Cognition 232 (2023) 105346

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A noisy-channel approach to depth-charge illusions

Yuhan Zhang [a,*], Rachel Ryskin [b], Edward Gibson [c]

[a] Department of Linguistics, Harvard University, Boylston Hall, Cambridge, MA 02138, USA
[b] Department of Cognitive & Information Sciences, University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA
[c] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar St., Cambridge, MA, 02139, USA

ARTICLE INFO

Keywords:
Depth-charge sentence
Semantic illusion
Noisy-channel framework
Language comprehension

ABSTRACT

The "depth-charge" sentence, *No head injury is too trivial to be ignored,* is often interpreted as "no matter how trivial head injuries are, we should not ignore them" while the literal meaning is the opposite – "we should ignore them". Four decades of research have failed to resolve the source of this entrenched semantic illusion. Here we adopt the noisy-channel framework for language comprehension to provide a potential explanation. We hypothesize that depth-charge sentences result from inferences whereby comprehenders derive the interpretation by weighing the plausibility of possible readings of the depth-charge sentences against the likelihood of plausible sentences being produced with errors. In four experiments, we find that (1) the more plausible the intended meaning of the depth-charge sentence is, the more likely the sentence is to be misinterpreted; and (2) the higher the likelihood of our hypothesized noise operations, the more likely depth-charge sentences are to be misinterpreted. These results suggest that misinterpretation is affected by both world knowledge and the distance between the depth-charge sentence and a plausible alternative, which is consistent with the noisy-channel framework.

1. Introduction

On the surface, sentence (1) seems like a reasonable thing to say:

(1) 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 (1) actually implies the opposite: "all head injuries should be ignored however trivial". The clash between 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. 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 – they explode after traveling in the water for a certain amount of time (e.g., Sanford & Emmott, 2012: 28).

The depth-charge interpretation was observed more than four decades ago (Wason & Reich, 1979) and the source of the illusion 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). These semantic illusions are easier to detect when the illusory words appear in the focus position of the sentence (Wang, Hagoort, & Yang, 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 the final interpretation, even after explicit instruction of how to interpret structures like *X is too Y to Z* (Giannouli, 2016; Kizach, Christensen, & Weed, 2016; Natsopoulos, 1985; O'Connor, 2015, 2017; Paape, Vasishth, & von der Malsburg, 2020).

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 and Thomas (1999), Vasishth, Suckow, Lewis, and Kern (2010) and Futrell, Gibson, and Levy (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 (2) (example from Frazier (1985),

* Corresponding author.
E-mail addresses: yuz551@g.harvard.edu (Y. Zhang), rryskin@ucmerced.edu (R. Ryskin), egibson@mit.edu (E. Gibson).

https://doi.org/10.1016/j.cognition.2022.105346
Received 1 June 2022; Received in revised form 24 October 2022; Accepted 28 November 2022
Available online 10 December 2022

reporting an intuition from Janet Fodor):

(2)
a missing verb phrase:
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 (2a) more acceptable than (2b) (Gibson & Thomas, 1999) in spite of the fact that (2a) is actually ungrammatical: (2a) 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 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 (2a) as a grammatical variant, as in (3) (see Hahn, Futrell, Levy, & Gibson, 2022, for an even further updated approach):

(3) 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 (1) 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 (e.g., Gibson, Bergen, & Piantadosi, 2013; 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.

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 (e.g., 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[1] 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, 1997).

Lastly, Paape et al. (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.

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 (4), *enough to* in (4a) and *so…that* in (4b) 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 (4c) presupposes that the degree of the adjective surpasses the

[1] In this paper, we use "the polarity of the final verb" to refer to whether the verb is positive or negative. Negative verbs can (1) have a negative prefix (e.g., *misinterpret*, *uninhabit*, *discontinue*), (2) mean the opposition to and the inhibition of some action (e.g., *ignore* means the inhibition of *treat*) and (3) from the sentiment analysis perspective, trigger negative unfavorable emotion toward the topic at issue (e.g., *reject*, *overlook*, *abandon*, *cancel*) (e.g., Mohammad & Turney, 2013; Turney & Littman, 2003). Note that the sense of 'polarity' here is not the same as that in negative polarity items (e.g., Ladusaw, 1980).

baseline above which the action would not be allowable or possible. Therefore, sentence (4c) 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*.

(4)
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, Hollingworth, Halliwell, & Ferreira, 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 offers a valuable insight for our approach of adopting the noisy-channel theory to explain the depth-charge illusion which we will elaborate on in section 1.5.

1.3. Negation

Wason and Reich (1979) suggested that the final negative verb might pose special difficulty 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, Logačev, Gattei, & Vasishth, 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, Fraundorf, Lee, & Watson, 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.

1.4. A construction-based non-illusory account

Some researchers working within a construction grammar framework (e.g., Goldberg, 1995; Kay & Fillmore, 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 (5a), *too…to* retains its canonical reading as in (4c) but in (5b), *too…to* means that the property of the subject exceeds some degree and the action denoted by the verb is realized as a consequence. Sentence (5b) 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*.

(5)
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., Dell & Reich, 1981; Gross, 1983). 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 (e.g., Pijpops, De Smet, & Van de Velde, 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.

1.5. Present work: a noisy-channel explanation of depth-charge sentences

In the current study, we adopt the information-theoretic noisy-channel framework (Shannon, 1948) to explore the depth-charge illusion. The noisy-channel framework views language comprehension as a process of rational Bayesian inference given uncertain input (Gibson et al., 2013; Levy, 2008; Levy, 2011; Ryskin et al., 2021; Ryskin, Futrell, Kiran, & Gibson, 2018).[2] Rational comprehenders reach an interpretation of the perceived utterances by weighing probabilities of their potential interpretations against the likelihood that the intended message is corrupted to the perceived signal by noise during information

[2] Two related frameworks are the good-enough processing approach (Christianson et al., 2001; Ferreira, 2003; Ferreira, Bailey, & Ferraro, 2002; Karimi & Ferreira, 2016) and the "shallow processing" approach (Sanford & Sturt, 2002) (see Traxler (2014) for a comprehensive comparison).

transmission. The schematic for this rational process is shown in Fig. 1, following Shannon (1948). The speaker intends to convey a meaning m_i by encoding it linguistically in the intended sentence s_i. The sentence is conveyed via a noisy channel and can be corrupted by noise due to producer or comprehender errors, or noise in the environment so that s_i and s_p may differ. For example, the speaker might have a slip of the tongue, changing the surface structure of s_i; the conversation could be occurring in a noisy room, or the listener could be inattentive, causing some words to be misheard, etc. The comprehender perceives the linguistic signal as s_p and recovers the intended meaning as m_p.

Successful communication takes place when m_p is the same as m_i but this is not always the case given the presence of noise. Rather, comprehenders infer the probability of a given s_i through Bayesian reasoning, which can be formalized by considering an ideal observer (Geisler, 1989; Geisler & Diehl, 2003) model of language comprehension:

$$P(s_i|s_p) \propto P(s_i)\, P(s_p|s_i) \quad (1)$$

In Eq. (1), s_p is the sentence perceived by the comprehender and s_i is one of the hypothesized sentences intended by the speaker to convey the message. This model is a simplification of the schematic in Fig. 1 in that we take s_i to represent both the linguistic strings and the intended meaning. The left-hand side of Eq. (1), $P(s_i|s_p)$, is the posterior probability assigned by the comprehender to the intended sentence s_i given the perceived input s_p. According to Bayes' rule, this is proportional to, on the right side of the equation, the prior probability $P(s_i)$ that the producer intends to communicate s_i times the likelihood, $P(s_p|s_i)$, which reflects the likelihood that the comprehender assigns to a noise corruption that would transform s_i into s_p during information transmission.

