical)

[**Few** authors [that **all** critics recommended]] have **ever** ...

b. *few-all, embedded licensor* (target)

∗[**All** authors [that **few** critics recommended]] have **ever** ...

c. *few-all, no licensor* (ungrammatical)

∗[**All** authors [that **All** critics recommended]] have **ever** ...

d. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

e. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

f. *no-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

In Experiment 3d, the critical determiner pair was *no* and *all*. The 2 x 3 critical manipulation setup was seen in (60).

(60) **Experiment 3d: no-all vs. no-the**

a. *no-all, matrix licensor* (grammatical)

[**No** authors [that **all** critics recommended]] have **ever** ...

b. *no-all, embedded licensor* (target)

∗[**All** authors [that **no** critics recommended]] have **ever** ...

c. *no-all, no licensor* (ungrammatical)

∗[**All** authors [that **All** critics recommended]] have **ever** ...

d. *no-the, matrix licensor* (grammatical)

[**No** authors [that **the** critics recommended]] have **ever** ...

e. *no-the, embedded licensor* (target)

∗[**The** authors [that **no** critics recommended]] have **ever** ...

f. *no-the, no licensor* (ungrammatical)

∗[**The** authors [that **the** critics recommended]] have **ever** ...

... received acknowledgments for a best-selling novel.

For the four sub-experiments, the design of the filler items, the acceptability rating paradigm, the Latin Square design, and the randomization of the trial list were the same as in Experiments 1 and 2.

4.5.3 Results

Figure 4.3 displays the sentence acceptability distributions for the four experiments. The descriptive analysis for each experiment in Experiment 3 is in Table 4.1.

Table 4.1: Descriptive analysis for Experiment 3

Experiment 3a: *few-the*			
Determiner	Licensor position	Mean	Confidence interval
few-the	Matrix licensor	4.68	[4.53, 4.83]
	Embedded licensor	3.30	[3.14, 3.47]
no-the	Matrix licensor	4.58	[4.42, 4.73]
	Embedded licensor	3.12	[2.96, 3.26]
the-the	No licensor	3.15	[3.01, 3.33]
Experiment 3b: *no-many*			
Determiner	Licensor position	Mean	Confidence interval
no-many	Matrix licensor	3.58	[3.39, 3.79]

	Embedded licensor	2.94	[2.77, 3.13]
	No licensor	2.91	[2.74, 3.09]
	Matrix licensor	4.30	[4.10, 4.51]
no-the	Embedded licensor	2.90	[2.71, 3.08]
	No licensor	2.71	[2.56, 2.87]

Experiment 3c: *few-all*

Determiner	Licensor position	Mean	Confidence interval
	Matrix licensor	3.38	[3.21, 3.54]
few-all	Embedded licensor	2.71	[2.55, 2.86]
	No licensor	2.54	[2.40, 2.69]
	Matrix licensor	3.87	[3.70, 4.04]
no-the	Embedded licensor	2.59	[2.45, 2.74]
	No licensor	2.85	[2.67, 3.01]

Experiment 3d: *no-all*

Determiner	Licensor position	Mean	Confidence interval
	Matrix licensor	3.31	[3.15, 3.49]
no-all	Embedded licensor	2.68	[2.53, 2.84]
	No licensor	2.67	[2.52, 2.83]
	Matrix licensor	4.41	[3.91, 4.29]
no-the	Embedded licensor	2.62	[2.47, 2.77]
	No licensor	2.90	[2.74, 3.04]

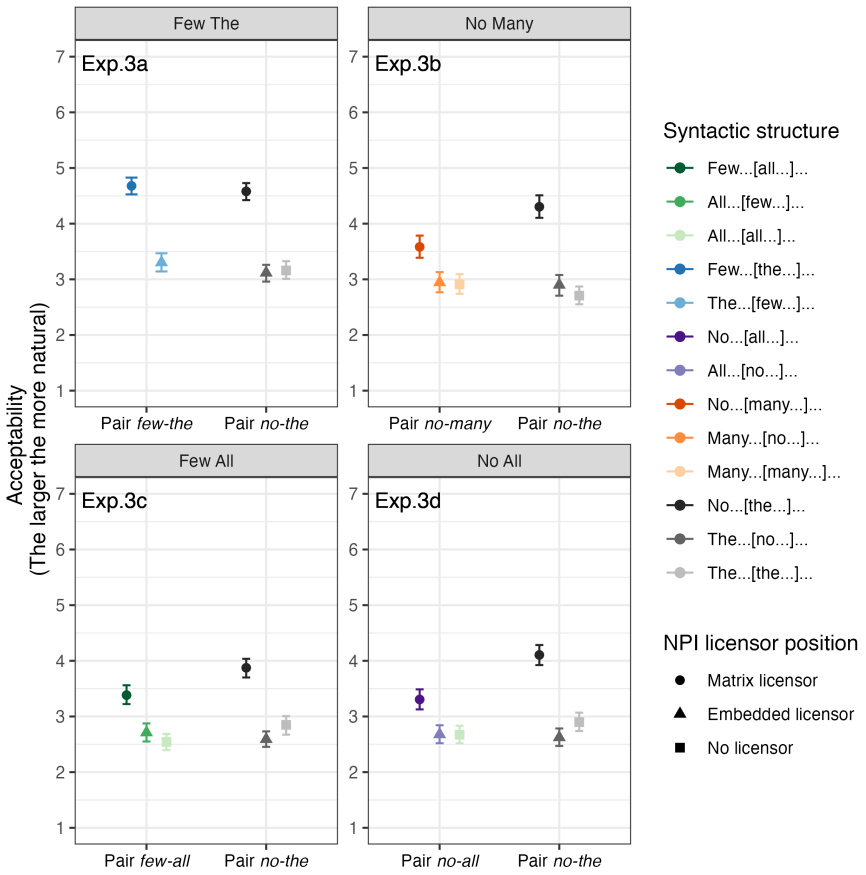


Figure 4.3: Acceptability ratings for four more determiner pairs.

Experiment 3a: *few-the* vs. *no-the*

Figure 4.3a shows that the critical Embedded licensor condition in both determiner pairs did not differ much from the ungrammatical condition. To statistically test this, we run Bayesian multilevel ordinal regression models using *brms* in R as in Experiments 1 and 2. Specifically, since Figure 4.3a shows that the grammatical Matrix licensor condition received overwhelmingly higher acceptability scores than the rest of the conditions and that the illusion effect has been measured by the increased acceptability in the Embedded licensor condition compared with the ungrammatical control, we only focused on the data in the Embedded licensor condition and the ungrammatical condition. For such a Bayesian model, the dependent variable was the raw sentence score per trial, the fixed effect was the dummy-coded three-level categorical variable with the ungrammatical control condition as the reference against the other two Embedded licensor conditions, and the random effect included a random intercept and a random slope of main effect for both items and subjects. The priors and meta-parameters of the model were the same as those in Experiment 1 and 2. All of the $\hat{R}$s of the fixed effects were 1.0, meaning a successful convergence of the sampling chain to the posterior distributions of the estimated coefficients.

Compared with the ungrammatical control condition, neither of the two Embedded licensor conditions received different ratings (for the *few-the* pair, β = 0.12, CrI = [-0.10, 0.35]; for the *no-the* pair, β = -0.06, CrI = [-0.27, 0.15]), which is consistent with the patterns in Figure 4.3a.

Experiment 3b: *No-many* vs. *No-the*

As shown in Figure 4.3b and Table 4.1, there was a significant overlap of confidence intervals between the Embedded licensor condition and the No licensor condition for both determiner pairs.

We used Bayesian multilevel ordinal regression models, similar to Experiments 1 and 2, for the statistical analysis. For one model, the dependent variable was the raw sentence score per trial and the ordinal level intervals were assumed to be non-equidistant. The fixed effects included the dummy-coded determiner pair (reference = "no...the"), the three-level categorical variable for the relative position of the determiner pair (reference = No licensor), and their interaction term. The random effects included the full structure of the fixed effects with a random intercept for the subject and the item. The model's priors and meta-parameters matched those from Experiments 1 and 2. For another model, we excluded the interaction term. For both models, all fixed effects converged successfully, indicated by $\hat{R}$s of 1.0. The leave-one-out cross-validation model comparison method shows that the model with the interaction term fitted better with the data pattern.

We used the *emmeans* package to investigate within each determiner pair the distribution of different licensor positions. Within the *no-the* pair, the ungrammatical No licensor condition did not differ from the Embedded licensor condition (β = -0.11, HPD = (-0.31, 0.12)) but both conditions were more unacceptable than the grammatical Matrix licensor condition (with the No licensor condition, β = -1.38, HPD = (-1.76, -1.04); with the Embedded licensor condition, β = -1.27, HPD = (-1.66, -0.90)). For the *no-many* pair, the results were identical: there was no difference between the No licensor and the Embedded licensor conditions (β = -0.01, HPD = (-0.26, 0.21)) and both received lower ratings than the grammat-

ical Matrix licensor condition (with the No licensor condition, β = -0.57, HPD = [-0.92, -0.23]; with the Embedded licensor condition, β = -0.56, HPD = [-0.92, -0.20]).

Experiment 3c: *Few-all* vs. *No-the*

As shown in Figure 4.3c and Table 4.1, both the grammatical Matrix licensor condition received the highest score within each determiner pair but there was still some confidence interval overlap between the Embedded licensor and the No licensor conditions for both determiner pairs.

Similar to Experiment 3b, we ran two Bayesian multilevel ordinal regression models to investigate the illusion effect. The dependent variable was the raw acceptability score (1–7) per trial with non-equidistant intervals. The fixed effects included the dummy-coded determiner pair (reference = "no...the") and the dummy-coded licensor position (reference = No licensor). The random variable included a random intercept and the random slope for the full fixed effects for the subject and the item. The only difference between the two models was the inclusion of an interaction term. With leave-one-out cross-validation, we chose the one with the interaction term which was a better fit. According to the *emmeans* result, within the *no-the* condition, the ungrammatical No licensor condition was judged to be more acceptable than the target Embedded licensor condition (β = 0.26, HPD = [0.08, 0.46]) and both conditions received lower ratings than the grammatical Matrix licensor condition (with No licensor, β = -1.08, HPD = [-1.38, -0.76]; with Embedded licensor, β = -1.34, HPD = [-1.64, -1.03]). Within the *few-all* condition, there was no difference between the No licensor and the Embedded licensor condition (β = -0.21, HPD = [-0.43, 0.01]) and both received lower ratings than the grammatical Matrix licensor condition (with No li-

censor, β = -0.99, HPD = [-1.30, -0.71]; with Embedded licensor, β = -0.77, HPD = [-1.09, -0.47]).

Experiment 3d: *No-all* vs. *No-the*

Figure 4.3d and Table 4.1 clearly show that there was no illusion effect for either of the two determiner pairs. Using the same statistical analysis method, we found, through a Bayesian multilevel ordinal regression model with an interaction term between the determiner pair and the licensor position, that (1) within the *no-the* control pair, the ungrammatical No licensor condition was more acceptable than the target Embedded licensor condition (β = 0.30, HPD = [0.11, 0.50]); both conditions were more unacceptable compared with the grammatical Matrix licensor condition (β = -1.18, HPD = [-1.48, -0.90] for the No licensor condition, β = -1.47, HPD = [-1.78, -1.18] for the Embedded licensor condition); (2) within the *no-all* condition, despite that the grammatical Matrix licensor condition received the highest acceptability rating, there was no difference between the Embedded licensor and the No licensor condition (β = -0.02, HPD = [-0.23, 0.19]).

4.5.4 Discussion

Experiment 3a shows that the target Embedded licensor sentence *The authors that few critics recommended have ever received acknowledgment for a best-selling novel* did not trigger an illusion effect. Together with Experiments 1 and 2, we know that the illusion effect is not triggered by the sole appearance of the NPI licensor *few*. In Experiment 3b, the lack of the illusion effect in ungrammatical sentences with *no* and *many* rejected the hypothesis that it is the determiner *many* that plays the most determining role in triggering the illusion effect

in Experiment 1. Experiments 3a and 3b suggest that it was the combination of *few* and a vague quantifier that triggered the illusion effect. Experiments 3c and 3d failed to find an illusion effect from the illusion sentence with *few*/*all* or *no*/*all*, suggesting that the semantic vagueness of the adjacent quantifier was a necessary condition for the negative polarity illusion.

In sum, Experiment 3 provided extra evidence that the illusion effect observed in Experiments 1 and 2 was triggered by the unique combination of *few* and *many*/*most*.

4.6 Experiment 4: A Lossy Memory Representation

Our memory-based account hypothesizes that comprehenders of sentences *Many authors that few critics recommended have ever...* store a lossy memory representation of the context before the NPI where they could transpose *many* and *few*. With the transposition, comprehenders would render the NPI *ever* licensed by *few* and judge the illusion sentence as acceptable. Experiment 4 featured a memorization and recall task to provide independent behavioral evidence that comprehenders were more likely to transpose *few* with *many* compared to *no* with *the*.

4.6.1 Participants

81 participants finished this task via Prolific. Each was paid $5 for their participation. We included in our main analysis those who (1) self-identified as native speakers of English from the United States, (2) finished at least 90% of all trials in the task, and (3) correctly finished at least 75% of the mathematical testing problems. 70 participants remained.

