Quadruplex negatio invertit? The on-line processing of depth charge sentences

Dario Paape,[1,a] Shravan Vasishth,[1] Titus von der Malsburg[1,2]

[1]Human Sciences Faculty, Department of Linguistics, University of Potsdam
[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[a] Corresponding author (paape@uni-potsdam.de)

Draft of March 13, 2020
Submitted to Journal of Semantics

Abstract

So-called "depth charge" sentences (*No head injury is too trivial to be ignored*) are interpreted by the vast majority of speakers to mean the opposite of what their compositional semantics would dictate. The semantic inversion that is observed for sentences of this type is the strongest and most persistent linguistic illusion known to the field (Wason & Reich, 1979). However, it has recently been argued that the preferred interpretation arises not because of a prevailing failure of the processing system, but rather because the non-compositional meaning is grammaticalized in the form of a stored construction (Cook & Stevenson, 2010; Fortuin, 2014). In a series of five experiments, we investigate whether the depth charge effect is better explained by processing failure due to memory overload (the *overloading hypothesis*) or by the existence of an underlying grammaticalized construction with two available meanings (the *ambiguity hypothesis*). To our knowledge, our experiments are the first to explore the on-line processing profile of depth charge sentences. Overall, the data are consistent with specific variants of the ambiguity and overloading hypotheses while providing evidence against other variants. As an extension of the overloading hypothesis, we suggest two heuristic processes that may ultimately yield the incorrect reading when compositional processing is suspended for strategic reasons.

Keywords depth charge, semantic illusion, negation processing, eye tracking

1 Introduction

Most English native speakers interpret the sentence *No head injury is too trivial to be ignored* to mean that *All head injuries, no matter how trivial they appear, should be treated* (Wason & Reich, 1979). The compositional semantics of the sentence, however, would dictate that it means *All head injuries should be ignored, even if they seem trivial enough to treat.* The commonly observed transformation of the nonsensical but correct meaning into a sensible but incorrect one is known as the depth charge effect. Wason & Reich speculated that three factors contribute to the meaning reversal:

1. The presence of (arguably) up to four negations: The initial *no* introduces a negative existential quantification over the entire clause. The particle *too* implicitly negates the following infinitive (*X is too Y to Z* → *X should not Z*). The word *trivial* can, as Wason & Reich argue, be analyzed as meaning *not serious*. Finally, the verb *ignore* arguably means *not treat*.[1] It is well known that multiple negation increases processing difficulty and causes misinterpretations (Sherman, 1976).
2. The absurdity of the scale invoked by *too trivial to be ignored*, which implies that something can be so trivial that one should *not* ignore it (compare *trivial enough to be ignored*, where the scale makes sense).
3. The incompatibility of the compositionally correct meaning (*Ignore all head injuries*) with world knowledge: Based on their experience, most people would agree that even minor head injuries are or at least can be a reason for concern.

Wason & Reich propose that these factors conspire and cause a switch from compositional to non-compositional processing. Specifically, Wason & Reich suggest that the stacking of four negations may "overload" the reader's working memory[2], with the other two factors – absurdity of the evoked scale and world knowledge – contributing to consistent misinterpretation. We call this the *overloading* account. Furthermore, Wason & Reich (1979) hypothesize that the final verb *ignore* is the point where the sentence becomes impossible to analyze compositionally and the switch occurs.

1.1 Previous empirical studies of the depth charge effect

In a series of small-sample experiments, Wason & Reich (1979) found that 4 out of 10 subjects were unable to assign the correct meaning to the depth charge construction even when world knowledge and lexical factors were taken out of the equation (*No WUG is too DAX to be ZONGED*), and that misinterpretation was more easily avoided when the compositional meaning agreed with world knowledge (gauged introspectively by the authors). The latter finding matches the observation of Fillenbaum (1974) that readers apply "pragmatic normalization" to sentences such as *Don't print that or I won't sue you!* in order to turn the compositionally nonsensical meaning into one that conforms with their a priori expectations about sensible propositions (see also Sanford & Sturt, 2002).

[1] Analyzing the words *trivial* and *ignore* as containing lexical negation is somewhat problematic, given that one could equally well argue *treat* to mean *not ignore*. What Wason & Reich (1979) appear to be using as a guideline is likely what O'Connor (2015) refers to as "affective negativity" and what is called "prior polarity" in automatic sentiment analysis (Wilson et al., 2005), namely the idea that some words are inherently more "negative" than others.