In daily communication, the prior $P(s_i)$ represents all the comprehender's existing linguistic and world knowledge about the intended meaning. In other words, it reflects the comprehender's estimates of the probability of the meaning and structure of the received sentence and the probability of the meaning and structure of all possible alternatives, which indicates that this term biases interpretation of the perceived input toward plausible meanings and forms. The noise model $P(s_p|s_i)$ represents the comprehender's estimates of how likely a given sentence is to be corrupted during transmission and received by the comprehender in a different form. In the simplified Eq. (1), the noise model doesn't distinguish between corruptions coming from the speaker and those coming from the hearer (or the environment, as in a loud room) though these are clearly important distinctions that the comprehender is likely attuned to. The noise model is also sensitive to specific linguistic structures in s_i and appears to assign higher probability to word deletions than insertions (Gibson et al., 2013; Poppels & Levy, 2016). For example, in Gibson et al. (2013), when comprehenders perceive the sentence "The mother gave the candle the daughter," which has low prior probability, they may entertain a near neighbor of the perceived sentence, "The mother gave the candle to the daughter", which is different only by a one-word deletion and has a much higher prior probability.[3] Since the likelihood of deletion is relatively high, the posterior probability of reaching an inferred non-literal meaning outweighs that of the literal interpretation. This prediction is supported by responses to comprehension questions in Gibson et al. (2013) which show that participants are likely to interpret the perceived sentence according to the meaning of its more plausible alternative.

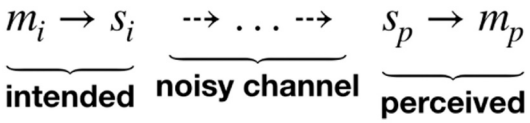

Fig. 1. Schematic representation of the noisy channel model.

Within this framework, we 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 $P(s_i|s_p)$.

Based on Eq. (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 and 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.[4] 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. Eq. (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 Sections 1.1–1.4) 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

[3] Other corruptions are possible as well. For instance, the noise here could be a switch between "candle" and "daughter" rather than a deletion of "to". In Ryskin et al. (2018), readers correcting implausible sentences assumed both exchanges and deletions for these kinds of sentences. Just looking at the inferences associated with this implausible double object (DO) structure is not enough to decide among many noise models. The reason why Gibson et al. (2013) concluded that a deletion was a likely source for the high inference rate in the implausible DO structures was from the inference rates associated with 10 different implausible materials (please see Table 2 for examples). In all of these, there is an exchange of two nouns, to result in a plausible meaning. Yet the inference rates were as low as ~5% (very low) for what Gibson et al. termed their active / passive materials and as high as ~70% for the implausible DO-benefactive materials, with a lot of inference rates in between (please see Figure 2 in Gibson et al. (2013)). Based on this pattern of data, Gibson et al. (2013) argued that deletions were a likely part of a noise model and were likely in the implausible DO and DO-benefactives (and in another structure).

[4] We also attempted another noise type involving negation. Please see Appendix B.

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 $P(s_p|s_i)$.

1.6. 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. Experiment 1

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.

2.1. Methods

2.1.1. 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.

2.1.2. Materials & procedure

The materials were translated to English from Paape et al.'s (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 (6). 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 (6) 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.

(6)

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 were plausible, whereas those in the other conditions were not. In the quantifier-*some*, negative-adjective condition, 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 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 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 commonsense 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 toward 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, Piantadosi, & Fedorenko, 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).

2.2. Results

The distributions of plausibility ratings by condition are presented in Fig. 2. The sentences in the quantifier-*some*, positive-adjective condition (6a) were largely rated as plausible (mean = 5.70, confidence interval (CI) = [5.54,5.84]). Sentences from the quantifier-*some*, negative-adjective condition (6b), were rated less plausible (mean = 3.17, CI =

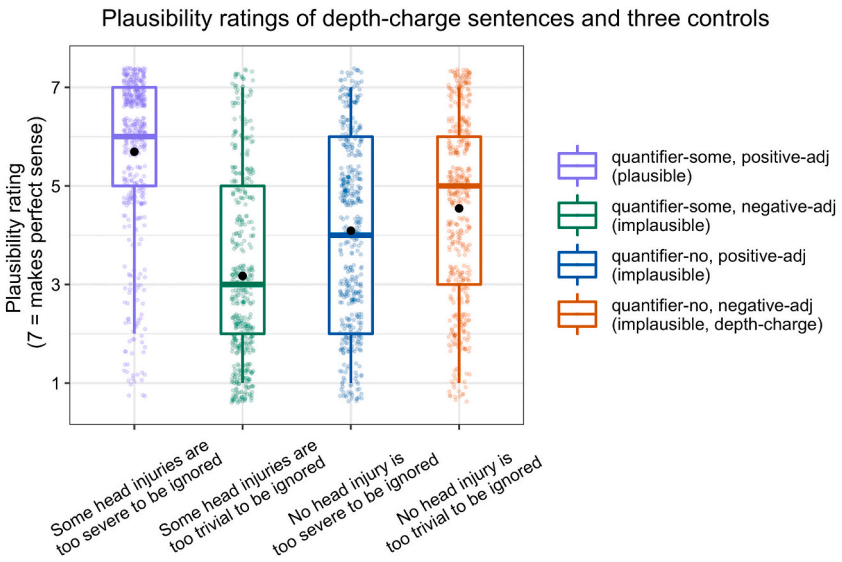

Fig. 2. Plausibility ratings for sentences by condition in Exp.1. The depth-charge sentences were rated as more plausible than the other implausible conditions. (The middle horizontal line in the boxplot represents the median; the black dot represents the mean; the jittered points represent single trial plausibility ratings.)

[3.01, 3.35]). Sentences from the quantifier-*no*, positive-adjective condition (6c), (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; (6d)) 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 (Core Team, 2018), following the same analytical strategy in Paape et al. (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 and items were included as random effects to obtain the maximal random effect structure for mixed effects models (Barr, Levy, Scheepers, & Tily, 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, Kurowicka, & Joe, 2009; Nalborczyk, Batailler, Lœvenbruck, Vilain, & Bürkner, 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 $\widehat{R}$ close to 1.0 marks the convergence of the sampling chain to the underlying posterior distribution of the target predictor (Gelman & Rubin, 1992). All $\widehat{R}$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. 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 (6c) 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.

3. Experiment 2

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 relative 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.

3.1. Methods

3.1.1. 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.

3.1.2. Materials & procedure

We created a sentence to represent the plausible intended meaning of each of the 32 depth-charge items in Experiment 1.[5] 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 (6d) is in (7):

(7) 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 associated 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 (RR and EG) 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.

3.2. Results

Figure 3 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 4 plots the plausibility ratings from Experiment 1 for the 32 critical items in 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, Singmann, Love, Buerkner, & Herve, 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 ($\beta = 0.27$, 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: $\beta = -0.14$, HPD = [−0.50, 0.22]; quantifier-*some* & negative-adjective: $\beta = 0.20$, HPD = [−0.04, 0.46]; quantifier-*no* & positive-adjective: $\beta = -0.07$, HPD = [−0.28, 0.18]). Please see the supplemental material for the full output from the Bayesian model.

3.3. 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.

4. Experiment 3

4.1. 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 operations – structural substitution and antonym substitution – that might lead to the generation of depth-charge sentences.[6] 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 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 (8) 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 (9) where the intended plausible sentence has the plausible *too trivial to be treated* but is produced as the implausible *too trivial to be ignored*.

