A noisy-channel explanation for depth-charge illusions

Yuhan Zhang (*Harvard University, yuz551@g.harvard.edu*), Rachel Ryskin (*University of California, Merced*), Edward Gibson (*Massachusetts Institute of Technology*)

"Depth-charge" sentences (e.g., *No head injury is too trivial to be ignored*) are overwhelmingly interpreted as plausible although their literal meanings are implausible (e.g., "head injuries should be ignored") (Wason & Reich, 1979). Previous attempts to explain the source of the illusion have appealed to shallow processing (Sanford & Sturt, 2002) and the complicated meaning composition of the sentential negation with the structure *too...to*, polar adjectives like *trivial*, and polar verbs like *ignore* (e.g., Fortuin, 2014; Paape et al., 2020). In this study, **we test whether comprehension of depth-charge sentences can be construed as a process of noisy-channel inference** (Levy, 2008; Gibson et al. 2013; Ryskin et al., 2018). On this view, readers may interpret the perceived depth-charge sentence, s_p, as a more plausible intended sentence, s_i, (e.g., *No head injury is so trivial as to be ignored*) which has been corrupted by noise during information transmission (e.g., a production error). We observe that the patterns of nonliteral comprehension (i.e., the probability of interpreting a depth charge as a more plausible alternative, $P(s_i \mid s_p)$) of these sentences are consistent with a rational inference modulated by the prior probability of the intended meaning $P(s_i)$ and a noise model that encodes how an intended sentence is likely to be corrupted during communication $P(s_p \mid s_i)$ (as in Eq.1 in (1)).

In **Exp.1**, participants (N = 58) rated the plausibility of 32 depth-charge sentences (translated from Paape et al.'s (2020) German items) and controls (Table 1). A linear mixed-effects regression (LMER) with quantifier (some vs. no), adjective (e.g., trivial vs. severe), and their interaction as predictors revealed a significant interaction ($\beta = 0.75$, SE = 0.04, $p < .001$, Fig.1), such that readers interpreted depth-charge materials as more plausible relative to other sentences with implausible literal meanings, replicating Paape et al. (2020).

In **Exp.2**, participants (N = 36) rated the consistency with world knowledge of the intended meaning (e.g., "head injuries are in general too severe to be ignored") of the 32 test items from Exp.1 to provide a measure of the prior, $P(s_i)$. An LMER with quantifier, world knowledge score, and their interaction as predictors (only negative adjective sentences were included) indicated that the higher the consistency with world knowledge, the more plausible the depth-charge sentence (interaction: $\beta = 0.28$, SE = 0.07, $p < .001$, Fig.2).

In **Exp.3**, we investigated alternative sentences which could lead to a depth-charge interpretation when corrupted by noise. In particular, we posited that a structural substitution (*so...as to* → *too...to*) is a plausible noise operation (Table 2) based on their similar local syntactic environments and semantic functions. Because substitutions are also more likely to occur on low-frequency words (Harley & MacAndrew, 2001), *too...to* is more frequent than *so...as to* (8828ct vs. 500ct in the *COCA* corpus), and deletion is more likely than insertion (Gibson et al. 2013), the corruption in the other direction (*too...to* → *so...as to*) is less likely. In line with this prediction, participants (N = 43) rated the likelihood that, e.g., *No head injury is so trivial as to be ignored* was corrupted to *No head injury is too trivial to be ignored* as higher than a corruption in the other direction (e.g., *No head injury is too trivial to be treated* → *No head injury is so trivial as to be treated*; $\beta = -0.74$, SE = 0.20, $p < .001$)

In **Exp.4**, participants (N = 47) were asked to answer yes/no comprehension questions (e.g., "According to this sentence, should head injuries be treated?") about 24 test items (all with high P(Si)) crossing plausibility and structure (Table 2) and 40 fillers. "No" indicates a literal interpretation (counterbalanced). While the plausible sentences were overwhelmingly interpreted literally, the implausible sentences with *too...to* elicited significantly more inferences than the implausible sentences with *so...as to* (interaction between plausibility and structure in logistic MER, $\beta = 1.84$, SE = 0.71, $p < .01$, Fig. 3).

Across Exp.1 to 4, depth-charge sentences are more likely to be interpreted as a more plausible alternative when 1) world knowledge strongly supports that alternative -- $P(S_i)$ is high -- and 2) when the noise corruption that could have transformed a plausible alternative into the perceived sentence is likely -- $P(S_i \rightarrow S_p)$ is high. Taken together, these results are consistent with a noisy-channel explanation for the depth-charge illusion.