[2] Wason & Reich use the term "channel capacity", but we equate the two concepts in the present work.

Natsopoulos (1985) and Kizach et al. (2015) investigated the depth charge effect in Greek and Danish, respectively. In his first study, Natsopoulos (1985) used eight depth charge sentences as stimuli, finding that none of his 64 subjects were able to derive the compositional meaning. Whether a sentence evoked "strong beliefs" or not did not influence the incidence of meaning reversal, contrary to Wason & Reich's finding. In a second study, when subjects were asked to pick a paraphrase out of three options, the compositionally correct (absurd) meaning was chosen about 50% of the time. The third experiment yielded some evidence that "strong beliefs" in favor of the illusory meaning led to more incorrect interpretations. Natsopoulos's studies used only a small number of items, and show a high level of variability between these: Some sentences were parsed correctly by a majority of subjects while others were almost always parsed incorrectly. Nevertheless, the results suggest that pragmatic reasoning can strengthen the depth charge effect while providing explicit paraphrasing choices will at least partly cancel it.

The experiment of Kizach et al. (2015) (29 subjects, 150 sentences) varied the three factors originally noted by Wason & Reich (1979) – number of negations, internal consistency of the scale and pragmatics – independently. Each stimulus sentence was followed by a conclusion, e.g. *No head injury is too trivial to be ignored. Therefore, we rarely treat head injuries.* Participants were asked to judge whether the conclusion made sense given the premise. After each judgment, they were also asked if their answer had been a guess. Results showed that participants resorted to guessing more often when negation was present on the adjective and the verb. However, even in the depth charge condition, where all three critical factors were present, participants' self-reported guessing rate was only about 20%, suggesting that subjects did not experience conscious processing failure on most trials. The depth charge condition showed the lowest comprehension accuracy (around 40%), followed by sentences with only a world knowledge violation (about 55%), the negation condition (about 65%), and finally the scale violation condition (about 75%). These results imply that the inverted meaning becomes entrenched once it has been generated, given that confidence in the erroneous interpretation is high, and that the world knowledge factor is a stronger driver of the inversion than the scale violation factor.

Furthermore, the results of Kizach et al. (2015) imply that not all of the problematic factors that are present in the classic depth charge configuration studied by Wason & Reich (1979) are necessary in order to obtain the meaning inversion effect. For instance, the "negation" of the adjective does not appear to be a prerequisite for meaning inversion, as examples of depth charge sentences without negative adjectives can be found in real-life corpora (e.g. *No challenge is too big to stop us from saving our children from polio*), as also observed by Fortuin (2014, p. 253f.). The overloading account therefore needs to accommodate the possibility that three negations may, in some circumstances, be enough to trigger meaning inversion.

O'Connor (2015; 2017) conducted another series of experiments in English investigating the contribution of the multiple negations to the depth charge effect. Results showed that especially the combination of global negation (*no head injury ...* compared to *all head injuries ...*) and the element *too* (compared to *enough*) led to a superadditive increase in misinterpretations, though a possible additional effect of adjectival negation was not investigated.

1.2 An alternative view: The ambiguity hypothesis

In opposition to the overloading account of Wason & Reich (1979), some scholars argue that there is a fundamental ambiguity to the meaning of sentences of the form *No X is too Y to Z*, and

that one of the available meanings is the inverted meaning (Cook & Stevenson, 2010; Fortuin, 2014). We call this account the *ambiguity hypothesis*. There are two variants of the ambiguity hypothesis that make somewhat different sets of predictions. The first variant is proposed in the form of a computational model by Cook & Stevenson (2010). It shares with the account of Wason & Reich (1979) the prediction that the final verb of the sentence is the source of the depth charge effect, as its polarity ("negative" or "positive") presumably signals which meaning of the *No X is too Y to Z* construction is intended. Furthermore, this version of the ambiguity hypothesis predicts that world knowledge should not have an effect on meaning inversion, as the information given by the verb is assumed to be sufficient to derive the intended meaning. The second variant of the ambiguity hypothesis is proposed by Fortuin (2014). It assumes that the origin of the depth charge effect lies before the lexical verb, when the word *too* is processed. Unlike the account of Cook & Stevenson (2010), Fortuin's account predicts that the plausibility of the presumably intended meaning can affect interpretation, as readers are assumed to use their world knowledge to identify the correct version of the stored construction.