(8) 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)

(9) 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

[5] In a preliminary version of Experiment 2 (see Appendix A) we used direct translations from Paape et al. (2020)'s materials with minor modifications for the plausible versions of depth charge items. For example, for the depth-charge item *No head injury is too trivial to be ignored*, we asked participants to rate whether they agree with "Head injuries are in general too severe to be ignored". We later noticed that this format of paraphrase provided awkward translations for many items and thus conducted this second version (presented in the main text) with more natural sounding materials. The results were consistent across both versions of the experiment.

[6] Because depth-charge sentences always have negative meanings, we initially hypothesized that the intended plausible alternatives might be (1a) or (1b), such that the depth-charge versions might be produced via the deletion or insertion of *not*. Following the noisy-channel hypothesis, perceived sentences resulting from deletions, as in (1a), should elicit more inference than those resulting from insertions, as in (1b). However, the results of a comprehension experiment were not consistent with this prediction (see Appendix B). One possible explanation is that people might be very unlikely to intend to produce multiple negations (e.g., as in the target utterance in (1a)) because they are extremely complex and hard to understand (Horn, 2009; Horn, 2010).(1) a. No head injury is too trivial to not be ignored.(plausible, the deletion of *not* results in the canonical depth-charge sentence)b. No head injury is too trivial to (# not) be treated.(plausible, the insertion of *not* results in a depth-charge construction)

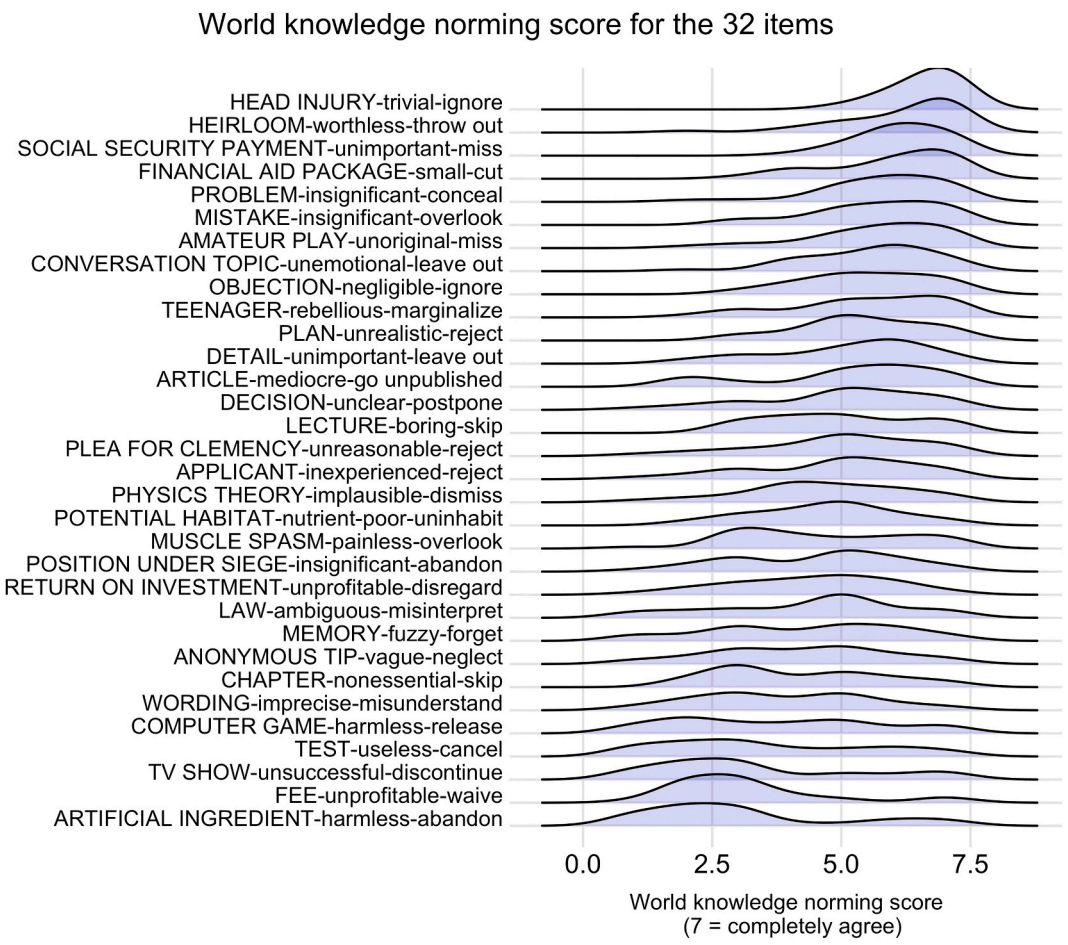

Fig. 3. The distribution of world knowledge norming score for the 32 items in Exp. 2.

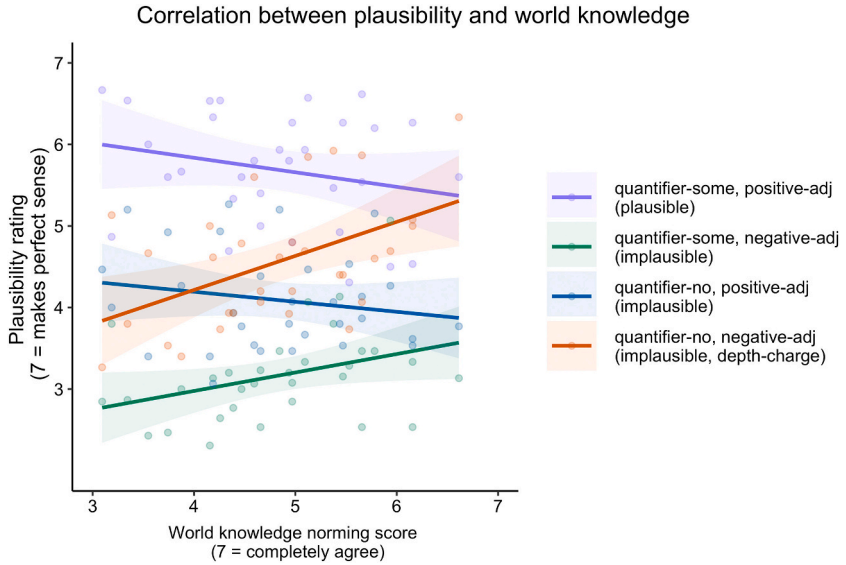

Fig. 4. The correlation between item-wise world knowledge norming score for the 32 items and the plausibility rating across the four conditions from Exp.1 (Lines represent a linear regression line; Shades represent 95% bootstrapped CI; The *x* axis was set from 1 to 7 but the result started from above 3; The *y* axis was set from 1 to 7 but the result started from above 2).

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, Dell, Verkuilen, & Schwartz, 2008, Stemberger, 1984, a.o.).

Antonym substitution, as in (9), 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 and 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 (8a) to be produced as (8b) compared with (10a) being produced as (10b). For the antonym substitution error, we predict it would be more likely for (9a) to be produced as (9b) compared with (11a) being produced as (11b).[7]

(10)

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)

(11)

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, 1996; Pérez, Santiago, Palma, & O'Seaghdha, 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 task in Experiment 3b by collecting noise ratings for all the noise operations posited in Gibson et al. (2013).

[7] Note we do not specify at which level of speech production (e.g., Garrett, 1980; Garrett, 1988) these errors occur in the actual production process but we assume that the errors we are hypothesizing here take place in the lexical selection and concept framing level.

4.1.1. 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.

4.1.2. 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 substitution) and the degree quantifier construction in the intended sentence (*so…as to* vs. *too…to*) (see Table 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 4.2). Before the task began, participants were asked to complete five English sentences. Responses to these sentence preambles were used to identify bots.

4.1.3. 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 Fig. 5. Overall, the structural substitution appeared to be a more probable noise operation (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 =

Table 1
Noise operation conditions tested in Exp.3a.