4.6.2 Materials & Procedure

We designed an online memorization-recall task with cognitive controls to answer how likely native speakers of English are to mistakenly recall the relative position of *few* and *many*. Participants were asked to read a targeted sentence (named exposure sentence), finish one arithmetic problem that taxed their working memory and created a time gap, and then recall whether a newly appeared sentence (named recall sentence) was identical to the sentence they had just seen. The exposure sentences also included *few* and *many*. Half of the corresponding recall sentences switched the relative position of *few* and *many*. If participants fail to recognize that a switch took place, we take it to represent that they hold a false memory about the relative position of *few* and *many*.

There were 20 critical items with a uniform syntactic structure where the subject had the structure *DET Xs with DET Ys* (DET represents determiners; *X* and *Y* were noun phrases), the verb phrase had the present perfect tense, and a transitive verb-object structure with a prepositional phrase modifying the event denoted by the verb phrase. The experiment followed a 2 x 2 x 2 within-subjects design. An example is (61). In the exposure phase, the exposure sentence contained either the pair *few-many* or the pair *no-the* while the latter was the control condition. For each sentence, the relative position of the determiner pair was manipulated to be "Negation first" or "Negation second" where "Negation first" had *few*/*no* appear before *many*/*the* and "Negation second" had *many*/*the* appear before *few*/*no*. The hypothesis here is that the relative position of the two determiners does not affect the recall accuracy. The other layer of manipulation took place in the recall sentence. Half of the recall sentences that a participant saw were identical to the exposure sentences and the other half switched the two determiners.

(61) An example in the exposure phase

a. *few-many*, negation first

Few students with **many** awards have gone to top PhD programs in the world.

b. *few-many*, negation second

Many students with **few** awards have gone to top PhD programs in the world.

c. *no-the*, negation first

No students with **the** awards have gone to top PhD programs in the world.

d. *no-the*, negation second

The students with **no** awards have gone to top PhD programs in the world.

40 filler items had similar structures compared to the critical items (i.e., present perfect tense, prepositional phrases at the end). The articles of the subjects included *few*, *no*, *some*, *many*, *lots of*, *the*, *most*, and the bare plural format. An example is *Most flight attendants for international airlines have acquired a second language to fulfill their career goal*. Each article type had five filler items. In the recall phase, 20 items were randomly chosen where the recall sentence was an exact match with the exposure sentence. The other 20 items went through various types of edits: five had two words substituted, five had three words substituted, five had four words substituted, and the rest five had the prepositional phrases deleted. The substitutions were all content words.

Between the exposure phase and the recall phase, participants also finished one arithmetic problem. The arithmetic problem was created with a template: Two integers from 0 to 50 were randomly selected to undergo an arithmetic operation randomly chosen from addition, subtraction, multiplication, and division. The third integer from 0 to 50 was then randomly chosen to randomly perform one of the four operations. One example is

(5+45)/2. Participants were asked to do the calculation and provide their answers round up to the integer. 60 math problems were generated and each was paired with one critical/filler item.

To encourage participants to finish the task with the required cognitive skills like memorization, mathematical calculation, and recall, on the instruction page, we explicitly noted that participants were not allowed to finish the recall task by saving a copy of the exposure sentence to avoid memorization. The entire survey was created by Qualtrics.

The experimental procedure was such that participants saw a randomized list of 60 trials. Within each trial, they were asked to comprehend an exposure sentence, finish a math calculation task, and then answer whether the newly appeared recall sentence was identical to the exposure sentence. They provided their answer via a Yes/No question "Is the following sentence exactly the same as the one you just saw?"

4.6.3 Results

Figure 4.4 shows that the percentage of choosing YES to the recall question "Is the following sentence exactly the same as the one you just saw?" across the eight conditions. Choosing YES means that participants believed that the recall sentence and the exposure sentence were the same. The bars are colored coded to represent whether, in reality, the determiners were switched in the recall sentence. For the switch condition, which represents that the recall sentence had the determiners switched and thus differed from the exposure sentence, choosing YES means that participants held a false memory. For the no switch condition, where the exposure sentence and the recall sentence were the same, choosing YES means that participants were correct in their memory. From Figure 4.4, we saw that when

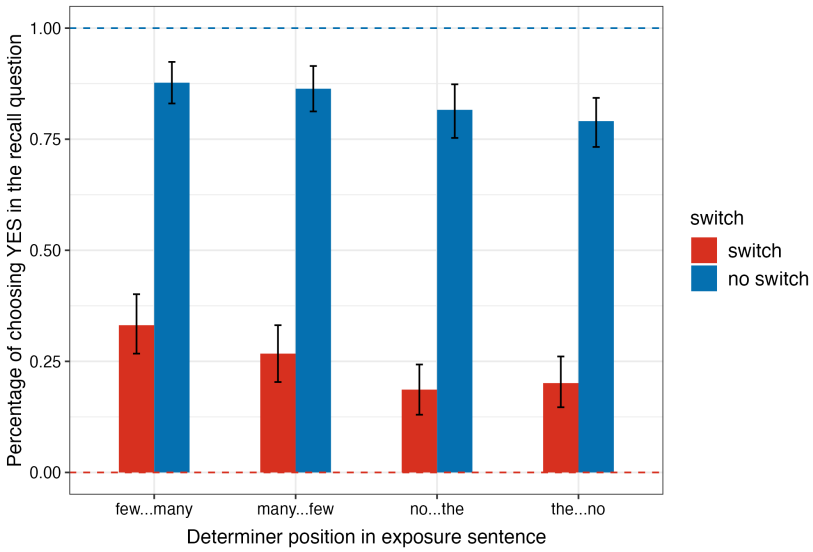


Figure 4.4: The percentage of choosing YES for the recall question.

the exposure sentences contained *few...many*, regardless of their relative position, participants were quantitatively more likely to choose YES. The most critical data point is the percentage of choosing YES for the switch condition, because it indicates the rate of false memory. From a visual inspection, it is clear that sentences with *few* and *many* are more likely to incur a false recall.

We ran Bayesian multilevel logistic regression models to statistically test this effect. The priors and the model parameters were set as the default. The dependent variable was the binary answer of choosing YES (1 = choosing YES); the main effects included the dummy-coded binary categorical variables of switch (reference = no switch), determiner pair (reference = *no...the*), and the determiner position (reference = negation second); the random effects included a random intercept and the full structure of the main effects for both subjects and items. Using leave-one-out cross-validation, we chose the model without interaction terms as a better fit, suggesting that there was no interaction effect among these factors. The

marginal comparison results from *emmeans* show that, for the main effects, sentences with *no* and *the* were less likely to induce a YES answer compared with those with *few* and *many* (β = -0.71, HPD = [-1.14, -0.29]); the position of negative determiner did not make a difference (β = -0.18, HPD = [-0.54, 0.18]); when there was no change between the exposure and the recall sentences, participants were more likely to answer YES for the recall question (β = 3.81, HPD = [3.17, 4.45]).

4.6.4 Discussion

The critical observation is that participants seemed more likely to believe there was no change for sentences containing *few* and *many* even when in fact there was, compared with sentences with *no* and *the*. This suggests that it might be harder for participants to track the veridical material in sentences with *few* and *many* and that they might hold a lossy memory of these materials.

The result in Experiment 4 also explains the acceptability rating patterns for the grammatical Matrix licensor condition in Experiment 1. There, as shown in Figure 4.1, the Matrix licensor condition for the *few-many* pair was rated worse than that for the *no-the* pair. One of the reasons for the degraded judgment could be that participants stored a lossy memory representation of the grammatical preamble *Few authors that many critics...*, represented it as the ungrammatical preamble *Many authors that few critics...*, and thus rated the grammatical sentence as worse than the *no-the* counterpart.

4.7 General Discussion

The negative polarity illusion is a fascinating phenomenon for understanding the nature of language processing because it reveals that comprehenders treat the ungrammaticality of ill-formed sentences differently. While an NPI needs a licensor and should appear in the scope of the licensor for the sentence to be acceptable, it is interesting from a processing algorithmic and mechanistic level to learn why an unlicensed NPI preceded by an inaccessible licensor could trick native speakers of a language.

Here, we provided novel data and a memory-based account for the negative polarity illusion. In Experiment 1, we demonstrated a robust illusion effect: in sentences such as *Many authors that **few** critics recommended have **ever** received acknowledgment for a best-selling novel* where the NPI *ever* is not in the domain of the potential NPI licensor *few*, comprehenders still favored this type of sentence compared to the one without any NPI licensor. We took the increased acceptability from the ungrammatical baseline as the illusion effect and we were the first to identify such an effect in an offline, untimed acceptability judgment task. In Experiments 2 – 4, we found that the illusion effect uniquely appeared with the NPI licensor *few* and when the initial quantifier of the matrix subject was *many* or *most*. The semantic property of *few*, *many*, and *most* as vague, proportional quantifiers inspired us to put forward a noisy-channel memory-based account. We hypothesized that readers store a lossy and distorted memory representation of the context that precedes the NPI during language comprehension. Specifically, the mental representation could feature a transposition of *few* and the vague quantifier such that the transposed *few* becomes the quantifier of the matrix subject and thus licenses the NPI *ever*. In Experiment 4, we offered independent behavior evidence to support this hypothesis. We show that in a similar structure, when *few*

was embedded in the postpositional modifier phrase of the matrix subject with the quantifier *many* (e.g., *Many students with few awards have gone to top PhD programs in the world.*), comprehenders were more likely to form a lossy memory of that sentence. They were more likely to represent the original sentence with a mental switch of the position of *few* and *many*. Altogether, we claim that the negative polarity illusion arises when comprehenders switch the embedded NPI licensor with the adjacent determiner/quantifier and rely on this noisy representation of the context to judge the condition of the NPI. This claim is in line with a noisy-channel framework of language processing (Futrell et al., 2020; Gibson et al., 2013a; Levy, 2008; Levy et al., 2009; Shannon, 1948).

This work is consistent with the finding that the negative polarity illusion only arises when the NPI licensor is inside a nominal phrase (de Dios-Flores et al., 2017; Orth et al., 2021). Instead of adopting the formal semantic tool of Quantifier Raising as the mechanistic explanation (Orth et al., 2021), we provide insights into the processing mechanism by assuming that language processing in general can be inflated with noises and errors and locating the errors in comprehenders' working memory of the sentence. This processing theory also predicts the reduced or lack of illusion effect in the canonical illusion sentence (49b) reported in Parker & Phillips (2016). First, the noise is realized as exchanges of similar functional words (i.e., *few* and *many* in Experiment 1; *few* and *most* in Experiment 2) and the exchange of similar words should be easier than the exchange of two distinct words (i.e., *no* and *the* in the canonical case, supported by Experiment 4). Therefore, comprehenders of (49b) would be less likely to hold a lossy memory representation and consequently less likely to experience an illusion in untimed judgment tasks. This could be changed when the task has time pressure, as shown in previous studies that the *no*/*the* illusion appeared in speeded

judgment tasks (Orth et al., 2021; Parker & Phillips, 2016; Xiang et al., 2009). The fact that the memory-based account makes clear predictions about the differences in *few*/*many* and *no*/*the* makes it take a step further compared with the cue-based retrieval approach and the pragmatic approach which did not make such specific prediction.

5

Evaluating Large Language Models on Language Illusions

From human experiments, language illusions reveal that language processing is not always accurate in the sense that the interpretation or the metalinguistic judgment of a sentence does not always align with its literal meaning or grammaticality. Using offline judg-

ments, I've shown that for the depth-charge illusion and the comparative illusion, the common interpretations of these anomalous sentences are not always the literal and compositional meaning of those sentences. For the negative polarity illusion, it seems that comprehenders could store a lossy memory representation of sentences containing *ever* and an inaccessible NPI licensor to the extent that they rated these anomalous sentences more acceptable than an ungrammatical control. From another perspective, the previous chapters show that the deviation from the literal meaning or the sentence's grammaticality can be viewed and explained by the rational aspect of sentence processing. Humans are not making mistakes in this scenario; instead, given limited cognitive resources and the extra cost to achieve an exact parse of a complex sentence, humans go with their prior knowledge about the topic and the linguistic structures to infer the meaning or the structure of the sentence. This indicates that humans have a way to deal with complexity and uncertainty given the cognitive and environmental constraints in communication. The question now becomes whether we can see similar processing performance from large language models (LLMs) that have arguably exhibited great linguistic capacities and triggered the debate of whether they can be viewed as cognitive models of human language processing.

The motivation and potential of this investigation are as follows. First, from the understanding of the nature of information processing, we learn from previous chapters that human brains take a rational route to extract the most relevant information from a sentence given limited cognitive resources. For natural language processing algorithms that theoretically do not have constraints such as memory shortage or distractions, will they generate the same processing output (e.g., in the form of a probabilistic score for a sentence string) that mirrors that from the human brain? While it is a giant research program to investigate the al-

gorithmic similarity between human brains and computer programs in language processing, this data point from language illusions could make a step towards understanding that.