(1) $P(S_i|S_p) \propto P(S_i)\, P(S_i \rightarrow S_p)$

Table 1: Critical item design in Exp.1

Quantifier	Adjective	Sentence	Plausibility
SOME	positive	Some head injuries are too severe to be ignored.	Yes
SOME	negative	Some head injuries are too trivial to be ignored.	No
NO	positive	No head injury is too severe to be ignored.	No
NO	negative	No head injury is too trivial to be ignored.	No (depth-charge)

Table 2: Noise operation conditions tested in Exp.3 & 4

Condition	Plausible	Implausible	Noise operation
so → too	No head injury is so trivial as to be ignored.	No head injury is too trivial to be ignored.	*so → too*, deletion of *as*
too → so	No head injury is too trivial to be treated.	No head injury is so trivial as to be treated.	*too → so*, insertion of *as*

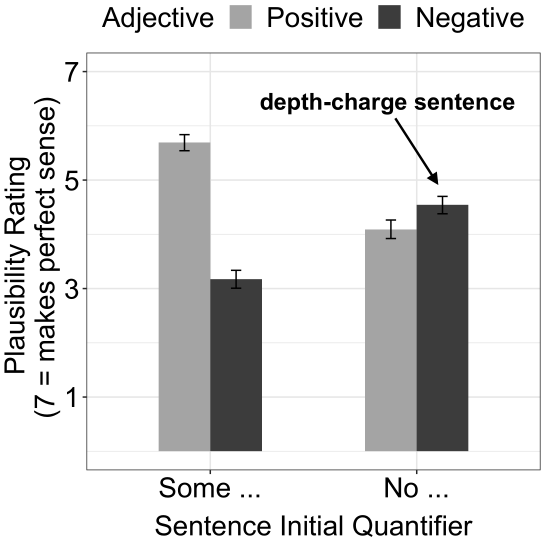


Fig.1 Plausibility rating by quantifier and adjective in Exp.1 (with 95% CI obtained via bootstrapping)

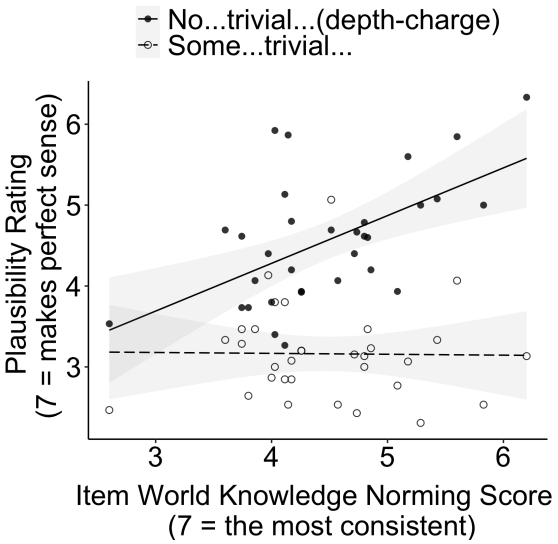
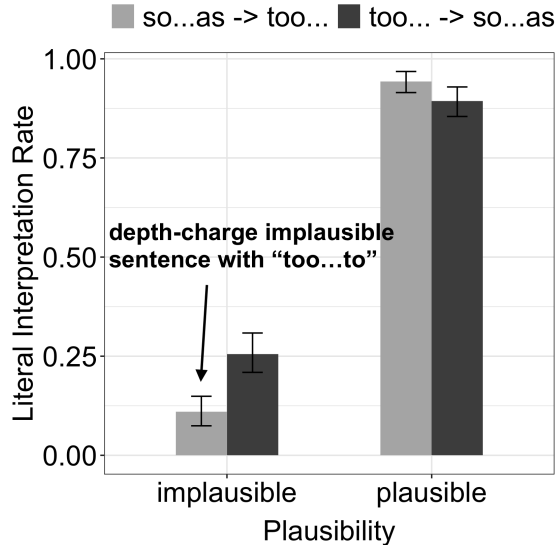


Fig.2 Plausibility rating is positively correlated with world knowledge in depth charges. (95% CI)



Selected References:
[1] COCA: *Corpus of Contemporary American English*. [2] Fortuin (2014). *Cognitive Linguistics*. [3] Gibson et al. (2013). *PNAS*. [4] Harley & MacAndrew (2001). *J of Psycholinguistic Research*. [5] Levy (2008). *EMNLP*. [6] Paape et al. (2020). *J of Semantics*. [7] Ryskin et al. (2018). *Cognition*. [8] Sanford & Sturt (2002). *Trends in Cognitive Sciences*. [9] Wason & Reich (1979). *Quarterly Journal of Experimental Psychology*.

Fig 3. Literal interpretation rate by plausibility and posited noise corruption (95% CI) (e.g., *No head injury is too trivial to be ignored* corresponds to 'so as..to' being substituted by 'too..to'). Comprehension question example: "According to this sentence, should head injuries be treated?"