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)

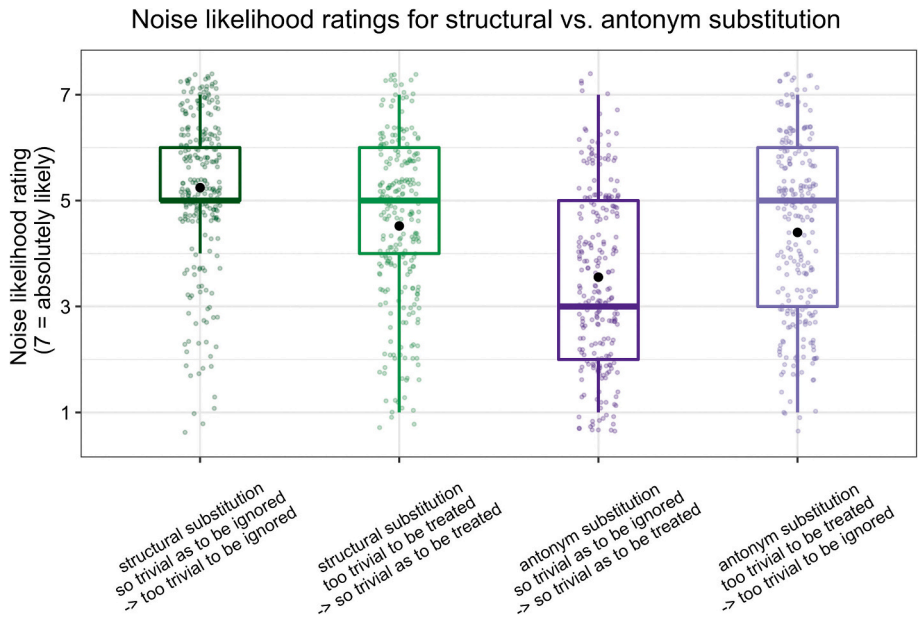

Fig. 5. Noise likelihood ratings for the depth-charge material crossing types of noise corruption and intended structure. (Black points are mean; horizontal lines are median, with dots representing individual trials).

±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 ($\widehat{R}$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*.

4.2. 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 deletions and insertions). Specifically, we used the critical items, across 5 syntactic alternations, from Gibson et al. (2013) (Table 2).

Gibson et al. (2013) proposed two general patterns of edit likelihood to account for the comprehension question responses that they observed (summarized in Fig. 6A). First, 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 – than Intransitives, PO goals, and PO benefactives – the (b)-(d)

Table 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.	c.The girl was kicked by the ball.	2 insertions
	b.The ball was kicked by the girl.	d.The ball ~~was~~ kicked ~~by~~ the girl.	2 deletions
Object-locative/ subject-locative	a.The cat jumped onto a table.	c.Onto the cat jumped a table.	1 insertion 1 deletion
	b.Onto the table jumped a cat.	d. The table jumped onto a cat.	1 deletion 1 insertion
Transitive/ intransitive	a.The tax law benefited the businessman.	c.The tax law benefited from the businessman.	1 insertion
	b.The businessman benefited from the tax law.	d.The businessman benefited ~~from~~ the tax law.	1 deletion
DO/PO goal	a.The mother gave the daughter the candle.	c.The mother gave the daughter to the candle.	1 insertion
	b.The mother gave the candle to the daughter.	d.The mother gave the candle ~~to~~ the daughter.	1 deletion
DO/PO benefactive	a.The cook baked Lucy a cake.	c.The cook baked Lucy for a cake.	1 insertion
	b.The cook baked a cake for Lucy.	d.The cook baked a cake Lucy.	1 deletion

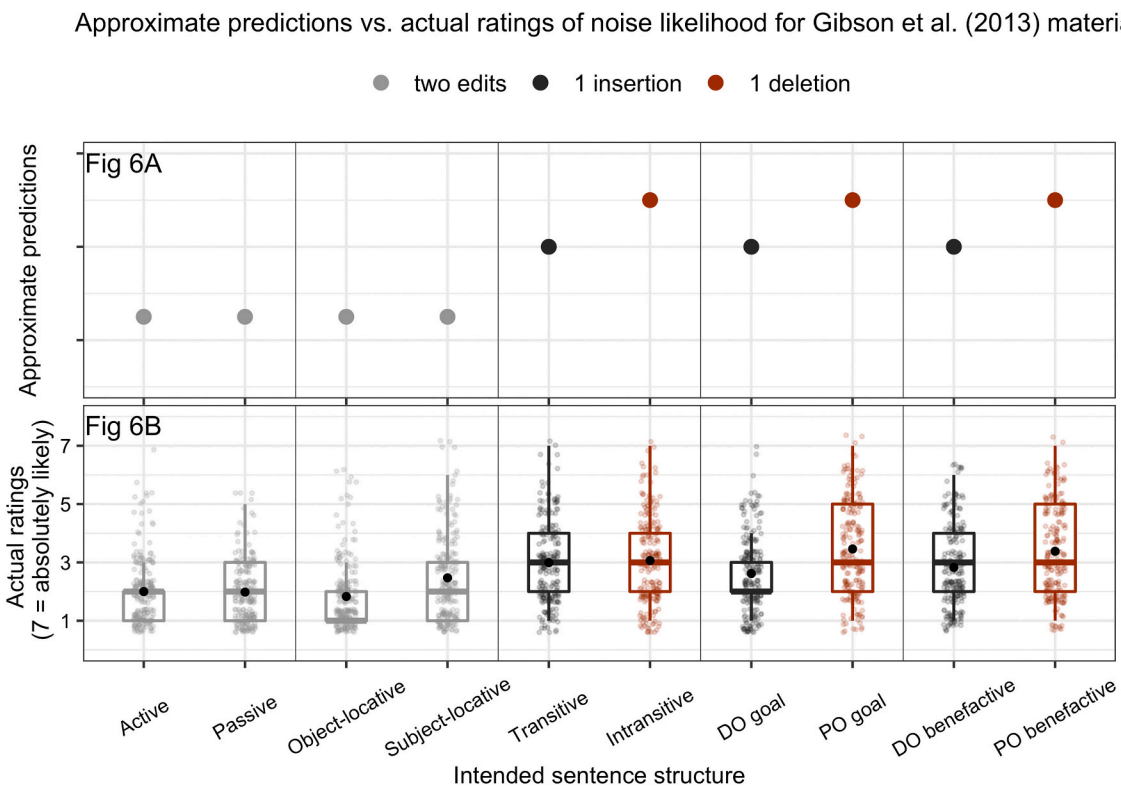

Fig. 6. A. The approximate prediction of people's expectations about the likelihood of a given edit (i.e., how likely it is that the plausible sentences were intended but produced as their implausible counterparts) across the five pairs of syntactic constructions in Gibson et al. (2013). B. Noise likelihood ratings from native English speakers. Each individual point is the raw rating for each trial in a jittered presentation. (The black dot indicates the mean of the distribution, and the boxplot indicates the quartiles of the distribution.)

pairs.

4.2.1. Participants

Participants were the same as in Experiment 3a.

4.2.2. Materials & procedure

The procedure was identical to Experiment 3a. From the materials in Gibson et al. (2013), we selected a random set of 5 pairs of items from each of the 10 conditions in Table 2. Each participant read the 50 pairs of sentences in a random order.

4.2.3. Results

Figure 6B 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 (see details in the supplemental material). All chains converged successfully ($\widehat{R}s = 1.0$).

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 ($\beta = -1.05$, 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 ($\beta = 0.51$, CrI $= [0.21, 0.81]$).