Second, from the status quo of the LLM evaluation work, current linguistic evaluations of LLMs use human language processing data (e.g., human norming data (Nair et al., 2020; Zhang et al., 2022), acceptability judgments (Linzen et al., 2016; Marvin & Linzen, 2018), and behavioral or neural measures of language processing (Schrimpf et al., 2021; Kauf et al., 2022)) as benchmarks to investigate whether LMs possess knowledge of language. These evaluations assume that humans always comprehend linguistic materials according to grammar and that they always understand the literal meaning of the sentence. But as shown by language illusions, that is not always true. Given noisy data that could contain errors, humans flexibly make rational inferences. Should we expect LLMs to behave in a rational way, the same as humans? Or should LLMs act like literal grammarians or proofreading checkers that detect the errors?

Consider, for example, the well-studied case of subject-verb agreement. While we expect an LM of Standard American English to prefer "the key to the cabinets **is** on the shelf" to "the key to the cabinets **are** on the shelf" (as discussed in Linzen et al., 2016), a wealth of psycholinguistic research has systematically documented that humans can ignore errors and accept globally ungrammatical strings (stemming from Bock & Miller, 1991). Should LMs follow the ideal grammar or mimic human's behavior?* Language illusions are a nice study case from which the investigation can lead to disambiguating conclusions.

In this study, basic examples of the language illusions are given in (62): the comparative illusion (62a), the depth-charge illusion (62b), and the negative polarity illusion (62c). All

*For additional critiques of the role of ideal grammatical knowledge in evaluations of LMs, see Pannitto & Herbelot (2020); Weissweiler et al. (2023).

three in (62) are literally unnatural English sentences, despite the fact that humans often find them surprisingly acceptable.

(62) a. More people have been to Russia than I have.

b. No head injury is too trivial to be ignored.

c. The hunter who no villager believed to be trustworthy will ever shoot a bear.*

In this paper, we relied on minimally different strings springing out from the basic illusion sentences that are either (a) considered fully **acceptable** by human participants, (b) considered fully **unacceptable** by human participants, or (c) rated **surprisingly acceptable** by humans (i.e. instances of the relevant illusion). We explored whether language models capture the basic contrast between acceptable and unacceptable strings, whether they rate illusion sentences as better than their unacceptable counterparts, and finally, whether models capture nuanced linguistic manipulations that influence human judgments of the illusion material. Further, we compared two ways of measuring models' preferences, one over the whole sentence (*perplexity*) and another of a privileged position in the sentence (*surprisal*).

If LMs pattern like human comprehension behavior that involves rational inferences, we expect to derive measures that similarly rate illusion sentences as more acceptable than typical unacceptable sentences. If, on the other hand, LMs align with ideal grammatical judgments, illusion sentences should be rated as unacceptable. Our findings indicate that none of the language models we investigated consistently exhibited illusion effects or demonstrated overall human-like judgment behaviors. Nor do they possess the necessary linguistic knowledge for literal sentence processing. These findings add more insights into the dis-

*During the time of this research, the *few-many* example had not been discovered in Chapter 4. Therefore, this research just focuses on the *no-the* sentences.

cussion of LMs' emulation of human behavior and their construal as cognitive models of human language processing.

5.1 Related work

5.1.1 LMs' linguistic abilities

We draw insights from evaluation work relying on acceptability tasks. The construction of minimal pairs has been used to evaluate models for a variety of linguistic processes, including subject-verb agreement (e.g., Linzen et al., 2016), filler-gap dependency (e.g., Wilcox et al., 2018), and binding (e.g., Davis, 2022). This basic template has been expanded into a variety of benchmarks, both for investigations of English (e.g., Warstadt et al., 2020), but also, other languages (e.g., Chinese (Song et al., 2022), Russian (Mikhailov et al., 2022), Japanese (Someya & Oseki, 2023)). While aggregated results suggest that models overlap with human acceptability judgments in a variety of cases (e.g. Hu et al., 2020), LMs can behave in distinctly non-human-like ways in capturing the intricacies of grammatical phenomenon (e.g. Lee & Schuster, 2022), the interaction between linguistic processes (e.g. Davis & van Schijndel, 2020), and in generalizing knowledge to infrequent items (e.g. Wei et al., 2021).

With respect to incremental processing behaviors, work has compared LMs to a range of human incremental behavioral measures, including fMRI (e.g. Schrimpf et al., 2021) and reading times (e.g. Shain et al., 2022). In our experiments, we are interested in cases where human interpretations and behaviors differ from what is expected given the literal content of the entire string. *Garden path* sentences are a classic example of this basic phenomenon. Strings like "The horse raced past the barn fell" are often difficult for humans on first read-

ing because the word *raced* is misparsed as a main verb (e.g., *the horse raced past*) rather than a reduced relative clause (e.g., *the horse that was raced past the barn fell*). LMs have been shown to similarly misprocess these sentences (van Schijndel & Linzen, 2021), though they fall short of capturing the magnitude of the processing cost (Arehalli et al., 2022). Here we expand these investigations to language illusions that similarly trigger errorful acceptable judgments in humans while being unnatural and unacceptable. We find that LMs do not pattern like humans in all cases.

5.2 Methods

5.2.1 Models and Measures

We analyzed four language models, two masked language models, and two autoregressive models: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020). BERT, RoBERTa, and GPT-2 were accessed via https://huggingface.co/ (Wolf et al., 2020), and GPT-3 via OpenAI's API.* We used two measures, sentence level perplexity and surprisal of specific target words. For autoregressive models, the surprisal of a specific word† is given by the following equation:

$$\text{Surp}(w_i) = -\log \text{Prob}(w_i | w_1 \ldots w_{i-1}) \tag{5.1}$$

Perplexity for a sentence of N words is:

*We used 'bert-base-cased', 'roberta-base', 'gpt2', and 'text-davinci-003'. Code for replicating the results, statistical tests, and figures can be found at `https://github.com/forrestdavis/LanguageIllusions.git`.

†For words that are subworded, the joint probability was calculated.

$$2^{\frac{1}{N}\sum_{i=1}^{N}\text{Surp}(w_i)} \tag{5.2}$$

For bidirectional models, we calculated the surprisal of a word in a context by using the masking technique in Kauf & Ivanova (2023), which corrects for words that are subworded.* Further, we used this masking technique to calculate the pseudo-perplexity of a sentence (Salazar et al., 2020).

5.2.2 Evaluation procedure

We treated LMs as psycholinguistic research subjects to generate both whole-sentence perplexity and surprisals at critical words for carefully controlled minimal pairs for each illusion (following, Futrell et al., 2019). Assuming these two scores are correlated to human acceptability judgments (Lau et al., 2017), we constructed mix-effects linear regression models from the *R* package *lme4* to test whether LMs were also sensitive to reported manipulations that affect human judgments.† For each scoring metric, we took it as the dependent variable and coded the manipulation condition representing a certain hypothesis into the independent variable. We read the estimated coefficient(s) of the tested condition variable(s) to infer whether LMs show sensitivity to the effect of that condition manipulation on the scoring metric. We evaluated language models in three broad aspects: acceptability differentiation, illusion effect, and sensitivity to manipulations.

- **Acceptability differentiation** We first asked whether language models could dis-

*For example, consider the word 'souvenir'. This is subworded by BERT into 'so', '##uven', and 'ir'. Rather than MASK each subpart, one at a time, (e.g., 'so' [MASK] 'ir'), the right context of the target subword is always masked (e.g., 'so' [MASK] [MASK]).

†We did not adopt the Bayesian statistics framework due to concerns with running time with demands of generating multiple models.

tinguish acceptable sentences from unacceptable sentences that humans have no trouble dealing with.* Models with relevant knowledge should assign lower perplexity/surprisal to acceptable sentences versus unacceptable ones.

- **Illusion effect** We took the results from the acceptability differentiation task as the foundation to test the illusion sentences. Here, we hypothesized that language models should either (i) align with humans' judgments, reflected by models' generating a lower perplexity/surprisal for illusion sentences than the unacceptable controls, or (ii) deviate from human behavior and show hints of being a literal processor, reflected by models' generating a higher or similar perplexity/surprisal score compared to the unacceptable condition. If models behave like humans, then we expected (i) to be the models' consistent behavior. If models conform to (ii), we take this as evidence of non-human-like behavior.

- **Sensitivity to manipulations** Lastly, we assessed whether language models were sensitive to illusion-specific linguistic manipulations that affect human judgments. For example, in the negative polarity illusion, would the switch of NPI licensor from *no* to sentential negation *n't* change the ratings of the illusive sentence? A greater degree of sensitivity indicates that the corresponding linguistic knowledge and how the knowledge affects sentence acceptability could be encoded in or learned by LMs. This allowed us to draw a fine-grained comparison between humans and LMs. If language models are insensitive, that indicates a difference between humans and LMs.

*According to finer-grained linguistic criteria, acceptable sentences are those that are grammatical, plausible, and felicious. Please refer to Tonhauser & Matthewson (2015) for detailed definitions and review.

5.3 Comparative Illusion

A canonical comparative illusion surfaces in sentences like *More people have been to Russia than I have*. People accept it at first glance but have trouble pinning down the exact meaning (Montalbetti, 1984). Potential rational nonliteral inference could be "people have been to Russia more times than I have" or "people have been to Russia but I haven't" (O'Connor, 2015; Christensen, 2016). Psycholinguistic research has found that various factors modulate the strength of the illusion, including the repeatability of the event described by the verb phrase (e.g., "vacationed in Florida" vs. "retired to Florida"), the subject form of the than-clause subject (e.g., "... than the student has" vs. "...I have"), as well as the number of that subject (e.g., "I have" vs. "we have")(Wellwood et al., 2018).

We adapted the experimental materials with 32 items from Zhang et al. (2023b) which were the same materials from Chapter 3. An example is in 63 where (63a) is the canonical comparative illusion, (63b) is the acceptable control, and (63c) is the unacceptable one.* We extract the surprisal of the final word for investigation (see the underlined word in (63)). The full experimental design can be seen in Table 5.1.

(63) a. (?) More teenagers have used Tiktok than I have. (illusion)

b. Many teenagers have used Tiktok more than I have. (acceptable)

c. (#) Many teenagers have installed Tiktok more than I have. (unacceptable)

*The repeatability of the verb phrase is responsible for this contrast, as it is more natural to say "use Tiktok more often or frequently" compared with "install Tiktok more often" when the action typically takes place once (in a while).

Illusion type	item	BERT		RoBERTa		GPT-2		GPT-3	
		PPL	Surp	PPL	Surp	PPL	Surp	PPL	Surp
Comparative	32	-0.36	-0.001	-0.56	-0.09	-0.22	-0.05	-0.30	-0.25
Depth-charge	32	-0.37	-0.15	-0.61	-0.45	-0.12	-0.41	-0.37	-0.98
NPI	32	-0.26	-2.46	-0.71	-2.60	-0.21	-1.73	-0.29	-2.55

Figure 5.1: Estimated coefficients of the main effect of acceptability differentiation.

5.3.1 Acceptability differentiation

We first ensured that LMs distinguish acceptable neighbors (63b) of the illusion sentence from unacceptable ones (63c). We ran statistical mixed-effects linear regression models on whole-sentence perplexity and the surprisal at the word *have* for the four language models. Either the perplexity or the surprisal was taken as the dependent variable with the condition "acceptability" as the fixed effect (reference level = the unacceptable condition, with a non-repeatable verb phrase vs. the acceptable condition, with a repeatable verb phrase) and the random intercept of each item as the random effect.*

Figure 5.1† shows the estimated coefficient for the main effect of each mixed-effect model for each LM and each illusion phenomenon. A significant negative estimated coefficient suggests that acceptable sentences received lower perplexity/surprisal compared to the unacceptable ones, indicating that LMs distinguish sentences based on acceptability. Except for surprisal values from BERT and GPT-2, the other six statistical models indicate that the LMs capture the acceptability difference of baseline sentences for the comparative illusion.

*The model syntax in *R* was PPL/SURP ~ acceptability + (1|item).

†If LMs rate acceptable sentences as more acceptable, the coefficients for perplexity or surprisal should be significantly negative. Cells color-coded in blue represent statistical significance level ($p < .05$) in the expected direction. White cells represent an insignificant main effect. In other words, blue cells indicate the statistical model output supports LMs' ability to distinguish sentences based on linguistic acceptability.

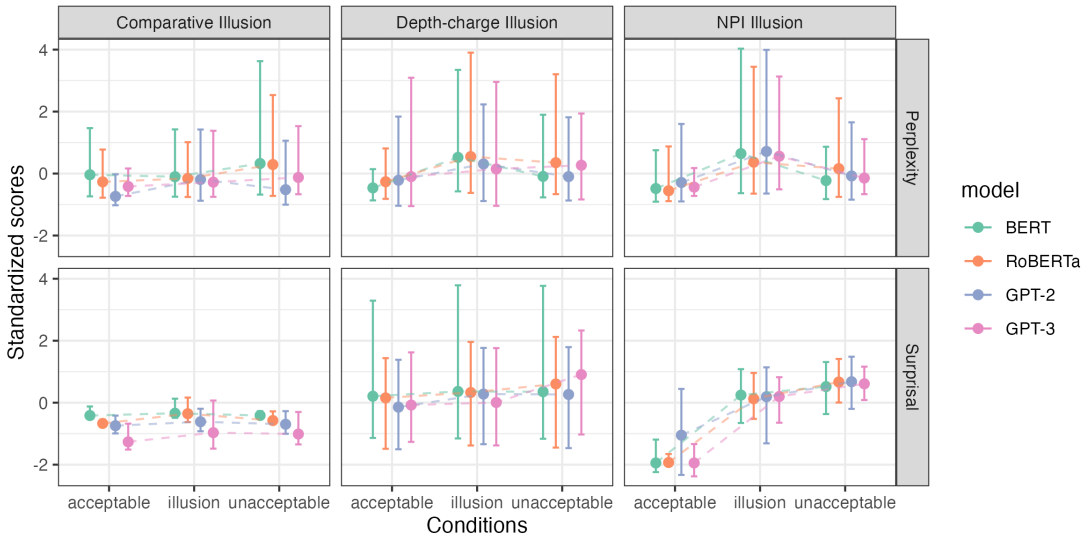


Figure 5.2: Perplexity and surprisal to test the illusion effect of three illusions.