Fortuin (2014) analyzes the classic depth charge configuration *No X is too Y to Z* as "[a] conventionalized combination of form-meaning elements" (p. 250), in the vein of construction grammar (e.g. Fillmore, 1985; Lambrecht, 1988; Goldberg, 1995; see also Cook & Stevenson, 2010). Using examples from real-life corpora, Fortuin argues that the construction can attain either a "negative" or a "positive" meaning. In Fortuin's terminology, "negative" meanings are cases in which inversion occurs, but as both meanings are properly licensed by the use of the construction, there is no "illusion" involved. For instance, the preamble *No detail is too small . . .* is shown to license both the continuation *. . . to be ignored* as well as its semantic opposite *. . . to pay attention to* with little change in the resulting meaning, as long as the context implies that details are important (p. 272). Fortuin (2014) also gives the example *No detail is too small to escape his notice or merit his attention*, where a "positive" and a "negative" reading are apparently licensed simultaneously (p. 275).

Fortuin concentrates on the presumed communicative (or rhetorical) intention behind the depth charge sentence. The "negative" version of the construction is arguably produced if the intent is for the reader (or hearer) to draw a negative inference. In the classic *No head injury is too trivial to ignore* example, the intended message is arguably that head injuries should *not* be ignored. On the other hand, if the writer (or speaker) intends a positive inference, he or she will accordingly produce the "positive" version (e.g. *No idea is too silly to discuss → discuss all ideas*, p. 272). From a processing perspective, this leaves the receiver with the (possibly challenging) task of reasoning about the intention of the utterer, which must be inferred from the context and the lexical items used. However, the task may be rendered less difficult by the fact that most instances of the construction can be identified as being "positive" or "negative" based on the lexical verb alone: The computational model of Cook & Stevenson (2010) reaches 88% classification accuracy using only the lexical features of the verb.

1.3 Open questions

Researchers' continued interest in the depth charge effect is likely due to the fact that it calls compositionality itself into question, which is a provocation for every formal-minded linguist; after all, without a rule system that determines possible interpretations, language dissolves into

"semantic soup" (Anderson, 2006).[3] The provocation becomes even greater when, as already observed by Wason & Reich (1979), both laypeople and fellow linguists stubbornly insist that they are interpreting the sentence normally and correctly. Here, it should be noted that the observed ability of the construction to attain both "negative" and "positive" meanings does not necessitate the assumption that this behavior is sanctioned by grammar, as assumed by the ambiguity hypothesis: Despite the observation that meaning inversion sometimes occurs and sometimes does not, and is not limited to sentences with four negations, the depth charge effect may nevertheless be a performance- rather than a competence-driven phenomenon (Chomsky, 1964), as claimed by the overloading hypothesis.

The ambiguity hypothesis of Fortuin (2014) and Cook & Stevenson (2010) draws explanatory force from construction grammar, which assumes complex, pre-compiled meanings as its central tenet. Interestingly, among the authors who argue for a processing-based explanation of the depth charge effect, none have proposed an explicit formal mechanism by which the unlicensed meaning in depth charge sentences is derived. One theory by O'Connor (2015) is that "comprehenders interpret implicit negation [introduced by *too*] as semantically inert [...] thus conflating the logical force of two or more negative elements of the sentence" (p. 168), based on Horn's (2009) general observation that multiple negation is often not interpreted as expected. Under the overloading hypothesis, one possible account is that the implicit negation is dropped from memory when the mental capacity limit is reached, resulting in the erroneous meaning. It is thus possible that processing difficulty is reduced when overloading is triggered compared to when a fully compositional interpretation is computed, given that fewer negations have to be taken into account. If the intuition of Wason & Reich (1979) is correct, such an effect should become visible at the final "negative" verb.