4.3. 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. (2013). The noise likelihood rating patterns are consistent with the findings in Gibson et al. (2013) 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.[8] The alignment between the experimentally rated error rates and the corresponding inference rates in Gibson et al. (2013) 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. (2013). Since structural substitutions were rated as more probable than antonym substitutions, we further investigated the former in Experiment 4 and

[8] The results on the subject-locative and intransitive conditions do not strictly match our prediction. This could be because we only selected 5 items out of 20 that were tested in Gibson et al. (2013) for each construction. It is possible that the selected sentences were not quite representative of the full sets. We did not design this experiment to test individual pairs, so we leave this potential issue for future work.

treated its noise likelihood rating as a proxy for $P(s_p|s_i)$ for the depth-charge materials.

5. Experiment 4

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. (2013). 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*.

5.1. Methods

5.1.1. 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.

5.1.2. Materials & 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 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 implausible 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.

Apart from the 24 items, there were also 40 filler items that all described the generic properties of or the common attitudes toward 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.

5.2. Results

Figure 7 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 sentences 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 included 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[9] (Nicenboim, Schad, & Vasishth, 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 $\widehat{R}$ 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 ($\beta = -2.22$, CrI = $[-3.16, -1.14]$). Within the plausible sentences, there was no difference in the literal interpretation rate between the two conditions ($\beta = 0.35$, 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) ($\beta = 1.36$, HPD = [0.43, 2.25]).

In addition, we evaluated the ideal observer model of language comprehension in Eq. (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 Fig. 8. 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 and probabilities.) Fig. 8 shows a positive correlation between the two measurements, as predicted (Spearman rank correlation: $r = 0.38$, $p = .008$).

5.3. 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.

Table 3
Conditions in Exp. 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.

[9] In brief, the priors of the logistic regression model are chosen such that forward-simulating data from the model using only the priors (i.e., without any data) results in a distribution of predicted outcomes that is consistent with the range of possible outcomes, given domain knowledge.

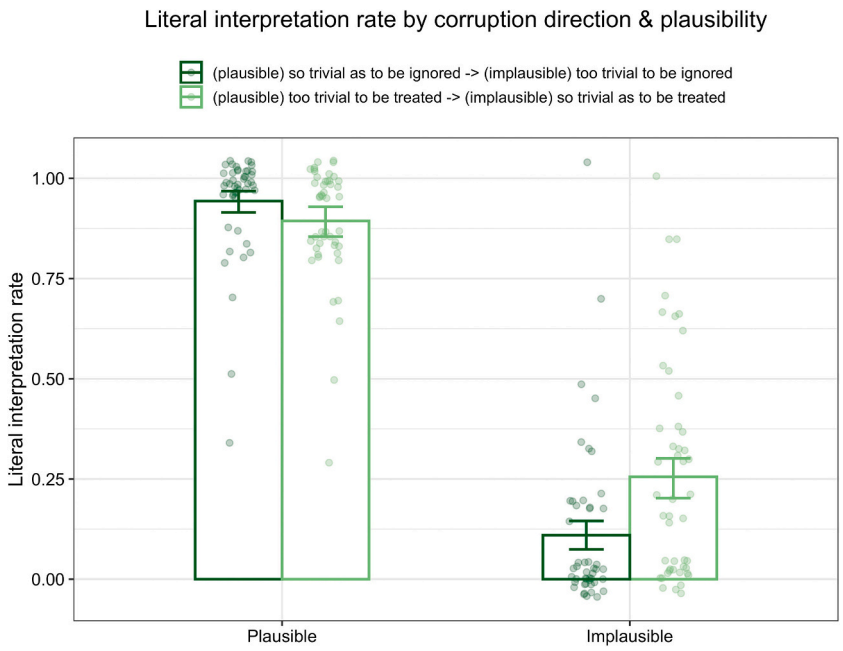

Fig. 7. The literal interpretation rate of plausible and implausible sentences crossing the noise corruption direction and structure (Error bars indicate 95% bootstrapped confidence interval over subject means; the points indicate average literal interpretation rate by subject).

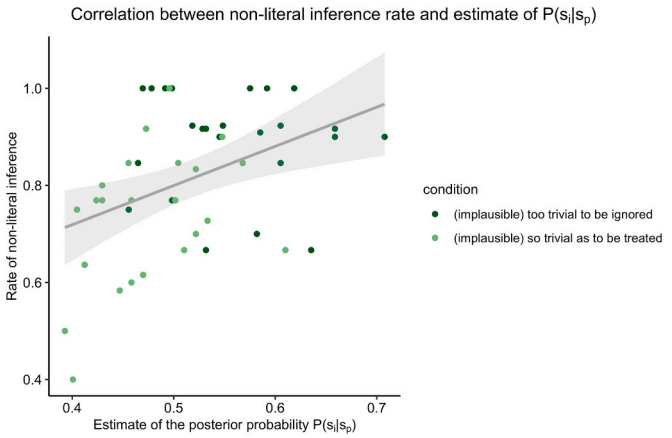

Fig. 8. Correlations between the non-literal inference rates and estimates of the posterior probability $P(s_i|s_p)$ (by multiplying normalized world knowledge rating and normalized noise likelihood constant). The grey line is a linear fit between the non-literal inference and the estimated posterior probability; the error bars represent standard deviation.

6. General discussion

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., 2013; 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 and 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, see details in Section 1.4) 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 Eq. (1) sheds light on why the interpretation of depth-charge

materials is heavily influenced by world knowledge (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., 2013; 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.

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.

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 (see Appendix B). 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.

Finally, the current work does not address how the noisy-channel framework is instantiated at the mechanistic level. So far, there is no concrete consensus 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, Bicknell, Slattery, & Rayner, 2009; Ryskin et al., 2021). 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. In addition, it would be interesting to examine the relationship between the likelihood of making the depth-charge inference and individual differences in exposure to written language or education level (e.g., Acheson, Wells, & MacDonald, 2008; Dąbrowska, 2012), in order to understand how language experience shapes the distributions of both the prior and the noise likelihood.

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., Christianson, Williams, Zacks, & Ferreira, 2006; Ferreira, 2003; Ferreira & Lowder, 2016; 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.

Overall, the noisy-channel approach provides a promising candidate explanation for depth-charge illusions and might explain other linguistic illusions in a similar vein. For instance, the noisy-channel framework may explain comparative illusions, such as *More people have been to Russia than I have,* which is literally incoherent but is accepted at initial processing (Leivada, 2020; O'Connor, Pancheva, & Kaiser, 2013, O'Connor, 2015; Wellwood, Pancheva, Hacquard, & Phillips, 2018). One speculation is that readers infer that a more plausible sentence such as *Many people have been to Russia more than I have* was intended by the speaker but was corrupted by noise (e.g., the substitution of *many* by *more* and the deletion of *more*). Further work is needed to test the full scope of language illusions for which the noisy-channel framework of human communication can provide an explanatory account. Together with more illusions to be explained, the noisy-channel framework could bring more perspectives in understanding properties of language processing.

Author contributions

YZ, RR, and EG designed the research; YZ performed the research; YZ and RR analyzed the data; YZ, RR, and EG wrote the paper. The authors declare no conflict of interest.

CRediT authorship contribution statement

Yuhan Zhang: Conceptualization, Methodology, Formal analysis, Investigation, Visualization, Data curation, Writing - original draft, Writing - review & editing. **Rachel Ryskin:** Conceptualization, Methodology, Formal analysis, Resources, Supervision, Investigation, Validation, Visualization, Writing - original draft, Writing - review & editing. **Edward Gibson:** Conceptualization, Methodology, Resources, Supervision, Project administration, Funding acquisition, Formal analysis, Investigation, Writing - original draft, Writing - review & editing.

Data availability

The experimental data and code scripts can be accessed through the Open Science Foundations repository https://osf.io/nhytx/.