5.3.2 Illusion effect

This task investigated whether language models pattern with humans in demonstrating illusion effects with the basic comparative illusion construction. The contrast involves the illusion condition (63a) with existing control conditions (63b) and (63c). The standardized metrics of the four LMs are displayed in Figure 5.2.*

To statistically evaluate whether LMs capture an illusion effect, we constructed another suite of statistical models across the four LMs and two metrics where the main effect has three levels – the illusion condition (reference), the acceptable condition, and the unacceptable condition – and the random effect included a random intercept for items.†

*If the illusion effect appears, the illusion condition should be rated more acceptable (thus lower in the graph) than the unacceptable condition and therefore has lower perplexity/surprisal. (Error bars are 95% bootstrapped confidence intervals).

†The model syntax in R was `PPL/SURP ~ condition + (1|item)` where `condition` had three levels.

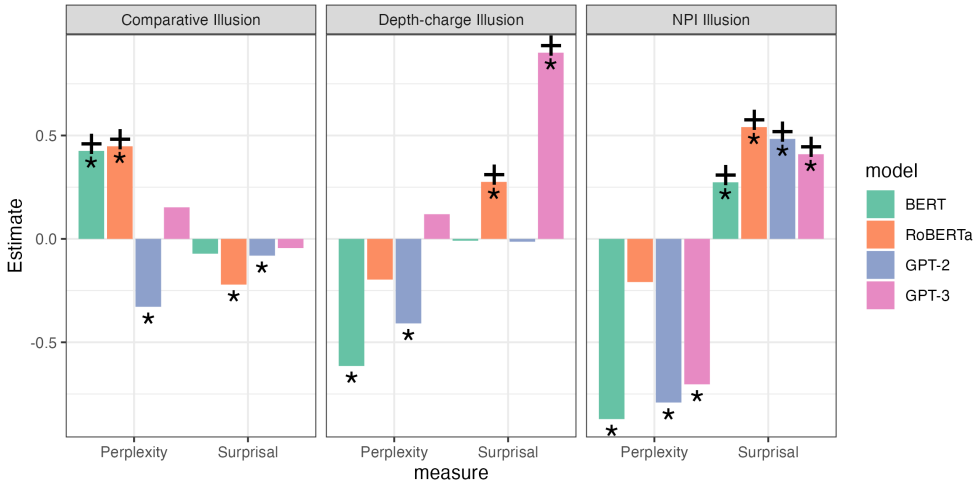


Figure 5.3: Estimated coefficients for the illusion effect in three illusions.

We analyzed the coefficient estimates of the main effect of the unacceptable condition compared with the illusion condition.* An illusion effect would appear with higher perplexity/surprisal for the unacceptable condition compared to the illusion case. In other words, the estimated coefficients for the unacceptable condition should be significantly positive.

Figure 5.3† displays the estimated coefficients for the unacceptable condition compared with the illusion condition. For the comparative illusion, only BERT and RoBERTa measured by perplexity show a human-like illusion effect. Other LM-metric combinations indicate that the illusion condition was rated either the same or worse than the unacceptable condition (contrary to humans).

*The coefficients for the acceptable condition generate similar conclusions. Further, no illusion sentences were rated better than acceptable ones.

†The *y* axis shows the coefficient estimates which represent the increase in perplexity/surprisal when the sentence is unacceptable compared to the illusion case, crossing three language illusions and four LMs. "+" marks a human-like behavior, in this case, an illusion effect where the unacceptable condition receives significantly higher perplexity/surprisal values than the illusion condition. "*" means that the estimated coefficient is significant.

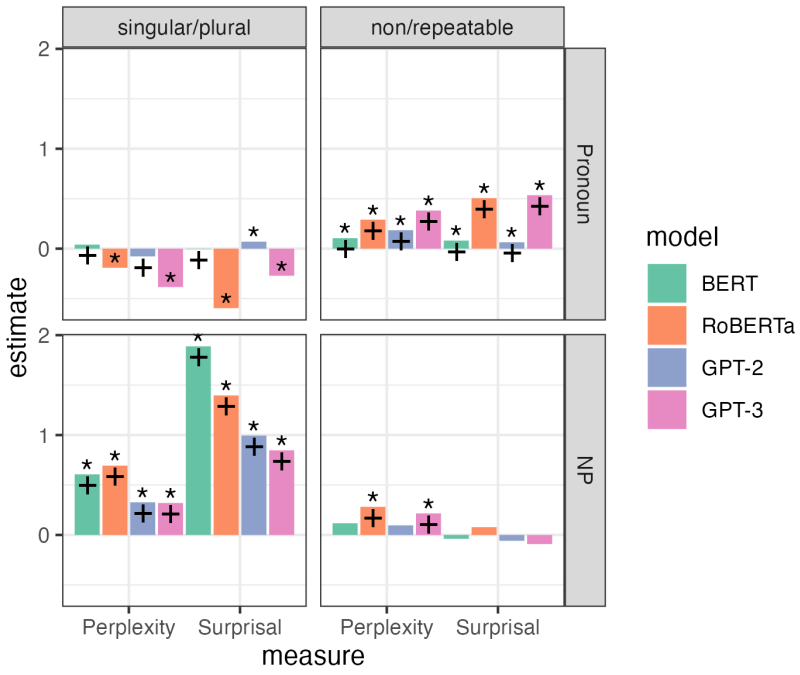


Figure 5.4: Estimated coefficients for critical linguistic manipulations in **comparative illusion**.

5.3.3 Sensitivity to manipulations

In this step, we evaluated whether language models were sensitive to sentence manipulations that affect human judgments. Three factors were investigated: (1) than-clause subject structure (pronoun vs. NP), (2) subject number (singular vs. plural), and (3) verb repeatability (repeatable vs. nonrepeatable). The example materials are represented in Table 5.1. For humans, plural than-clause subjects are more acceptable than singular ones only in the NP case. Overall, repeatable verbs are more acceptable than nonrepeatable ones (O'Connor, 2015; Wellwood et al., 2018; Zhang et al., 2023b).

Figure 5.4* displays the estimated coefficients for the main effects from the statistical

*The *y* axis shows the estimated coefficients for the increase in perplexity/surprisal with respect

COMPARATIVE ILLUSION

Number	**VP**	**Examples**
When the *than*-clause subject is **noun phrase**:		
Singular	Repeatable	More students *have been to Russia* than **the teacher** has.
Singular	Non-repeatable	More students *have escaped from Russia* than **the teacher** has.
Plural	Repeatable	More students *have been to Russia* than **the teachers** have.
Plural	Non-repeatable	More students *have escaped from Russia* than **the teachers** have.
Control	Repeatable	More students *have been to Russia* than **teachers** have. (Good)
Control	Non-repeatable	More students *have escaped from Russia* than **teachers** have. (Good)
When the *than*-clause subject is **pronoun**:		
Singular	Repeatable	More students *have been to Russia* than **I** have.
Singular	Non-repeatable	More students *have escaped from Russia* than **I** have.
Plural	Repeatable	More students *have been to Russia* than **we** have.
Plural	Non-repeatable	More students *have escaped from Russia* than **we** have.
Control	Repeatable	Many students *have been to Russia* more than **I** have. (Good)
Control	Non-repeatable	Many students *have escaped from Russia* more than **I** have. (Bad)

Table 5.1: Full manipulation for the comparative illusion (the same as Tables 3.1 and 3.2)

models.* As for the subject number, when the than-clause subject was a pronoun, only BERT and GPT-2 (with perplexity) aligned with human-like behavior: there is no difference between singular and plural than-clause subjects. When it comes to NP subjects, all four LMs with both metrics showed human-like behavior where the singular NP subject was more unacceptable than the plural NP subject. As for repeatability, all four LMs captured this distinction in the pronoun condition but in the NP condition, only RoBERTa and GPT-3 achieved human-like results with perplexity.

In general, we found substantial overlap between LMs and humans. But none of the language models fully captured all the manipulations.

5.4 Depth-charge Illusion

Consider the most famous depth-charge sentence *No head injury is too trivial to be ignored* Wason & Reich (1979). People overwhelmingly interpret it as meaning "no matter how trivial head injuries are, we should not ignore them", while the literal meaning is the opposite as "we should ignore them".

To understand the depth-charge sentence requires knowing meaning composition rules, multiple negation processing (Wason & Reich, 1979), adequate world knowledge reasoning (Paape et al., 2020), and the neighboring constructions of *too...to* such as *so...that*, *so...as to* and *enough to...* (Zhang et al., 2023c). Since existing research already shows that language

to singular vs. plural than-clause subjects, or nonrepeatable vs. repeatable verb phrases, respectively. "*" means statistically significant contrasts; "+" means human-like results.

*More statistic model information: Iterating over LMs, metrics, and the subject structure (NP vs. pronoun), we initiated statistical models taking both repeatability (reference = repeatable) and subject number (reference = plural) as the main effects with the random effect including a random intercept for the items.

DEPTH CHARGE ILLUSION		
Conditions		**Examples**
Canonical depth-charge		No head injury is too trivial to be ignored.
Plausible control		Some head injury is too severe to be ignored.
Implausible control		Some head injury is too trivial to be ignored.
too...to	plausible	No head injury is too trivial to be treated.
too...to	implausible	No head injury is too trivial to be ignored.
too...to not	plausible	No head injury is too trivial to not be ignored.
too...to not	implausible	No head injury is too trivial to not be treated.
so...as to	plausible	No head injury is so trivial as to be ignored.
so...as to	implausible	No head injury is so trivial as to be treated.

Table 5.2: Full manipulation of materials for the depth-charge illusion

models are quite limited in processing negation (e.g. Kassner & Schütze, 2019; Ettinger, 2020), we speculate that LMs might encounter difficulty in the more complicated case of depth-charge sentences.

The evaluation materials were adapted from Zhang et al. (2023c) and Chapter 2 with 32 items. An example is (64) where we took the surprisal of the sentence-final word for comparison. Table 5.2 shows, in the example of head injuries, the condition manipulations of the materials across the three tests.

(64) a. (?) No head injury is too trivial to be ignored. (depth-charge sentence)

b. Some head injury is too severe to be ignored. (plausible, acceptable)

c. (#) Some head injury is too trivial to be ignored. (implausible, unacceptable)

5.4.1 Acceptability differentiation

Utilizing the same methodology as the comparative illusion, we found, as depicted in Figure 5.1, that all combinations of LMs and metrics, except GPT-2 (perplexity), captured the acceptability difference between (64b) and (64c) with a significantly lower perplexity/surprisal for the acceptable sentences like (64b).

5.4.2 Illusion effect

Next, we studied if LMs "experience" the illusion effect by assigning lower perplexity/surprisal scores to the depth-charge sentence (64a) compared to the unacceptable one (64c).

Our statistical results show, in Figure 5.3, that only RoBERTa and GPT-3 demonstrated an illusion effect (for surprisal) by assigning a significantly higher score to the unacceptable control sentences. This means that it is not easy to "trick" LMs with the depth-charge illusion. Similar results have led concurrent work to suggest that LMs are better at deriving the literal meaning of a sentence, which is in sharp contrast with the overwhelming illusion effect from humans (Paape, 2023, a.o.).

5.4.3 Sensitivity to manipulations

This task tested LMs' sensitivity to the plausibility contrast of three near-neighbor pairs of the depth-charge sentence. These pairs differ by the degree quantifier construction (*too…to* vs. *so…as to* vs. *too…to not*) (see Table 5.2). Competent language models should differentiate plausible sentences from implausible ones.

Figure 5.5* displays estimated coefficients of statistical models' main effect. We expect implausible sentences to receive higher perplexities/surprisals when the illusion occurs.† We find that LMs captured some of the distinctions in the *too...to* condition and the *so...as to* condition. However, implausible sentences with *too...to not* were rated as more acceptable than their plausible counterparts, which flouts what linguistic rules predict.‡ The fact that "No head injury is too trivial to be treated" and "No head injury is too trivial to not be ignored" generate opposite results while having the same meaning suggests LMs still struggled with negation, antonyms, and meaning composition (She et al., 2023; Truong et al., 2023).