An alternative view is that subjects enter into a different mode of processing when compositional semantics fails. This mode may be driven by world knowledge and possibly semantic associations between the lexical items in the sentence. It could be envisioned as treating the sentence as a "bag-of-words" rather than as an internally structured utterance (*no* + *head injury* + *trivial* + *ignore* ≈ *Ignore no trivial head injury*). If this seems extreme, note that in computational natural language processing, negation also poses a challenge (Wiegand et al., 2010), but respectable accuracy in sentiment detection (between 80% and 90%) can be achieved using bag-of-words or local n-gram representations (Pang et al., 2002; Ng et al., 2006). This suggests that, especially when applied in tandem with general world knowledge, such representations may often be sufficient to ensure successful communication (e.g. Jackendoff & Wittenberg, 2014). In terms of on-line processing, treating the sentence as less structured than it actually is would also predict reduced processing difficulty in depth charge sentences, given that the compositional computation of meaning is likely more effortful than using a "bag-of-words" approach. The ambiguity hypothesis makes the same prediction, given that the *No X is too Y to Z* construction does not receive a (fully) compositional interpretation when the inference is negative.

If overloading is the correct explanation for the depth charge effect, there may be individuals with enough cognitive capacity to overcome the challenge posed by depth charge sentences. It has been suggested that high working memory capacity can help subjects overcome retrieval difficulties during the completion of verb-argument dependencies in high-interference contexts (King & Just, 1991; Nicenboim et al., 2016), and that individuals with low working memory

[3] See, however, McClelland et al. (1989) and Rabovsky et al. (2018) for a model of sentence processing that does not require explicit compositional representations of meaning.

capacity may construct less detailed syntactic representations of sentences, as evidenced by their reduced sensitivity to garden-path structures (von der Malsburg & Vasishth, 2013). High working memory capacity could also help subjects avoid overload in depth charge sentences and preclude the resulting switch from compositional to non-compositional processing that is thought to bring about the illusion. Note, however, that overloading of the reader's working memory is only one possible way in which the failure of compositional processing can be thought of. We will propose a variant of the overloading hypothesis that does not rest on the assumption of memory overload, but instead assumes that readers arrive at a limit to their intrinsic motivation and apply a "stop rule" (e.g. Simon, 1972), after which they switch to non-compositional processing.

The point of contention between proponents of the overloading hypothesis (Wason & Reich, 1979; Kizach et al., 2015) and those of the ambiguity hypothesis (Cook & Stevenson, 2010; Fortuin, 2014) is not whether the inverted meaning of depth charge sentences is fully compositional in nature or not: Both accounts assume that the inverted meaning cannot be derived by combining the lexical meanings of the words in the sentence according to a rule-based system. The disagreement is about whether the inverted reading is, fundamentally, due to an error in the processing system or whether it is, despite its non-compositionality, licensed by grammar by way of a pre-stored construction with different meanings. As noted by Cook & Stevenson, under a construction-based approach the inverted reading is not a bug but a feature: As opposed to being an edge case in terms of compositional processing, the depth charge configuration is seen as increasing the expressive power of the grammar beyond the compositional interpretation.[4] Note, however, that the assumption of a stored construction does not mean that there is no compositional processing whatsoever in depth charge sentences. Constructions may be grammaticalized to different degrees, and thus show different degrees of compositionality (Trousdale, 2012). Furthermore, in order for the construction to be understood in the intended way, it first needs to be recognized, which should only be possible after at least some of the words in the sentence have been read and analyzed compositionally.[5]

1.4 Contributions of the present work

Below, we present a series of four on-line experiments and one off-line experiment in German, as well as a sketch of possible non-compositional heuristics that may explain the depth charge effect. The main theoretical accounts to be investigated are the overloading hypothesis of Wason & Reich (1979) on the one hand and the ambiguity hypothesis on the other. The main research questions for each study are summarized in Table 1. To our knowledge, our experiments are the first to investigate the depth charge effect using on-line measures such as reading times and eye-tracking data. While previous experimental investigations have yielded valuable informa-

[4] In a sense, depth charge sentences can also be seen as "ambiguous" under the overloading hypothesis, because both the compositional and the inverted meaning can be derived without the comprehender noticing an error. However, this type of (subjective) ambiguity is more "accidental" in nature (O'Connor, 2017) and not sanctioned by grammar.

[5] In research on idiom processing, the related concepts of the idiom key – a word in the sentence that signals the presence of an idiomatic expression, such as *be in seventh* (heaven) (Cacciari & Tabossi, 1988) – and the idiom's recognition point – the point where most readers have access to the idiomatic meaning (Cacciari & Corradini, 2015) – are used to make predictions about when compositional processing is suspended. It is not clear what the recognition point of the *No X is too Y to Z* construction is under the ambiguity hypothesis. Based on the account of Cook & Stevenson (2010), it should lie at the lexical verb, that is, at the very end of the sentence.

tion regarding the final interpretation of depth charge sentences, our studies are designed to also shed light on how and when these interpretations arise during processing.