Acknowledgements

We thank the three anonymous reviewers, Roger Levy, Richard Futrell, members at the Language Lab at Massachusetts Institute of Technology, audience at the 2021 AMLaP conference for their helpful comments and suggestions.

Funding

The work was supported by a grant from the National Science Foundation (Award 2121074) to E. Gibson and R. Levy.

Appendix A. World knowledge norming experiment (replication)

This experiment had the same goal as Experiment 2 which was to gather a set of world knowledge norming scores and test whether the plausibility ratings for the depth-charge sentences in Experiment 1 were related to the world-knowledge-based norming score.

A.1. Methods

A.1.1. Participants

40 participants were recruited from Amazon's Mechanical Turk to finish the world knowledge norming study. The participants were paid $2 for their participation. After excluding those who took the survey twice, who did not answer at least 75% of the comprehension questions correctly, and who self-identified as a non-native speaker of English, 32 participants contributed to the final data analysis.

A.1.2. Materials & procedure

The current norming study created 32 sentences each targeting one critical depth-charge sentence in Experiment 1. All the sentences were of the form, "[TOPIC] are in general/on average too [POS ADJ] to be [VERBed]", as in (A1). Half of them had "in general" and the other half "on average". [TOPIC] is the subject noun phrase from each depth-charge sentence (e.g., *head injuries*, *problems*, *mistakes*). [POS ADJ] is the positive adjective used in the control sentences in Experiment 1 (e.g., *severe*, *significant*, *important*). [VERBed] is the sentence-final verb for each depth-charge sentence (e.g., *ignored*, *missed*, *overlooked*). In order to increase the scoring variability across items, no filler items were added. Each participant read a randomized presentation of the 32 sentences and was asked to answer a comprehension question (e.g., *Does "head injuries" appear in this sentence*?). Then they were asked to rate "to what degree they agree or disagree with the sentence" on a 7-point Likert scale (1 = "completely disagree", 4 = "intermediate", 7 = "completely agree").

(A1) Head injuries are in general too severe to be ignored.

A.2. Results

Figure A1 displays the distribution of the world knowledge norming score across the 32 items. The relative ordering of the items is not identical to that in Experiment 2, but a Pearson correlation of the item-wise norming score shows that the two experiments collected highly correlated ratings across items ($r = 0.38$, 95% CI = [0.33, 0.43], $p < .001$).

Figure A2 shows the average plausibility rating for each item in the four conditions (from Experiment 1) against the item-wise world knowledge norming score. As Experiment 2, the depth-charge condition represented by *No head injury is too trivial to be ignored* shows a clear positive correlation between world knowledge and plausibility rating. The Bayesian multilevel cumulative model had the same structure and prior setup as that in Experiment 2: it took the raw plausibility score per trial as the dependent variable; the fixed effects were the dummy-coded condition, the centered mean of world knowledge score per item, and their interaction; the random effect contained random intercepts as well as random slopes of all fixed effects for both subjects and items. As world knowledge scores increased, the plausibility ratings in the depth-charge condition increased, too ($\beta = 0.38$, HPD = [0.15, 0.59]). Similar to Experiment 2, HPDs for the effect of world knowledge on plausibility rating all contained zero for the other three conditions (quantifier-*some* & positive-adjective: $\beta = 0.19$, HPD = [−0.26, 0.64]; quantifier-*some* & negative-adjective: $\beta = -0.05$, HPD = [−0.33, 0.22]; quantifier-*no* & positive-adject: $\beta = -0.29$, HPD = [−0.57, 0.004]).

A.3. Discussion

The results of this study are consistent with the results in Experiment 2, which used a different set of sentences to reflect the intended meaning, suggesting that world knowledge has an impact on depth-charge sentence interpretation but its impact on the interpretation of the other types of sentences is unclear.

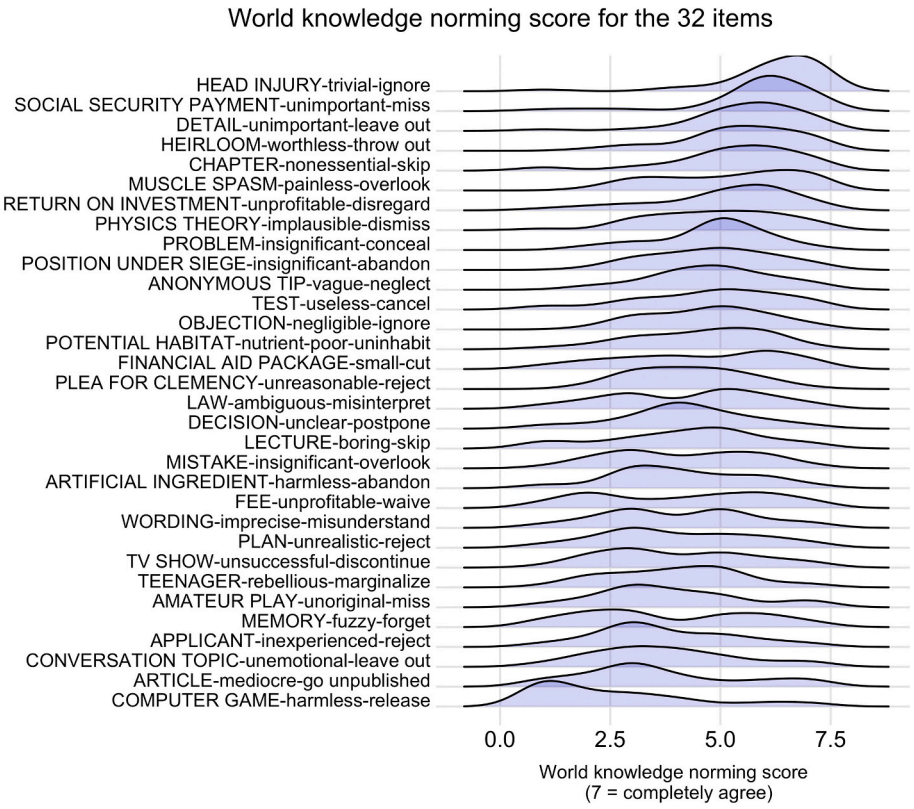

Fig. A1. Distribution of world knowledge norming scores across the 32 items.

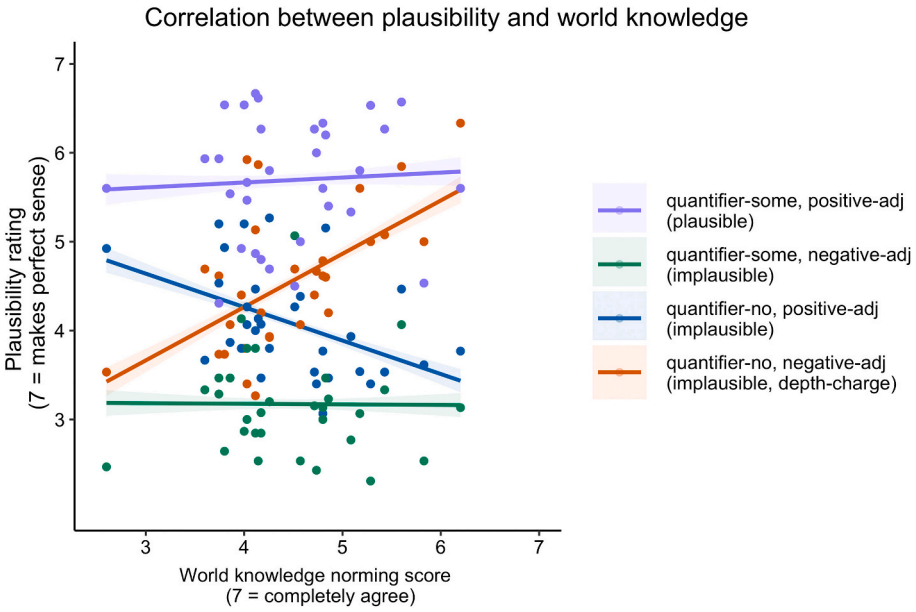

Fig. A2. Plausibility rating against the item-wise world knowledge norming score across the four conditions (Both the points and the regression line are average score per critical item).