5.5 Negative Polarity Illusion

Negative polarity items and their licensing conditions have been investigated in prior work with language models. For a canonical NPI (e.g., *ever*, *any*) to be acceptable, it has to be in a downward entailing environment – in this specific case, in the scope of negation.§ Existing computational research has shown that the syntactic dependency between the licensor and the NPI is captured by language models (Jumelet & Hupkes, 2018; Jumelet et al., 2021; Shin et al., 2023) but with more difficulty as compared to subject-verb agreement or other syntactic dependencies (Marvin & Linzen, 2018; Warstadt et al., 2019, 2020). In this

*The *y* axis shows the increase in perplexity/surprisal when the sentence is implausible vs. plausible. "*" means statistically significant contrasts; "+" means human-like behavior. While we see differences among LMs and metrics in the "no...so...as to" and the "no...too...to" conditions, the condition of "no...too...to not" yielded completely opposite results to humans.

†Iterating over sentence pairs, LMs, and metrics, we ran mixed-effects linear regression models on scores over the plausibility contrast (reference = plausible).

‡The sentence *No head injury is too trivial to not be ignored* should be plausible because compositionally, "too trivial to not be ignored" means "too trivial to be treated" which yields a plausible sentence given the sentential negation.

§The licensing conditions of negative polarity items are far more than in the scope of negation. We focus on the classic licensing condition and refer to Penka (2015) for a review.

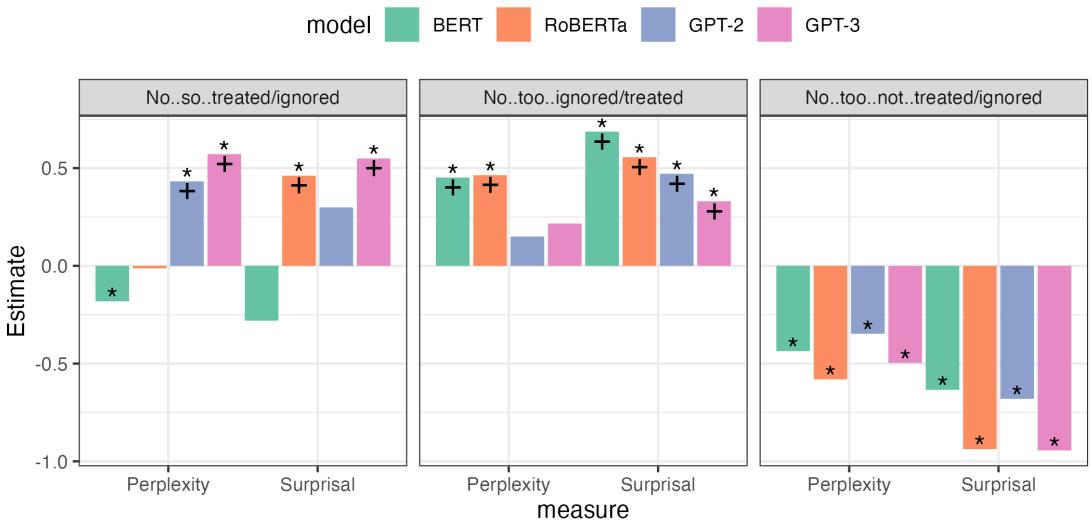


Figure 5.5: Estimated coefficients for the plausibility contrast in **depth-charge illusion**.

task, we expanded the suite of LMs and metrics and explored sensitivities to four types of licensors.

Our materials came from Orth et al. (2021) with 32 items. The essential triad is (65) where the illusion condition has the NPI *ever* not in the scope of the negation word *no*.

(65) a. The hunter who no villager believed to be trustworthy will ever shoot a bear. (negative polarity illusion)

b. No hunter who the villager believed to be trustworthy will ever shoot a bear. (Matrix No, acceptable)

c. (*) The hunter who the villager believed to be trustworthy will ever shoot a bear. (Licensor Absent, unacceptable)

5.5.1 Acceptability differentiation

Figure 5.1 shows that all the four LMs could capture the acceptability difference of control sentences (65b) and (65c) (with both perplexity and surprisal).

5.5.2 Illusion effect

Figure 5.3 shows that only in the case of surprisal did we see an illusion effect where the unacceptable sentences (65c) received significantly higher surprisals than the illusion sentence (65a). This finding replicates Shin et al. (2023) in that, for the illusion condition (65a) where *no* linearly precedes *ever* but is in an unlicensing position, *ever* incurs higher surprisal. It is interesting to see the sharp discrepancy between surprisal and perplexity, which we leave to Section 5.6.4 for discussion.

5.5.3 Sensitivity to variations

The linguistic manipulations we explored in this section concern the illusion effect with different NPI licensors. Among the ones we tested, *didn't*, *did not*, and *never*. Please refer to Table 5.3 for the full experimental conditions. Human research shows that none of these triggers illusion effects (Orth et al., 2021).

Iterating over licensors, LMs, and metrics, we ran statistical models with the same structure in section 5.5.2. We plotted the estimated coefficients of the unacceptable main effect in Figure 5.6* and predicted that a significantly positive coefficient indicates an illusion effect.

*The *y* axis shows the increase in perplexity/surprisal when the sentence is ungrammatical vs. is in the illusion condition. "+" marks an illusion effect while none of the three licensors should trigger an illusion effect according to human behavior; "*" means a significant contrast.

NPI ILLUSION	
Conditions	**Examples**
Matrix No	**No** hunter who the villager believed to be trustworthy will ever shoot a bear.
Licensor Absent	The hunter who the villager believed to be trustworthy will ever shoot a bear.
Relative No	The hunter who **no** villager believed to be trustworthy will ever shoot a bear.
Relative Didn't	The hunter who **didn't** believe the villager to be trustworthy will ever shoot a bear.
Relative Did not	The hunter who **did not** believe the villager to be trustworthy will ever shoot a bear.
Relative Never	The hunter who **never** believed the villager to be trustworthy will ever shoot a bear.

Table 5.3: Full manipulation for the negative polarity illusion

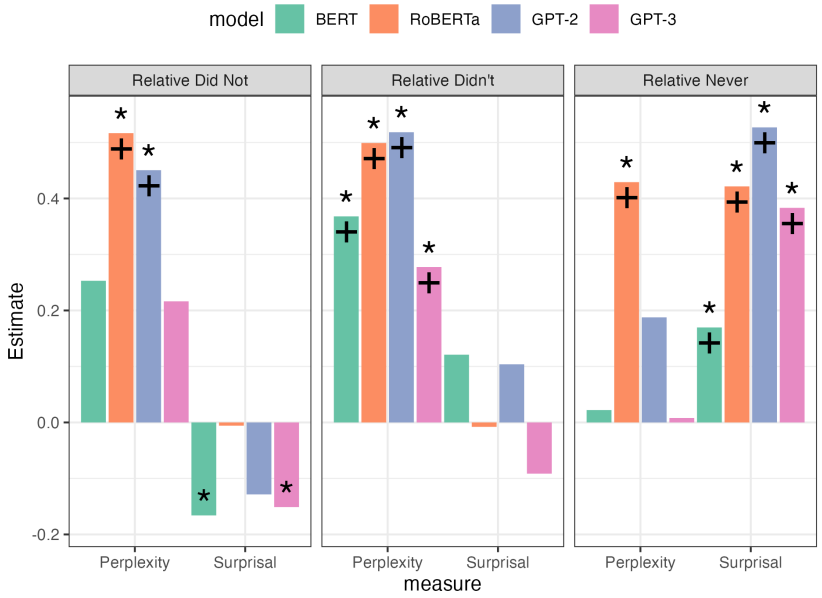


Figure 5.6: Estimated coefficients for the illusion effect in **negative polarity illusions**.

Contrary to human-like behavior, for all three licensors there were some LM-metric combinations that indicate an illusion effect: for the licensor *did not*, RoBERTa (perplexity) and GPT-2 (perplexity) show an illusion effect; for *didn't*, all four LMs with perplexity show an illusion effect; for *never*, all four LMs with surprisal, plus RoBERTa with perplexity, show an illusion effect. This pattern shows that with negative polarity illusions, LMs are more easily tricked than humans.

5.6 Discussion

5.6.1 Illusion effect

Successful language processing requires a dynamic integration of lexical knowledge, grammatical knowledge, logical reasoning, and world knowledge, among other cognitive abilities and sources of knowledge. An illusion effect in humans where unacceptable sentences receive unexpectedly high acceptability presents a unique case where the comprehender might prioritize different processing mechanisms or linguistic constraints during processing. Studying how language models process language illusions helps us understand (1) from a superficial level, whether LMs appear to be human-like in the acceptability judgment tasks of language, and (2) from a deeper level, whether we can make inferences about LMs employing the same set of resources and abilities to process a sentence (i.e., whether they can serve as cognitive models). For example, as argued by the previous chapters, one question to ask is whether LMs process illusive materials in a rational way as humans do.

In this research, by studying four language models' acceptability judgments of three language illusions, we found that LMs were good at differentiating basic acceptable sentences from unacceptable ones, and yet no LMs showed consistent human-like illusion effects

		BERT		RoBERTa		GPT-2		GPT-3	
		PPL	Surp	PPL	Surp	PPL	Surp	PPL	Surp
Comparative illusion	Acceptability differentiation	✓		✓		✓	✓	✓	✓
	Illusion effect	✓		✓					
	Number effect: Pronoun	✓	✓			✓			
	Number effect: NP	✓	✓	✓	✓	✓	✓	✓	✓
	Repeatability: Pronoun	✓	✓	✓	✓	✓	✓	✓	✓
	Repeatability: NP			✓				✓	
Depth-charge illusion	Acceptability differentiation	✓	✓	✓	✓		✓	✓	✓
	Illusion effect				✓				✓
	Plausibility contrast (so...as to)				✓	✓		✓	✓
	Plausibility contrast (too...to)	✓	✓	✓	✓		✓		✓
	Plausibility contrast (too...to not)								
NPI illusion	Acceptability differentiation	✓	✓	✓	✓	✓	✓	✓	✓
	Illusion effect for Relative No		✓		✓		✓		✓
	Illusion effect for Relative Did not	✓	✓		✓		✓	✓	✓
	Illusion effect for Relative Didn't		✓		✓		✓		✓
	Illusion effect for Relative Never	✓				✓		✓	

Figure 5.7: Language models' performance on all three illusions. (✓ means LMs show human-like behavior.)

among three illusion phenomena by any metric (Figure 5.7). We conclude from this result that LMs might not be a good cognitive model of human language processing. At least, they do not adopt a rational approach to language comprehension.

With this said, we do observe a divergence between the comparative/depth-charge illusion and the negative polarity illusion – it seems more likely for LMs to be tricked by the negative polarity illusion compared to the former two. Since the negative polarity illusion is more relevant to the hierarchical structure of language and its interaction with meaning, whereas both the comparative illusion and depth-charge illusion clearly emphasize semantic nuances, we tentatively conclude that LMs are more easily tricked by the illusion which is caused by a structural anomaly, compared to the illusion which is caused by meaning. Yet, since the inner working of LMs is still a black box, we cannot make causal reasoning of whether LMs learn to compose the meaning of the depth-charge sentence or the comparative illusion sentence, or whether LMs are bad at detecting the hierarchical relation in a sentence.

5.6.2 Human-like behaviors & Potential processing mechanisms

For both the comparative illusion and depth-charge illusion, the illusion effect test did not show human-like behavior. This could either mean that LMs strictly abide by linguistic rules to compose the language literally or that LMs have trouble understanding this complicated set of sentences overall. For the comparative illusion, the sensitivity task (section 5.3.3) suggests that they might have some capacity to process comparative structures. For the depth-charge illusion, that LMs seem to have trouble understanding the literal contrast between plausible/implausible pairs (section 5.4.3) suggests sentences involving multiple negations could pose a challenge to LMs. The two cases indicate we still need to develop more robust evaluations to gauge LMs' semantic capabilities in various semantic domains.

For the negative polarity illusion, the interpretation could be more complicated. On one hand, the illusion test for the licensor *no* yields human-like results (with surprisal) but other licensors also elicit non-human-like illusion effect (cf. Orth et al., 2021). On the other hand, the discrepancy between sentence perplexity and surprisal makes it difficult to conclude to what degree LMs and humans overlap (cf. Shin et al., 2023).

Ultimately, we want to address whether LMs are like humans that utilize not only grammatical rules but also contexts, frequencies, and semantic priors to rationally process language, or LMs are like grammarians that interpret string inputs in a strict compositional manner. Our investigation does not yield consistent results given the three language illusions but the behavioral inconsistency suggests that language models are far from being a cognitive model of human language.

5.6.3 Language models' performance in general

All four language models performed on par with each other. If we tallied the number of tests where LMs reported expected results from Figure 5.7 and averaged between perplexity and surprisal, we have a ranking order from RoBERTa (N=10) and GPT-3 (N=9), to BERT (N=8.5) and GPT-2 (N=8). The successors of both the masked language model and the autoregressive model perform better than their predecessors.

5.6.4 Perplexity & Surprisal

It is surprising to see that the two widely used probability-based metrics can generate different results for a given hypothesis and a given language model. Future work should (i) investigate both mathematically and practically why the difference could occur and (ii) check if better definitions for the critical regions exist to capture surprisals. Future evaluation work that utilizes one metric should be mindful of the intrinsic limitations of that metric.