Study	Method	Research question(s)
Experiment 1	Whole-sentence reading + ratings	Can non-compositional processing be detected in on-line measures?
Experiment 2A	Eye-tracking + ratings	Is the final verb the source of the illusion? – Wason & Reich (1979): **yes** – Cook & Stevenson (2010): **yes** – Fortuin (2014): **no** Is meaning inversion affected by working memory capacity? – Wason & Reich (1979): **yes** – Cook & Stevenson (2010): **no** – Fortuin (2014): **no**
Experiment 2B	Ratings	Is meaning inversion affected by world knowledge? – Wason & Reich (1979): **yes** – Cook & Stevenson (2010): **no** – Fortuin (2014): **yes**
Experiment 3*	Whole-sentence reading + ratings	Is the depth charge effect limited to a specific construction? – Wason & Reich (1979): **possibly no** – Cook & Stevenson (2010): **possibly yes** – Fortuin (2014): **possibly no**
Experiment 4*	Sentence completions	Does meaning inversion occur in the absence of the final verb? – Wason & Reich (1979): **no** – Cook & Stevenson (2010): **no** – Fortuin (2014): **yes**

Table 1: Methods used and main research questions investigated across the five experimental studies.
* Experiments 3 and 4 were conducted in the reverse order but have been reordered for expository ease.

Experiment 1 tests whether the depth charge effect can be observed during on-line processing. Here, participants read depth charge as well as control sentences and assigned ratings of perceived sensibleness. Experiment 1 shows that depth charge sentences cause no additional processing difficulty compared to sentences with fewer negations, consistent with a partly non-compositional interpretation mechanism. Experiment 2A (eye tracking during reading) tests the prediction that the verb of the complement clause is the source of the illusion, either because it introduces the final negation (Wason & Reich, 1979), or because it signals the intended meaning of the construction (Cook & Stevenson, 2010). Experiment 2A also investigates whether readers with

high working memory capacity are more resistant to the illusion, which would be expected if the depth charge effect is due to memory overload. Results are compatible with the final verb being the source of the effect, but show no evidence of working memory capacity having an influence on meaning inversion. Thus, Experiment 2A does not support the account put forward by Wason & Reich (1979) as a specific instance of the overloading assumption. Experiment 2B investigates whether world knowledge plays a role in meaning inversion, as predicted by the overloading account as well as by one version of the ambiguity hypothesis (Fortuin, 2014). Experiment 2B suggests that world knowledge does indeed affect the magnitude of the meaning inversion effect. Experiment 3 tests whether the depth charge illusion generalizes beyond the classic *No X is too Y to ...* construction to related constructions (e.g. *No X is so Y that ...*). In contrast to the ambiguity account, the overloading account predicts that the illusion occurs more generally with multiple negation. Results suggest that the effect generalizes and occurs with equal strength as long as the element *too* is present, thus providing evidence against a strong version of the ambiguity hypothesis. As a possible extension of the overloading account, we propose two candidate heuristics, *negation cancellation* and *negate the verb*, that may be involved in creating the depth charge effect. Finally, Experiment 4 is a sentence completion study that serves as a more stringent test of the claim that the origin of the illusion lies at the verb. In contrast to the results from Experiment 2A, Experiment 4 shows that meaning inversion reliably occurs when the verb is missing.

To summarize, we find that non-compositional processing of depth charge sentences can be detected in on-line measures, and that the final verb is likely not the source of meaning inversion. Furthermore, we find that the depth charge effect is influenced by world knowledge, but find no evidence that high working memory capacity grants partial immunity to the effect. Finally, our results show that the depth charge effect generalizes to other constructions besides the *No X is too Y to Z* construction.