Appendix B. Deletion/insertion of *not* as a possible noise operation

We propose that during production of the intended meaning underlying depth-charge sentences, speakers might lose track of the number of negations (cf. Horn, 2009) and cause the deletion or insertion of *not* in the perceived sentence. For instance, the speaker may utter *I didn't say nothing* while intending *I didn't say anything* or *I said nothing* (Fay, 1981). Readers may consider a deletion or insertion of *not* to be a plausible noise operation and interpret depth-charge sentences according to a more plausible alternative that differs by such an operation. We hypothesize that the intended sentence s_i could be *No head injury is too trivial not to be ignored* where *not* is deleted during information transmission and the perceived sentence s_p becomes the depth-charge sentence. Or the intended sentence s_i could be *No head injury is too trivial to be treated* where *not* is inserted and the perceived sentence s_p becomes the implausible *No head injury is too trivial not to be treated*. Following the predictions of the noisy-channel framework, the offline

comprehension task in Experiment A2 predicts that perceived sentences resulting from the deletion of *not* should receive more inference than sentences resulting from the insertion of *not*.

B.1. Methods

B.1.1. Participants

A total of 80 participants were recruited from Amazon's Mechanical Turk and the payment for participation was $3. We deleted the data from participants who self-identified or were detected by our sentence completion task as non-native speakers of English and whose answers to the attention-checking question of the filler trials were less than 90% correct. We analyzed data from 56 participants.

B.1.2. Materials & procedure

Table B1 provides an example item across the 4 conditions. The experiment consisted of 24 critical items with a high world-knowledge norming score according to Appendix A. We manipulated sentence plausibility and the proposed noise operation within-subjects, using a Latin Square design. Each trial was accompanied by a comprehension question (e.g., *According to the sentence, should head injuries be ignored/treated?*) with a YES/NO forced choice. The polarity of the verb was counterbalanced within each item and participants' choices were converted to indicate whether they interpreted the sentence literally or not.

Table B1
Example item across conditions in Experiment A2.

Noise	Plausible	Implausible
deletion	No head injury is too trivial to not be ignored.	No head injury is too trivial to be ignored (depth-charge).
insertion	No head injury is too trivial to be treated.	No head injury is too trivial to not be treated.

Besides the critical items, there were 60 plausible fillers each describing a generic property of or an attitude toward a topic (e.g., *Never can we walk away without resolving problems which have immeasurable importance*.). Each filler was accompanied by a comprehension question and the YES/NO answer was counterbalanced across the fillers.

B.2. Results

As shown in Fig. B1, the overall rates of literal interpretation for implausible sentences were very low (~10%). Implausible sentences which are assumed to be produced after the deletion of *not* appeared to receive higher literal interpretation rate and lower inference rate, compared with the implausible sentences in the insertion condition. Notably, the rates of literal interpretation for the plausible sentences in the deletion condition were also lower than expected (~75%).

We ran a Bayesian multilevel ordinal model for statistical analysis. A literal interpretation was coded as 1 and an inferential reading was 0. The model included the dummy-coded error type (deletion vs. insertion), the plausibility of the sentences, and the interaction term as the fixed effects; the random effects included all of the fixed effects for subjects and items. The prior and all the other settings were the same as Experiment 4. Implausible sentences were less likely to receive literal interpretations ($\beta = -1.41$, CrI = $[-2.53, -0.32]$). Within the implausible sentences, those that resulted from the deletion of *not* were not less likely to receive literal interpretation than those that resulted from the insertion of *not*, in fact the effect could be more likely in the opposite direction ($\beta = 0.945$, HPD = $[-0.05, 1.98]$).

B.3. Discussion

Contrary to the noisy-channel prediction, the condition with the deletion of *not* did not trigger more inference than the insertion of *not*. The reason for the contradictory finding could be that adding or deleting *not* in a sentence which already has many negative meanings might be hard and overly unnatural. This explanation is consistent with the relatively low rates of literal interpretation of the plausible sentences in the deletion condition which contain two explicit negation terms (*no* and *not*). Sentence structures of this type may be much lower in prior probability such that readers may infer that some other sentence was intended even when they contain no semantic implausibility (Keshev & Meltzer-Asscher, 2021; Poliak et al., 2022). Note that the fact that these results are inconsistent with the noisy-channel prediction does not mean that the noisy-channel framework fails to explain depth-charge illusions more generally. These findings suggest that the noise model involving a deleted negation may not be that which readers are considering during comprehension of depth-charge sentences and that multiple negations may affect the structural prior.

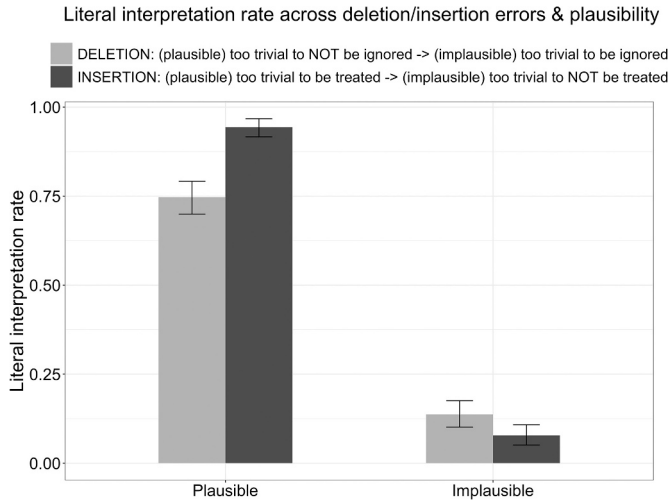

Fig. B1. Literal interpretation rate in Appendix B by noise operations and sentence plausibility (with 95% bootstrapped CI).

References

Keshev, M., & Meltzer-Asscher, A. (2021). Noisy is better than rare: Comprehenders compromise subject-verb agreement to form more probable linguistic structures. *Cognitive Psychology*, *124*, 101359.

Poliak, M., Ryskin, R., Braginsky, M, & Gibson, E. (in prep). It's not what you say but how you say it: Evidence from Russian shows robust effects of the structural prior on noisy channel inferences.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods, 40*(1), 278–289.

Allen, M. P. (1997). The problem of multicollinearity. *Understanding Regression Analysis*, 176–180.

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. https://doi.org/10.1016/j.jml.2012.11.001

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.

Bock, K. (1996). Language production: Methods and methodologies. *Psychonomic Bulletin and Review, 3*, 395–421.

Box, G. E., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis* (Vol. 40). New York: John Wiley & Sons.

Bürkner, P. C. (2017). Brms: An *R* package for Bayesian multilevel models using *Stan*. *Journal of Statistical Software, 80*(1). https://doi.org/10.18637/jss.v080.i01

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.

Büttner, A. C. (2007). Questions versus statements: Challenging an assumption about semantic illusions. *Quarterly Journal of Experimental Psychology, 60*(6), 779–789. https://doi.org/10.1080/17470210701228744

Cantor, A. D., & Marsh, E. J. (2017). Expertise effects in the Moses illusion: Detecting contradictions with stored knowledge. *Memory, 25*(2), 220–230. https://doi.org/10.1080/09658211.2016.1152377

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology, 42*(4), 368–407.