5.6.5 Limitations

Considering the research methodology, acceptability judgment tasks (even with carefully controlled minimal pairs) are indirect measures of language comprehension and it is hard to infer the exact interpretation based on probability-based measures. Further studies should work on direct comprehension measures (e.g., generating paraphrases) that reveal LMs' hidden knowledge. Another question is whether the research paradigm of human judgments would make a difference when being aligned with LM performance. For instance, the human data for the depth-charge and the comparative illusion sentences is based on untimed acceptability rating (Zhang et al., 2023b,c) while the negative polarity illusion is based on

speeded judgment tasks. A better alignment of comparable meta-linguistic features between humans and LM algorithms is needed for making replicable, appropriate, and scalable evaluations.

5.7 Conclusion

We tested four language models' ability to process three language illusions and asked (1) whether they judge unacceptable illusion sentences to be more acceptable as humans (termed an illusion effect) and (2) whether they are sensitive to linguistic manipulations that modulate human judgments. Our results are based on whole-sentence perplexity and critical word surprisal. We show that none of the LMs demonstrated consistent illusion effects or exhibited overall human-like judgment behaviors. We conclude that given the case of language illusions, language models neither behave like humans with full sets of cognitive abilities and rational behavior, nor possess the necessary linguistic knowledge for error-free, literal sentence processing.

6

Concluding Remarks and Future Work

6.1 Summary of the Thesis

This thesis examines the processing patterns of three linguistic phenomena called language illusions and provides a unified account from the rational approach of language comprehension (Futrell et al., 2020; Gibson et al., 2013a; Hahn et al., 2022; Levy, 2008; Shannon, 1948). Language illusions refer to sentences that are ungrammatical, semantically ill-

formed, or implausible but are accepted by the native speakers of that language.

The three language illusions under investigation are the depth-charge illusion, the comparative illusion, and the negative polarity illusion, examples shown in (66).

(66) a. No head injury is too trivial to be ignored. (Depth-charge illusion)

b. More people have been to Russia than I have. (Comparative illusion)

c. Many authors that few critics recommended have ever received acknowledgment for a best-selling novel. (Negative polarity illusion)

The problem with the depth-charge illusion as in (66a) is that comprehenders almost always interpret it as, e.g., "head injuries should **not** be ignored" while the literal meaning is the opposite, as "head injuries should be ignored." (e.g., Wason & Reich, 1979; Paape et al., 2020). The uniqueness of this example is that it is filled with expressions with a negative meaning which increases the linguistic complexity immensely. Regardless of *no*, *too X to Y* means something is so X that it becomes harder to realize Y, *trivial* conveys the lack of importance, and *ignore* conveys the lack of action. In addition, the common takeaway is heavily influenced by world knowledge. In **Chapter 2**, with four experiments, my coauthors and I identified the comprehension patterns of different depth-charge items and we found that apart from the famous and canonical example of head injury, other items exhibit a wide range of plausibility ratings which were correlated with the world knowledge rating. We claim that the depth-charge sentence can be regarded as a sentence distorted by production errors which could have been, for example, *No head injury is **so** trivial **as to** ignored*. The structural substitution from the intended *so...as to* to the produced *too...to* is motivated by first the structural and semantic similarity between the two structures and second the higher frequency of *too...to*. Then we assume that while comprehenders process the depth-charge

sentence, they rationally infer what the perceived sentence is intended to mean. The probability of making that inference can be mathematically modeled in the Bayesian framework with quantitative predictions that can be tested via behavioral experiments:

$$P(s_i|s_p) \propto P(s_i)P(s_p|s_i) \tag{6.1}$$

There, we found that the probability of rational inference over the depth-charge sentence, $P(s_i|s_p)$, was positively correlated with the world knowledge prior, $P(s_i)$, and how likely comprehenders conceive an error could take place which results in the perceived sentence, $P(s_p|s_i)$.

Interestingly, the same approach could also apply to the comparative illusion. There, the problem with the comparative illusion is that if the verb ellipsis is reconstructed, as in (67), there is no appropriate quantity or degree in the subordinate clause that can be compared with the matrix clause.

(67) More people have been to Russia than I have been to Russia.

In **Chapter 3**, we hypothesized that the comparative illusion can be regarded as the perceived s_p that can have multiple s_is, given empirical evidence that comprehenders did interpret the sentence differently. One example is *People have been to Russia more than I have* where it is the frequency or times of being in Russia that is compared between other people and the speaker. The theory is that comprehenders would choose the interpretation that has a high likelihood of being produced into the comparative illusion sentence. In other words, assuming similar $P(s_i)$s across these interpretations, higher $P(s_p|s_i)$s lead to higher $P(s_i|s_p)$. This is what we found through four experiments in Chapter 3.

The depth-charge illusion and the comparative illusion suggest that humans rely on their prior world knowledge and linguistic knowledge to decipher uncertain linguistic input that comes with errors. This suggests that language comprehension is a rational behavior where comprehenders choose the more efficient route of integrating multiple resources to achieve an interpretation, which could just be good-enough (e.g., Ferreira et al., 2002; Ferreira, 2003; Ferreira & Patson, 2007), rather than stick to incremental and compositional processing of complicated materials with the cost of more memory or attention consumption.

While the depth-charge sentence and the comparative illusion sentence deal with offline interpretations, the negative polarity illusion focuses on the online processing of sentences, especially with a dependency structure (e.g., The NPI licensor *few* and the NPI *ever*). In **Chapter 4**, we show that the noisy-channel framework can be generalized to explain characteristics of the online processing. Here, the hypothesis is that while comprehenders process sentences like (66c) and before reaching the NPI, they could mentally store a distorted representation of the preceding context such that this distorted one is (68). This distorted mental representation of sentences is extensively studied by Futrell et al. (2020) and Hahn et al. (2022). The distortion is manifested as a position exchange between *few* and *many* and the result sequence licenses the NPI *ever*. We also designed an empirical study to support the distortion of *few* and *many*. The reason for a possible switch of *few* and *many* is that they belong to "vague quantifiers" whose external quantity is not always clear and is heavily dependent on contexts (e.g., Partee, 1988; Solt, 2011). *Few* and *many* are similar in this dimension, which exactly facilitates the position exchange at little cost of interpretation accuracy. This aspect supports a rational processing account of the negative polarity illusion – while there is little or negligible cost of misrepresentation, comprehenders could mentally

switch the position of *few* and *many* to obtain an easier processing strategy.

(68) **Few** authors that **many** critics recommended have ever received acknowledgment for a best-selling novel.

With the knowledge of these language illusions, **Chapter 5** was set out to investigate whether large language models (e.g., BERT, RoBERTa, GPT-2, GPT-3) can also be tricked by language illusions. We yielded inconsistent results mainly due to the discrepant performances by the whole-sentence perplexity and the critical-region surprisal for a single model and a single test. While more understanding is needed before reaching a conclusion, we argue that this study shows that LMs cannot be viewed as cognitive models of human language processing, given the inconsistent results.

6.2 Discussion & Future Work

6.2.1 Language Illusion

The investigation of language illusions has been going on for a while. Apart from the three illusions in this thesis, some famous ones are listed in (69), each presenting a different online processing and illusion profile (Phillips et al., 2011).

(69) a. How many of each type of animal did Moses take on the ark?

(Erickson & Mattson, 1981)

b. When an airplane crashes, where should the survivors be buried?

(Barton & Sanford, 1993)

c. The key to the cabinets are on the table.

(Bock & Miller, 1991)

d. The runners who the driver see each morning always wave.

(Bock & Miller, 1991)

e. The patient who the nurse who the clinic had hired met Jack.

(Janet Fodor's observation mentioned in Frazier (1985))

These language illusions differ from each other in at least two aspects. First, the anomalies occur at different levels of linguistic representations. The Moses illusion (69a) and its variant (69b) concern a lexical anomaly. The subject-verb agreement attraction in (69c) and (69d) deals with grammatical features within the dependency structures. The structural forgetting sentence (69e) and the negative polarity illusion (Chapter 4) involve long-distance dependency structures with complicated structural design. The comparative illusion (Chapter 3) involves the well-formedness in a semantic level. The depth-charge illusion (Chapter 2) involves meaning plausibility that is evaluated under world knowledge, similar to the Moses illusion. This multi-level representation is intuitive because language can certainly induce errors at every level of its representation.

The second dimensionality of difference is how much each illusion is related to language production versus language comprehension. While the subject-verb agreement error and the depth-charge sentence have been reported to appear in production studies or the production corpora (Fortuin, 2014; Phillips et al., 2011), to the best of my knowledge, other illusions have relatively fewer spots in production. Distinguishing the side of occurrence might be helpful for us to understand the similarity and the distinction between language production and language comprehension.

Another direction with important theoretical contributions but challenging implementation is to ask how much of the processing involves a conscious detection of the anomaly.

While it is not hard to envision that mental error correction is part of rational inference, a more critical question is whether we can still call it rational inference if the errors are not detected. According to Levy (2008), rational inference can go under consciousness but there seems to be different interpretations of that theory (e.g., Paape, 2024). Future work could say more about the mechanistic level of rational inference and how error detection is involved in this process.

A very exciting direction for future research is whether we can find a systematic explanation to account for the reasons behind all language illusions and at the same time preserve individual specificity. This brings the discussion of the next section on the rational processing of language, which could posit a potential universal solution to these language illusions.

6.2.2 Rational Processing of Language

I have had the great fortune to learn about rationality in this line of work. The more I get to the fundamentals, the sooner I realize that I need to catch up with the history of rationality and have a good working definition of rational processing. It seems that without a clear definition of rationality, the rational processing of language illusions seems to lack an essential piece in the jigsaw puzzle. In this section, I will attempt to provide my own potential definition of rationality in the context of language processing, by synthesizing existing work on the rational processing of language (e.g., Gibson et al., 2013b; Levy, 2008; Futrell et al., 2020; Hahn et al., 2022) and cross-disciplinary work on rational analysis (Anderson, 1990), bounded rationality (Simon, 1964) and broader economics (e.g., Mankiw, 2018).

The term "Rationality" or "rational behavior" is loaded with multiple senses and enjoys

a long history of academic investigation.* There are two major senses as described in Anderson (1990). The first one can be called "normative rationality" which states that humans explicitly engage or should engage in logically correct reasoning or that human behaviors are guided by reason. This notion is closer to the etymology of the word *rational* which is "pertaining to or springing from reason"†. As we know, this is not always true and people sometimes adopt heuristics that lead to severe and systematic errors (e.g., Tversky & Kahneman, 1974). The second sense can be called "adaptive rationality" which means that human behavior is optimal in terms of achieving human goals. Adaptive also hints that humans adjust their strategies to adapt to the changing environments given certain conditions and constraints (Simon, 1964). This notion of rationality is directly relevant to the concept of "rational agent" in economics which essentially means that "rational people systematically and purposefully do the best they can to achieve their objectives, given the available opportunities." (e.g., Mankiw, 2018). Anderson (1990) further argues that the criteria for normative rationality can deviate from the criteria for adaptive rationality, which gives us another reason to focus on the adaptive notion of rationality.

While the adaptive notion of rationality emphasizes goal achieving and optimization, another necessary property of rationality is its boundedness (Simon, 1964). "Bounded rationality" points out that during the optimization of a certain strategy or decision, humans cannot enumerate all the possible options and calculate the best one due to constraints on external factors such as time and internal factors such as their information processing capacity. Just like playing chess, the player cannot mentally calculate all the next steps (even with the successive steps) and choose the best one because the chess match does not allow

*Wikipedia page on "Rationality": https://en.wikipedia.org/wiki/Rationality.
†Sources from Online Etymology Dictionary.

strategizing for too long and the player usually cannot think through all the possibilities. Therefore, in Simon (1964)'s term, they fall back to the rule of thumb, i.e., heuristics, or conventions of the chess game to make a move out of a subset of possible options.

In general, rational behaviors entail a goal, an optimizing procedure to achieve the goal, and limitations on the resources employable to execute the optimization. Because the resources are limited and heuristics are taken, it is not always possible to truly **optimize**. Instead, humans settle with a **satisficing** result (Simon, 1972), in other terms, a good enough result. The way to quantitatively predict rational human behavior is through the probabilistic Bayesian methods (Chater & Oaksford, 2021). Against this background, we have the rational approach to language comprehension. The essence of this approach is that, with the goal to maximize information gain and comprehension accuracy, and with the constraints from human cognition as well as the environment, comprehenders integrate their knowledge about the grammar, the context, and the world to arrive at an interpretation. When the resource is limited and 100% accuracy cannot be achieved, they resort to their probabilistic knowledge to assign a good-enough interpretation.

It is under such scaffolding of rational processing of language that I present an account for the three types of language illusions. For the depth-charge illusion, since spending an extra amount of energy to do compositional processing does not yield excessive return in information, and relying on world knowledge seems enough to derive a piece of meaning that seems to fulfill the communicative purpose, the understanding of *No head injury is too trivial to be ignored* would stop at "we should not ignore head injuries no matter how trivial they are". For the comparative illusion, since the goal is to maximize our understanding of the sentence and we (probably) had better generate useful information from it, even though

comprehenders might detect the anomaly (Paape, 2024), they still reconstruct a plausible meaning as the result of rational processing. For the negative polarity illusion, since *few* and *many* are so similar that exchanging their position might not cause detrimental information inaccuracy or loss, comprehenders mentally do the switch to allow the dependency to be formed by the NPI *ever* and its licensor *few*. Of course, all three cases show that there are inaccuracies in information representation, but the rational story tells us that as long as comprehenders control the inaccuracy cost under a tolerable threshold, they are satisfied with making inferences over linguistic input with noises and uncertainty.