2 Experiment 1

Both the overloading hypothesis (Wason & Reich, 1979) and the ambiguity hypothesis (Cook & Stevenson, 2010; Fortuin, 2014) predict that non-compositional processing should occur in depth charge sentences. The main purpose of Experiment 1 is to establish whether this non-compositional processing can be made visible using an on-line processing paradigm, as previous empirical studies only provided off-line data. We use a twofold approach to detect non-compositional processing: By varying the number of negations in the sentence, we can test whether processing difficulty increases monotonically with each added negation, as would be expected under compositionality. In addition, by having participants assign ratings of sensibleness to each sentence, we can probe if incongruous sentences are "normalized" to have a sensible meaning when the depth charge effect occurs.

In our experimental design, we manipulated negation on the adjective as well as global negation of the sentence in potential depth charge configurations. As noted earlier, a negated or otherwise "negative" adjective is not a prerequisite for meaning inversion. However, looking at the data provided by Fortuin (2014, p. 268) as well as the results of Kizach et al. (2015), it appears that the presence of a "negative" adjective significantly increases the likelihood of the sentence receiving a "negative" (that is, inverted) interpretation.

2.1 Method

Participants Twenty native speakers of German were recruited from the local student population. Subjects either received credit points or were paid €5 as compensation.

Materials We constructed 32 items according to the design in (1). The presence of global negation as well as negation of the adjective were manipulated according to a 2 × 2 scheme. In 26 out of 32 items, adjectival negation was signaled by the presence of an overt negative affix such as *un-* or *-less* on the adjective.

(1) **Global negation absent, adjectival negation absent**
(NO NEGATION)

a. Manch eine Kopfverletzung ist zu gefährlich, um ignoriert zu werden.
Some a head injury is too dangerous to ignored to get
"Some head injuries are too dangerous to be ignored."

Global negation absent, adjectival negation present
(ADJECTIVAL NEGATION)

b. Manch eine Kopfverletzung ist zu ungefährlich, um ignoriert zu werden.
Some a head injury is too un-dangerous to ignored to get
"Some head injuries are too innocuous to be ignored."

Global negation present, adjectival negation absent
(GLOBAL NEGATION)

c. Keine Kopfverletzung ist zu gefährlich, um ignoriert zu werden.
No head injury is too dangerous to ignored to get
"No head injury is too dangerous to be ignored."

Global negation present, adjectival negation present
(DOUBLE NEGATION)

d. Keine Kopfverletzung ist zu ungefährlich, um ignoriert zu werden.
No head injury is too un-dangerous to ignored to get
"No head injury is too innocuous to be ignored."

The two types of negation control the "pragmatic" and "semantic" coherence of the sentence: When global negation is present, the compositional meaning demands that one ignore all head injuries, contrary to world knowledge. When adjectival negation is present, the scale evoked by the *too Y to Z* phrase becomes nonsensical, as more trivial head injuries are claimed to be more worthy of attention. The double negation condition, which exhibits both types of incoherence, corresponds to the classical depth charge configuration.

The experimental items were mixed with 64 unrelated fillers. 36 of the fillers contained the negative polarity element *jemals*, 'ever', along with either a structurally accessible licensor, a structurally inaccessible licensor or no licensor, similar to the design of Drenhaus et al. (2005). We chose these specific items as our fillers in hopes that the ungrammatical conditions may jump out at participants and serve to distract them from the actual purpose of the experiment,

and to provide another sentence type whose acceptability rests on the correct use of negation, as otherwise the experimental sentences may have been too noticeable among the fillers.

Procedure The experiment was run on a PC using the Linger software (Rohde, 2003). The stimulus sentences were rotated through the conditions according to a Latin-square procedure. Presentation order was randomized at runtime. Each trial started with the presentation of the sentence in the center of the screen. Participants were instructed to press the space bar once they had finished reading the sentence in order to proceed to a rating task. Here, they indicated on a scale from 1 to 7 whether the sentence had "made clear sense and contained no grammatical errors" (1 = incomprehensible or contains error, 7 = very clear, no errors). We chose a seven-point scale in order to give subjects the opportunity to choose the "middle ground" (4), and also to allow for enough gradation within the "upper" and "lower" parts of the scale. Both the time from the initial sentence presentation to the pressing of the space bar and the time taken to assign the rating were recorded.