Core Team, R. (2018). *R: A language and environment for statistical computing. R Foundation for Statistical Computing. Version 3.4.4*.

Christianson, K., Williams, C. C., Zacks, R. T., & Ferreira, F. (2006). Younger and older adults'" good-enough" interpretations of garden-path sentences. *Discourse Processes, 42*(2), 205–238.

Cook, P., & Stevenson, S. (2010). No sentence is too confusing to ignore. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground* (p. 9).

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism, 2*(3), 219–253.

Davies, M. (2015). *Corpus of Contemporary American English (COCA)*. https://doi.org/10.7910/DVN/AMUDUW

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. https://doi.org/10.1016/S0022-5371(81)90202-4

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. https://doi.org/10.1016/S0022-5371(81)90165-1

Fay, D. (1981). Substitutions and splices: A study of sentence blends. *Linguistics, 19*(7–8). https://doi.org/10.1515/ling.1981.19.7-8.717

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*(2), 164–203.

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11–15. https://doi.org/10.1111/1467-8721.00158

Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. In , *Vol. 65. Psychology of learning and motivation* (pp. 217–247). Academic Press.

Ferreira, F., & Patson, N. (2007). The "good enough" approach to language comprehension. *Lang & Ling Compass, 1*, 71–83. https://doi.org/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). https://doi.org/10.1515/cog-2014-0014

Frazier, L. (1985). Syntactic complexity. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language processing: Psychological, computational and theoretical perspectives* (pp. 129–189). Cambridge, UK: Cambridge University 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), Article e12814.

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Vol. 1. Language production* (pp. 177–220). London: Academic Press.

Garrett, M. F. (1988). Processes in language production. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge survey: III. Language: Psychological and biological aspects* (pp. 69–96). Cambridge: Cambridge University Press.

Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review, 96*(2), 267.

Geisler, W. S., & Diehl, R. L. (2003). A Bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science, 27*(3), 379–402.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 457–472.

Giannouli, V. (2016). A verbal illusion reexamined. *Acta Neuropsychologica, 14*(4), 324–329.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences, 110*(20), 8051–8056. https://doi.org/10.1073/pnas.1216438110

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. https://doi.org/10.1111/j.1749-818X.2011.00295.x

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language & Cognitive Processes, 14*(3), 225–248.

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. https://doi.org/10.1016/j.tics.2022.01.005

Gross, H. S. (1983). Errors in linguistic performance: Slips of the tongue, ear, pen, and hand. *The Journal of Nervous and Mental Disease, 171*(12), 753. https://doi.org/10.1097/00005053-198312000-00008
Hacquard, V. (2005). Aspects of "too" and "enough" constructions. *Semantics and Linguistic Theory*, 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). https://doi.org/10.1073/pnas.2122602119
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. https://doi.org/10.1016/S0364-0213(84)80001-4
Harley, T. A., & MacAndrew, S. B. G. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research, 30*(4), 395–418. https://doi.org/10.1023/A:1010421724343
Heim, I. (2000). Degree operators and scope. *Semantics and Linguistic Theory*, 40–64.
Horn, L. R. (2009). Hypernegation, hyponegation, and parole violations. *Annual Meeting of the Berkeley Linguistics Society, 35*(1), 403. https://doi.org/10.3765/bls.v35i1.3628
Horn, L. R. (2010). Multiple negation in English and other languages. In *The expression of negation* (pp. 111–148). De Gruyter Mouton.
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*, 97–109.
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. https://doi.org/10.1016/j.jml.2018.05.006
Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior, 10*(3), 244–253. https://doi.org/10.1016/S0022-5371(71)80051-8
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.
Kapatsinski, V. (2010). Frequency of use leads to automaticity of production: Evidence from repair in conversation. *Language and Speech, 53*(1), 71–105. https://doi.org/10.1177/0023830909351220
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. https://doi.org/10.1080/17470218.2015.1053951
Kay, P., & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: The What's X doing Y? construction. *Language, 75*(1), 1–33. https://doi.org/10.1353/lan.1999.0033
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. https://doi.org/10.1080/02643290701674851
Kizach, J., Christensen, K. R., & Weed, E. (2016). A verbal illusion: Now in three languages. *Journal of Psycholinguistic Research, 45*(3), 753–768. https://doi.org/10.1007/s10936-015-9370-6
Ladusaw, W. A. (1980). *Polarity sensitivity as inherent scope relations*. New York: Garland.
Leivada, E. (2020). Language processing at its trickiest: Grammatical illusions and heuristics of judgment. *Languages, 5*(3), 29. https://doi.org/10.3390/languages5030029
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *Package 'emmeans'*.
Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In Mirella Lapata, & Hwee Tou Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
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.
Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: Formal techniques and empirical results. In Dekang Lin, et al. (Eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 1055–1065). Association for Computational Linguistics.
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.
Meier, C. (2003). The meaning of too, enough, and so…that. *Natural Language Semantics, 11*, 69–107. https://doi.org/10.1023/A:1023002608785
Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence, 29*(3), 436–465.
Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.
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. https://doi.org/10.1007/BF01067882
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.
Nicenboim, B., Schad, D. J., & Vasishth, S. (2021). An introduction to Bayesian data analysis for cognitive science. In *Under contract with Chapman and Hall/CRC Statistics in the Social and Behavioral Sciences Series*.
O'Connor, E., Pancheva, R., & Kaiser, E. (2013). Evidence for online repair of Escher sentences. *Proceedings of Sinn und Bedeutung, 17*, 363–380. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/350.
O'Connor, E. (2015). *Comparative illusions at the syntax-semantics interface*. University of Southern California.
O'Connor, E. (2017). The (accidental) ambiguity of inversion illusions. *Proceedings of NELS, 47*, 329–342.
Paape, D., Vasishth, S., & von der Malsburg, T. (2020). Quadruplex negatio invertit? The on-line processing of depth charge sentences. *Journal of Semantics, 37*(4), 509–555. https://doi.org/10.1093/jos/ffaa009
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*(3), 207–235.
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. https://doi.org/10.1075/cf.00021.pij
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. https://doi.org/10.1515/flin-2016-0020
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).
Ryskin, R., Futrell, R., Kiran, S., & Gibson, E. (2018). Comprehenders model the nature of noise in the environment. *Cognition, 181*, 141–150. https://doi.org/10.1016/j.cognition.2018.08.018
Ryskin, R., Stearns, L., Bergen, L., Eddy, M., Fedorenko, E., & Gibson, E. (2021). An ERP index of real-time error correction within a noisy-channel framework of human communication. *Neuropsychologia, 158*, Article 107855.
Sanford, A. J., & Emmott, C. (2012). *Mind, Brain and Narrative*. Cambridge University Press.
Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences, 6*(9), 382–386. https://doi.org/10.1016/S1364-6613(02)01958-7
Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423.
Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior, 15*(2), 143–157. https://doi.org/10.1016/0022-5371(76)90015-3
Stemberger, J. P. (1984). Structural errors in normal and agrammatic speech. *Cognitive Neuropsychology, 1*(4), 281–313. 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 B. J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition*. New York: Plenum Press.
Traxler, M. J. (2014). Trends in syntactic parsing: Anticipation, Bayesian estimation, and good-enough parsing. *Trends in Cognitive Sciences, 18*(11), 605–611.
Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*(2), 127–154.
Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315–346.
Wang, L., Hagoort, P., & Yang, Y. (2009). Semantic illusion depends on information structure: ERP evidence. *Brain Research, 1282*, 50–56. https://doi.org/10.1016/j.brainres.2009.05.069
Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *Quarterly Journal of Experimental Psychology, 31*(4), 591–597. https://doi.org/10.1080/14640747908400750
Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics, 35*(3), 543–583.
Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language & Cognitive Processes, 25*(4), 533–567.