Future work on the rational processing of language should go both deeper and broader. Theoretically, it is worthwhile to test when the goals, environmental conditions, and cognitive constraints change, how to mathematically predict the language processing behavior. Broadly speaking, it is necessary and extremely exciting to see how generalizable the framework is to account for other language processing phenomena. Practically, getting the empirical landscape right would require the maturity from the theory, the mathematical model, the experimental methodology (e.g., judgments, reading behaviors, and neuroimaging), and a holistic understanding of language. Each and every combination of these areas would bring fruitful advances and I am very looking forward to the next chapter in the study of human language processing.

A

Other figures

Figure A.1: The visualization of the noise likelihood task in Experiment 4 for the comparative illusion

References

Allen, M. P. (2004). *Understanding regression analysis*. Springer Science & Business Media.

Anderson, J. R. (1990). *The adaptive character of thought*. New Jersey: Lawrence Erlbaum Associates, Inc.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, 29(3), 313–341.

Anderson, J. R. & Lebiere, C. J. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates, Inc.

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic Surprisal From Neural Models Predicts, But Underestimates, Human Processing Difficulty From Syntactic Ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (pp. 301–313). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.

Barton, S. B. & Sanford, A. J. (1993). A case study of anomaly detection: Shallow semantic processing and cohesion establishment. *Memory & cognition*, 21(4), 477–487. Publisher: Springer.

Bock, K. & Levelt, W. (1994). Grammatical encoding. (pp.40).

Bock, K. & Miller, C. A. (1991). Broken agreement. *Cognitive psychology*, 23(1), 45–93.

Box, G. E. & Tiao, G. C. (2011). *Bayesian inference in statistical analysis*. John Wiley & Sons.

Bremnes, H. S., Szymanik, J., & Baggio, G. (2022). Computational complexity explains neural differences in quantifier verification. *Cognition*, 223, 105013. Publisher: Elsevier.

Bremnes, H. S., Szymanik, J., & Baggio, G. (2023). The interplay of computational complexity and memory load during quantifier verification. *Language, Cognition and Neuroscience*, (pp. 1–23).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.

Bürkner, P. (2024). Interpreting results from categorical() with brm - Interfaces / brms.

Bürkner, P.-C. (2017). **brms** : An *R* package for Bayesian multilevel models using *Stan*. *Journal of Statistical Software*, 80(1).

Bürkner, P.-C. & Vuorre, M. (2019). Ordinal regression models in psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1), 77–101. Publisher: SAGE Publications Inc.

Büttner, A. C. (2007). Questions versus statements: Challenging an assumption about semantic illusions. *Quarterly Journal of Experimental Psychology*, 60(6), 779–789. Publisher: SAGE Publications.

Cantor, A. D. & Marsh, E. J. (2017). Expertise effects in the Moses illusion: detecting contradictions with stored knowledge. *Memory*, 25(2), 220–230. Publisher: Routledge _eprint: https://doi.org/10.1080/09658211.2016.1152377.

Chater, N. & Oaksford, M. (2021). Bayesian rationality in the psychology of reasoning. In M. Knauff & W. Spohn (Eds.), *The Handbook of Rationality* (pp. 285–294). The MIT Press.

Chen, S., Nathaniel, S., Ryskin, R., & Gibson, E. (2023). The effect of context on noisy-channel sentence comprehension. *Cognition*, 238, 105503.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Christensen, K. R. (2010). Syntactic reconstruction and reanalysis, semantic dead ends, and prefrontal cortex. *Brain and Cognition*, 73(1), 41–50.

Christensen, K. R. (2016). The dead ends of language: The (mis) interpretation of a grammatical illusion. *Let us have articles betwixt us–Papers in Historical and Comparative Linguistics in Honour of Johanna L. Wood. Aarhus: Aarhus University*, (pp. 129–160).

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42(4), 368–407. Publisher: Elsevier.

Cook, P. & Stevenson, S. (2010). No sentence is too confusing to ignore. In *Proceedings of the 2010 Workshop on NLP and Lingusitics: Finding the Common Ground* (pp.9). Uppsala, Sweden.

Dahl, Ö. (1981). On the definition of the telic-atelic (bounded-nonbounded) distinction. In *Tense and aspect* (pp. 79–90). Brill.

Davies, M. (2015). Corpus of Contemporary American English (COCA).

Davis, F. (2022). Incremental processing of principle b: Mismatches between neural models and humans. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)* (pp. 144–156).

Davis, F. & van Schijndel, M. (2020). Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 396–407). Online: Association for Computational Linguistics.

de Dios-Flores, I., Muller, H., & Phillips, C. (2017). Negative polarity illusions: licensors that don't cause illusions, and blockers that do.

Dell, G. S. & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, 20(6), 611–629.

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, predictive parsing with psycholinguistically motivated Tree-Adjoining Grammar. *Computational Linguistics*, 39(4), 1025–1066.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, (pp. 145–165).

Erickson, T. D. & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, 20(5), 540–551.

Ettinger, A. (2020). What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8, 34–48. Place: Cambridge, MA Publisher: MIT Press.

Farkas, D. (1981). Quantifier scope and syntactic islands. In *Papers from the Regional Meeting. Chicago Ling. Soc. Chicago, Ill*, number 17 (pp. 59–66).

Fay, D. (1981). Substitutions and splices: a study of sentence blends. *Linguistics*, 19(7-8).

Fedorenko, E., Ryskin, R., & Gibson, E. (2022). Agrammatic output in non-fluent, including Broca's, aphasia as a rational behavior. *Aphasiology*, (pp. 1–20).

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2), 164–203.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), 11–15.

Ferreira, F. & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1(1-2), 71–83. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1749-818X.2007.00007.x.

Fortuin, E. (2014). Deconstructing a verbal illusion: The 'No X is too Y to Z' construction and the rhetoric of negation. *Cognitive Linguistics*, 25(2).

Frazier, L. (1985). Syntactic complexity. *Natural language parsing: Psychological, computational, and theoretical perspectives*, (pp. 129–189).

Fromkin, V. (1980). *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*, volume 171. Academic Press.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, 44(3), e12814. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12814.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. arXiv:1903.03260 [cs].

Garey, H. B. (1957). Verbal aspect in French. *Language*, 33(2), 91–110. Publisher: Linguistic Society of America.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4), 457–472.

Giannakidou, A. (2002). Licensing and sensitivity in polarity items: from downward entailment to nonveridicality. *CLS*, 38, 29–53.

Giannakidou, A. (2006). Only, emotive factive verbs, and the dual nature of polarity dependency. *Language*, (pp. 575–603).

Giannouli, V. (2016). A verbal illusion reexamined. *Acta Neuropsychologica*, 14(4), 324–329.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Doctoral dissertation, Carnegie Mellon University.

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

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000, 95–126.

Gibson, E. (2024). *Syntax: A cognitive approach*. MIT Press.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013a). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20), 8051–8056. Publisher: Proceedings of the National Academy of Sciences.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using mechanical turk to obtain and analyze english acceptability judgments: Linguistic acceptability on mechanical turk. *Language and Linguistics Compass*, 5(8), 509–524.

Gibson, E., Piantadosi, S. T., Brink, K., Bergen, L., Lim, E., & Saxe, R. (2013b). A noisy-channel account of crosslinguistic word-order variation. *Psychological Science*, 24(7), 1079–1088. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Gibson, E., Sandberg, C., Fedorenko, E., Bergen, L., & Kiran, S. (2016). A rational inference approach to aphasic language comprehension. *Aphasiology*, 30(11), 1341–1360.

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't underestimate the benefits of being misunderstood. *Psychological Science*, 28(6), 703–712.

Gibson, E. & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3), 225–248. Publisher: Routledge _eprint: https://doi.org/10.1080/016909699386293.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E. & Ferreira, F. (2022). Good-enough language production. *Trends in Cognitive Sciences*, 26(4), 300–311.

Hacquard, V. (2005). Aspects of 'too' and 'enough' constructions. *Semantics and Linguistic Theory*, (pp. 80–97).

Hahn, M., Futrell, R., Levy, R., & Gibson, E. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43), e2122602119. Publisher: Proceedings of the National Academy of Sciences.

Harley, T. A. (1984). A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science*, 8(3), 191–219.

Harley, T. A. & MacAndrew, S. B. G. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research*, 30(4), 395–418.

Heim, I. (2000). Degree operators and scope. *Semantics and Linguistic Theory*, (pp. 40–64).

Heim, I. & Kratzer, A. (1998). *Semantics in generative grammar*. Blackwell.

Heim, S., Peiseler, N., & Bekemeier, N. (2020). "few" or "many"? an adaptation level theory account for flexibility in quantifier processing. *Frontiers in psychology*, 11, 382.

Horn, L. R. (2009). Hypernegation, hyponegation, and parole violations. *Annual Meeting of the Berkeley Linguistics Society*, 35(1), 403.

Horn, L. R. (2010). Multiple negation in english and other languages. *The Expression of Negation*, (pp. 111–148).

Hornstein, N. (1995). *Logical form: From GB to minimalism*. John Wiley & Sons.

Hornstein, N. (1999). Minimalism and quantifier raising. In *Working Minimalism*, volume 32 of *Current Studies In Linguistics*. The MIT Press.

Hotopf, W. H. N. (1980). Semantic similarity as a factor in whole-word slips of the tongue. *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand*, (pp. 97–109). Publisher: Academic Press New York.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A Systematic Assessment of Syntactic Generalization in Neural Language Models. *arXiv:2005.03692 [cs]*. arXiv: 2005.03692.

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, 102, 155–181.

Jumelet, J., Denic, M., Szymanik, J., Hupkes, D., & Steinert-Threlkeld, S. (2021). Language Models Use Monotonicity to Assess NPI Licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 4958–4969). Online: Association for Computational Linguistics.

Jumelet, J. & Hupkes, D. (2018). Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 222–231). Brussels, Belgium: Association for Computational Linguistics.

Just, M. A. & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 10(3), 244–253.

Just, M. A. & Clark, H. H. (1973). Drawing inferences from the presuppositions and implications of affirmative and negative sentences. *Journal of Verbal Learning and Verbal Behavior*, 12(1), 21–31. Publisher: Elsevier.

Kapatsinski, V. (2010). Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech*, 53(1), 71–105.

Karimi, H. & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly Journal of Experimental Psychology*, 69(5), 1013–1040.

Kassner, N. & Schütze, H. (2019). Negated LAMA: birds cannot fly. *CoRR*, abs/1911.03343.

Kauf, C. & Ivanova, A. (2023). A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 925–935). Toronto, Canada: Association for Computational Linguistics.

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2022). Event knowledge in large language models: the gap between the impossible and the unlikely. *arXiv preprint arXiv:2212.01488*.

Kay, P. & Charles J., F. (1999). Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. *Language*, 75(1), 1–33.

Kelley, P. (2018). *More People Understand Eschers than the Linguist Does: The Causes and Effects of Grammatical Illusions*. Ph.D. ISBN: 9780355930733.

Kim, N. & Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. *CoRR*, abs/2010.05465.

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture naming errors. *Cognitive neuropsychology*, 25(4), 463–492.

Kizach, J., Christensen, K. R., & Weed, E. (2016). A Verbal Illusion: Now in Three Languages. *Journal of Psycholinguistic Research*, 45(3), 753–768.

Kotek, H., Sudo, Y., & Hackl, M. (2015). Experimental investigations of ambiguity: the case of most. *Natural Language Semantics*, 23, 119–156.

Ladusaw, W. A. (1979). *Polarity Sensitivity as Inherent Scope Relations.* The University of Texas at Austin.

Langsford, S., Stephens, R. G., Dunn, J. C., & Lewis, R. L. (2019). In Search of the Factors Behind Naive Sentence Judgments: A State Trace Analysis of Grammaticality and Acceptability Ratings. *Frontiers in Psychology*, 10.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5), 1202–1241.

Lee, S.-H. & Schuster, S. (2022). Can language models capture syntactic associations without surface cues? a case study of reflexive anaphor licensing in English control constructions. In *Proceedings of the Society for Computation in Linguistics 2022* (pp. 206–211). online: Association for Computational Linguistics.

Leivada, E. & Westergaard, M. (2020). Acceptable ungrammatical sentences, unacceptable grammatical sentences, and the role of the cognitive parser. *Frontiers in Psychology*, (pp. 364). Publisher: Frontiers.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Emmeans: estimated marginal means, aka least-squares means (Version 1.3. 4). *Emmeans Estim. Marg. Means Aka Least-Sq. Means https://CRAN. R-project. org/package= emmeans*.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08* (pp. 234). Honolulu, Hawaii: Association for Computational Linguistics.

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 1055–1065). Portland, Oregon.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, 106(50), 21086–21090. Publisher: Proceedings of the National Academy of Sciences.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9), 1989–2001. Publisher: Elsevier.