Data analysis Bayesian linear mixed-effects models with full variance-covariance matrices for the random effects (Schielzeth & Forstmeier, 2008; Barr et al., 2013) were fitted to whole-sentence reading times and rating times using the brms package (Bürkner, 2017), which provides a front-end for Stan (Stan Development Team, 2018) in R (R Core Team, 2018). A shifted lognormal distribution was assumed as the generating distribution for these dependent variables. The lognormal distribution was chosen after applying the Box-Cox procedure (Box & Cox, 1964) using the boxcox function from the MASS package (Venables & Ripley, 2002), which suggested a λ value of zero. Given that the amount of time it takes participants to press the response key is included in each measurement, a shifted distribution provides a more accurate model of the generative process behind the data than an unshifted one (Rouder, 2005). Trials with rating times below 150 ms and above 10 s were removed prior to the analysis. The sensibleness ratings were analyzed by fitting a fully hierarchical cumulative logit model with non-equidistant cut-points in brms. While many researchers fit metric models to ordinal data such as data collected using Likert scales, such models often yield suboptimal fits to the actual distribution of ratings, and can lead to inflated rates of Type I and Type II errors (Liddell & Kruschke, 2018).

Across models, the factors global negation and adjectival negation were sum-coded, with presence of negation coded as 1 and absence of negation coded as -1, the interaction term being the product of the respective values. Along with the posterior means, we report the 95% credible interval (percentile-based) of each parameter estimate (back-transformed to the original measurement scales) and treat effects as reliable if 95% of the posterior probability are either above or below zero. We also report as $\hat{\Delta}$ the estimated difference between conditions, back-transformed to the original scale, along with its 95% credible interval.

Across all experiments, we use priors that serve to mildly restrict the possible values for each parameter, but nevertheless allow for considerable variation should the data support large differences between conditions. We set Normal(0,5) priors across all fixed-effect parameters for the (shifted) lognormal models and Normal(0,2) priors for the fixed-effect parameters of the cumulative logit model.[6] For the correlation matrices, we used LKJ priors (Lewandowski et al., 2009) with the ν parameter set to 2; with this setting, higher correlation values are treated as being *a priori* less likely than lower ones, without the prior being overly restrictive.

[6] On the original scales, these priors assume a low but non-zero likelihood of obtaining reading-time differences of up to 44 seconds and rating differences across the entire rating scale, respectively, ignoring intercepts).

Four sampling chains with 2000 iterations each were run for each model. The first 1000 samples were discarded as warmup. $\hat{R}$ values close to 1 were used to monitor for any cases of non-convergence (Gelman & Rubin, 1992). The model function calls, along with the complete fixed-effects output, can be found in Appendix A. The appendices, experimental data and analysis code for all experiments are available at `https://osf.io/rb748`.

2.2 Predictions

The no negation condition (1a) is free of both global and adjectival negation, so that a sensible, compositionally-derived meaning is available.[7] In the adjectival negation condition (1b), the scalar relationship between the adjective and the verb is incongruous (*more un-dangerous (trivial)* → *less reason to ignore*). Anecdotally, sentences in this condition are most easily recognized as being nonsensical. In the global negation condition (1c), world knowledge is contradicted (*Ignore all head injuries*), but the scalar relationship between the adjective and the verb is congruent (*more dangerous* → *less reason to ignore*). Finally, the double negation condition (1d) is the classic depth charge configuration for which meaning inversion should occur: Global negation as well as the negative prefix on the adjective are present and a "negative" lexical verb appears in the complement of the copula. The compositional meaning of the sentence implies that all head injuries should be ignored, which contradicts world knowledge, and the evoked scale is also incongruent. When inversion occurs, sentences of this type are interpreted similarly to sentences in condition (1a) (*Some/all head injuries should be treated*).

Given that multiple negation is known to be problematic for readers, processing time should increase as more negations are added – assuming that processing is compositional – which would predict the highest reading times in the double negation condition and the shortest reading times in the no negation condition, with the two remaining conditions in between. However, if overloading occurs in the double negation condition, as predicted by the account of Wason & Reich (1979), reading times may instead be similar to or even lower than in the global negation and adjectival negation conditions, as readers are assumed to abort compositional interpretation.