Lewis, R. L. & Vasishth, S. (2005). An Activation-Based Model of Sentence Processing as Skilled Memory Retrieval. *Cognitive Science*, 29(3), 375–419. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog0000_25.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*, 10(10), 447–454. Publisher: Elsevier.

Lieder, F. & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, e1.

Linebarger, M. C. (1987). Negative polarity and grammatical representation. *Linguistics and philosophy*, (pp. 325–387).

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs].

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2020). Structural frequency effects in noisy-channel comprehension. In *Presentation at the Penn Linguistics Conference*.

Lohse, B., Hawkins, J. A., & Wasow, T. (2004). Domain minimization in english verb-particle constructions. *Language*, (pp. 238–261).

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, 4.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Mankiw, N. G. (2018). *Principles of Microeconomics*. Cengage Learning, 8th edition.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Marty, P., Chemla, E., & Sprouse, J. (2020). The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A Journal of General Linguistics*, 5(1), 72.

Marvin, R. & Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. In *EMNLP 2018*. arXiv: 1808.09031.

May, R. (1985). *Logical form: Its structure and derivation*, volume 12. MIT press.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Meier, C. (2003). The meaning of too, enough, and so...that. *Natural Language Semantics*, 11, 69–107.

Mendia, J. A., Poole, E., & Dillon, B. (2018). Spurious NPI licensing and exhaustification. *Semantics and Linguistic Theory*, (pp. 233–250).

Mikhailov, V., Shamardina, T., Ryabinin, M., Pestova, A., Smurov, I., & Artemova, E. (2022). RuCoLA: Russian corpus of linguistic acceptability. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5207–5227). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Montalbetti, M. M. (1984). *After binding: On the interpretation of pronouns*. PhD Thesis, Massachusetts Institute of Technology.

Muller, H., de Dios-Flores, I., & Phillips, C. (2019). Not (just) any licensors cause negative polarity illusions.

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Nair, S., Srinivasan, M., & Meylan, S. (2020). Contextualized Word Embeddings Encode Aspects of Human-Like Word Sense Knowledge. *arXiv:2010.13057 [cs]*. arXiv: 2010.13057.

Nalborczyk, L., Batailler, C., Lœvenbruck, H., Vilain, A., & Bürkner, P.-C. (2019). An introduction to Bayesian multilevel models using brms: A case study of gender effects on vowel variability in Standard Indonesian. *Journal of Speech, Language, and Hearing Research*, 62(5), 1225–1242.

Natsopoulos, D. (1985). A verbal illusion in two languages. *Journal of Psycholinguistic Research*, 14(4), 13.

Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in psychology*, 7, 280. Publisher: Frontiers.

Nicenboim, B., Schad, D., & Vasishth, S. (2021). An introduction to bayesian data analysis for cognitive science. *Under contract with Chapman and Hall/CRC statistics in the social and behavioral sciences series*.

O'Connor, E. (2015). *Comparative illusions at the syntax-semantics interface*. PhD thesis, University of Southern California.

O'Connor, E. (2017). The (accidental) ambiguity of inversion illusions. *Proceedings of NELS*, (47), 329–342.

O'Connor, E., Pancheva, R., & Kaiser, E. (2013). Evidence for online repair of Escher sentences. *Proceedings of Sinn und Bedeutung*, 17, 363–380.

Orth, W., Yoshida, M., & Sloggett, S. (2019). NPI illusions are a problem of quantification. In *The 32nd CUNY Conference on Human Sentence Processing* University of Colorado Boulder.

Orth, W., Yoshida, M., & Sloggett, S. (2020). Polarity illusions are quantifier illusions. volume 5 (pp. 6–00).

Orth, W., Yoshida, M., & Sloggett, S. (2021). Negative polarity item (NPI) illusion is a quantification phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(6), 906–947. Publisher: American Psychological Association.

Paape, D. (2020). Too strong to eliminate or negate is surely no illusion.

Paape, D. (2023). When Transformer models are more compositional than humans: The case of the depth charge illusion. *Experiments in Linguistic Meaning*, 2(0), 202–218. Number: 0.

Paape, D. (2024). How do linguistic illusions arise? Rational inference and good-enough processing as competing latent processes within individuals.

Paape, D., Vasishth, S., & Von Der Malsburg, T. (2020). Quadruplex negatio invertit? The on-line processing of depth charge sentences. *Journal of Semantics*, 37(4), 509–555.

Pannitto, L. & Herbelot, A. (2020). Recurrent babbling: Evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning* (pp. 165–176). Online: Association for Computational Linguistics.

Parker, D. & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, 157, 321–339.

Partee, B. (1988). Many quantifiers. In *Proceedings of ESCOL*, volume 5 (pp. 383–402).

Partee, B. (1989). Many quantifiers. In J. Powers & K. de Jong (Eds.), *ESCOL 89: Proceedings of the Eastern States Conference on Linguistics* Columbus, OH: Department of Linguistics, Ohio State University.

Penka, D. (2015). Negation and polarity. *The Routledge handbook of semantics*, (pp. 303–319).

Pérez, E., Santiago, J., Palma, A., & O'Seaghdha, P. G. (2007). Perceptual bias in speech error data collection: Insights from spanish speech errors. *Journal of psycholinguistic research*, 36, 207–235.

Pezzelle, S. (2018). Learning the Meaning of Quantifiers from Language and Vision.

Pezzelle, S., Bernardi, R., & Piazza, M. (2018). Probing the mental representation of quantifiers. *Cognition*, 181, 117–126.

Pham, L. (2022). *Acceptability of Comparative Illusions as a function of interactions between Repeatability of a Verb Phrase and Active production task*. PhD thesis, University of Amsterdam.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. *Experiments at the Interfaces*, 37, 147–180.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language*, 24(5), 554–585.

Pijpops, D., De Smet, I., & Van de Velde, F. (2018). Constructional contamination in morphology and syntax: Four case studies. *Constructions and Frames*, 10(2), 269–305.

Pijpops, D. & Van de Velde, F. (2016). Constructional contamination: How does it work and how do we measure it? *Folia Linguistica*, 50(2), 543–581.

Poliak, M., Ryskin, R., Braginsky, M., & Gibson, E. (2023). It is not what you say but how you say it: Evidence from Russian shows robust effects of the structural prior on noisy channel inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Publisher: American Psychological Association.

Poppels, T. & Levy, R. P. (2016). Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, Texas.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

Ryskin, R., Bergen, L., & Gibson, E. (2021a). Agreement errors are predicted by rational inference in sentence processing. Publisher: PsyArXiv.

Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition*, 181, 141–150.

Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021b). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia*, 158, 107855.

Salazar, J., Liang, D., Nguyen, T. Q., & Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2699–2712). Online: Association for Computational Linguistics.

Sanford, A. J. & Emmott, C. (2012). *Mind, Brain and Narrative*. Cambridge University Press. Google-Books-ID: iAYhAwAAQBAJ.

Sanford, A. J. & Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences*, 6(9), 382–386.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. P. (2022). Large-Scale Evidence for Logarithmic Effects of Word Predictability on Reading Time.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3), 379–423. Publisher: Nokia Bell Labs.

She, J. S., Potts, C., Bowman, S. R., & Geiger, A. (2023). Scone: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. *arXiv preprint arXiv:2305.19426*.

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, 15(2), 143–157.

Shin, U., Yi, E., & Song, S. (2023). Investigating a neural language model's replicability of psycholinguistic experiments: A case study of NPI licensing. *Frontiers in Psychology*, 14, 937656.

Simon, H. (1972). Theories of bounded rationality. In C. McGuire & R. Radner (Eds.), *Decision and Organization* (pp. 161–176). North-Holland Publishing Company.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69(1), 99–118. Publisher: Oxford University Press.

Simon, H. A. (1964). Theories of bounded rationality. *CIP Working Paper*, 66, 1–23.

Solt, S. (2011). Vagueness in Quantity: Two Case Studies from a Linguistic Perspective. In P. Cintula, C. G. Fermueller, L. Godo, & P. Hajek (Eds.), *Understanding Vagueness. Logical, Philosophical and Linguistic Perspectives* (pp. 157–174). London: College Publications.

Solt, S. (2016). On measurement and quantification: The case of" most" and" more than half". *Language*, (pp. 65–100).

Someya, T. & Oseki, Y. (2023). JBLiMP: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 1581–1594). Dubrovnik, Croatia: Association for Computational Linguistics.

Song, Y., Krishna, K., Bhatt, R., & Iyyer, M. (2022). SLING: Sino linguistic evaluation of large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4606–4634). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Stemberger, J. P. (1984). Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology*, 1(4), 281–313. Publisher: Routledge _eprint: https://doi.org/10.1080/02643298408252855.

Stemberger, J. P. (1992). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In *Experimental slips and human error: Exploring the architecture of volition* (pp. 195–215). Springer.

Syrett, K. (2015). Qr out of a tensed clause: Evidence from antecedent-contained deletion. *Ratio*, 28(4), 395–421.

Szymanik, J. (2016). *Quantifiers and Cognition: Logical and Computational Perspectives*, volume 96 of *Studies in Linguistics and Philosophy*. Cham: Springer International Publishing.

Szymanik, J. & Zajenkowski, M. (2010). Comprehension of Simple Quantifiers: Empirical Evaluation of a Computational Model. *Cognitive Science*, 34(3), 521–532. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-6709.2009.01078.x.

Tonhauser, J. & Matthewson, L. (2015). Empirical evidence in research on meaning.

Townsend, D. J. & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. MIT Press.

Traxler, M. J. (2014). Trends in syntactic parsing: anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences*, 18(11), 605–611.

Truong, T. H., Baldwin, T., Verspoor, K., & Cohn, T. (2023). Language models are not naysayers: An analysis of language models on negation benchmarks. *arXiv preprint arXiv:2306.08189*.

Turner, M. L. & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, 28(2), 127–154. Publisher: Elsevier.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases, biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157), 1124–1131.

van Schijndel, M. & Linzen, T. (2021). Single-Stage Prediction Models Do Not Explain the Magnitude of Syntactic Disambiguation Difficulty. *Cognitive Science*, 45(6), e12988.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, 32(4), 685–712. Publisher: Wiley Online Library.

Vasishth, S., Drenhaus, H., Saddy, J., Potsdam, R., Lewis, & Michigan (2005). Processing negative polarity. *Talk Presented at the 18th CUNY Sentence Processing Conference*.

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4), 533–567. Publisher: Taylor & Francis.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.

Von Fintel, K. (1999). Npi licensing, strawson entailment, and context dependency. *Journal of semantics*, 16(2), 97–148.

Wang, L., Hagoort, P., & Yang, Y. (2009). Semantic illusion depends on information structure: ERP evidence. *Brain Research*, 1282, 50–56.

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., Alsop, A., Bordia, S., Liu, H., Parrish, A., Wang, S.-F., Phang, J., Mohananey, A., Htut, P. M., Jeretič, P., & Bowman, S. R. (2019). Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. arXiv:1909.02597 [cs].

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377–392.

Wason, P. C. & Reich, S. S. (1979). A Verbal Illusion. *Quarterly Journal of Experimental Psychology*, 31(4), 591–597.

Wei, J., Garrette, D., Linzen, T., & Pavlick, E. (2021). Frequency Effects on Syntactic Rule Learning in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 932–948). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Weissweiler, L., He, T., Otani, N., Mortensen, D. R., Levin, L., & Schütze, H. (2023). Construction grammar provides unique insight into neural language models. *arXiv preprint arXiv:2302.02178*.

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The Anatomy of a Comparative Illusion. *Journal of Semantics*, 35(3), 543–583.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of BlackboxNLP 2018*. arXiv: 1809.00042.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Online: Association for Computational Linguistics.

Wright, S. (1954). The death of lady mondegreen. *Harper's Magazine*, 209(1254), 48–51.

Xiang, M., Dillon, B., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain and Language*, 108(1), 40–55.

Xu, F. & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, 114(2), 245. Publisher: American Psychological Association.

Zajenkowski, M. & Szymanik, J. (2013). Most intelligent people are accurate and some fast people are intelligent.: Intelligence, working memory, and semantic processing of quantifiers from a computational perspective. *Intelligence*, 41(5), 456–466.

Zajenkowski, M., Szymanik, J., & Garraffa, M. (2014). Working memory mechanism in proportional quantifier verification. *Journal of psycholinguistic research*, 43, 839–853.

Zhan, M., Chen, S., Levy, R., Lu, J., & Gibson, E. (2023). Rational Sentence Interpretation in Mandarin Chinese. *Cognitive Science*, 47(12), e13383.

Zhang, Y., Chen, W., Zhang, R., & Zhang, X. (2022). Representing affect information in word embeddings. *arXiv preprint arXiv:2209.10583*.

Zhang, Y., Gibson, E., & Davis, F. (2023a). Can Language Models Be Tricked by Language Illusions? Easier with Syntax, Harder with Semantics. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)* (pp. 1–14). arXiv:2311.01386 [cs].

Zhang, Y., Kauf, C., & Gibson, E. (2023b). A noisy-channel explanation of the comparative illusion. *Architectures and Mechanisms for Language Processing*.

Zhang, Y., Ryskin, R., & Gibson, E. (2023c). A noisy-channel approach to depth-charge illusions. *Cognition*, 232, 105346.