The ambiguity hypothesis predicts that readers will notice that they have encountered an instance of the *No X is too Y to Z* construction in the double negation condition and abort the computation of any compositional semantics. They would then interpret the construction according to the most plausible intended meaning, which according to Cook & Stevenson (2010) would be indicated by the "negative" polarity of the verb.[8]Assuming that accessing the stored construction is cognitively less effortful than computing the compositional meaning, processing

[7]We opted for *some (a) X* as the non-negated counterpart of *no X* because *some* involves simple existential as opposed to negative existential quantification (*Some head injuries are X* → ∃y [head_injury(y) ∧ X(y)]). Note that *some* also carries an implicature that more than one head injury has the property X. This is unfortunate with regard to the aim of using minimal pairs across conditions. However, the conceivable alternatives (*All head injuries are X*, *A head injury is X*) would introduce different, and possibly worse, problems: *all* is not a minimal counterpart to *no*, given that *no* is the negation of *at least one*, not of *all*. The indefinite article *a*, on the other hand, makes the sentence infelicitous out of the blue, as it implies the existence of a specific head injury that has not been previously mentioned. Given these caveats, we opted for the variant that produces a minimal pair at least with regard to formal semantics.

[8]It should be noted that the model of Cook & Stevenson (2010) was developed for computational natural language processing and not intended as a model of human behavior. However, as the authors theorize about the role of the *No X is too Y to Z* construction in the grammar, we believe that it is fair to test the model's predictions with human participants.

times are not expected to be longer in the double negation condition than in the global and adjectival negation conditions. The predictions of the ambiguity hypothesis are thus in agreement with those of the overloading hypothesis for this dependent measure.

Under the overloading hypothesis, it is possible that the double negation condition will show shorter reading times but longer rating times compared to the global negation and adjectival negation conditions: Failures to compute a compositional interpretation could lead participants to abort reading, press the space bar and then compute a non-compositional interpretation during the rating time. However, we assume that the current task will not allow a neat division into "interpretive" and "post-interpretive" processing (Caplan & Waters, 1999). Rather, we assume that readers will start their deliberation while reading the sentence. If there is processing spillover from the reading process into the rating process, rating times should pattern analogously to reading times.

We purposely avoided incorporating a task that directly probes the final interpretation due to the concern that it might bias participants to do deeper semantic processing than they would do in a more naturalistic setting. Our task shows indirect evidence for meaning reversal if sensibleness ratings are higher in the double negation condition than in the two conditions with one negation each, even though its compositional interpretation combines the problematic elements (world knowledge violation, absurd scale) of the other two. As the no negation condition is also expected to receive high ratings of sensibleness under all accounts being investigated, there should be a crossover interaction between the experimental factors if meaning inversion occurs.

2.3 Results

Figure 1 shows arithmetic condition means for whole-sentence reading times and rating times; Figure 2 shows the distribution of sensibleness ratings by condition.

Reading times Reading times were increased in the conditions with global negation compared to the conditions without global negation ($\hat{\Delta} = 974$ ms, CrI: [405 ms, 1566 ms], $\Pr(\beta > 0) \approx 1$), as well as in the conditions with adjectival negation compared to those without adjectival negation ($\hat{\Delta} = 773$ ms, CrI: [295 ms, 1259 ms], $\Pr(\beta > 0) \approx 1$). There was also an interaction ($\hat{\Delta} = -731$ ms, CrI: [−1260 ms, −201 ms], $\Pr(\beta > 0) = 0.01$), revealed by nested contrasts to be driven by higher reading times due to adjectival negation when global negation was absent ($\hat{\Delta} = 1515$ ms, CrI: [766 ms, 2301 ms], $\Pr(\beta > 0) \approx 1$).

Rating times Rating times showed an interaction between the negation types ($\hat{\Delta} = -109$ ms, CrI: [−225 ms, 5 ms], $\Pr(\beta > 0) = 0.03$), driven by longer rating times in the presence of adjectival negation when global negation was absent ($\hat{\Delta} = 175$ ms, CrI: [30 ms, 323 ms], $\Pr(\beta > 0) = 0.99$).

Sensibleness ratings Global negation had a negative effect on ratings ($\hat{\Delta} = -0.6$, CrI: [−1.2, 0.01], $\Pr(\beta > 0) = 0.03$), as did adjectival negation ($\hat{\Delta} = -1.49$, CrI: [−2.26, −0.68], $\Pr(\beta > 0) \approx 0$). There was also an interaction ($\hat{\Delta} = 3.08$, CrI: [2.12, 3.84], $\Pr(\beta > 0) \approx 1$), which nested contrasts revealed to be due to a negative effect of adjectival negation in the absence of global negation ($\hat{\Delta} = -3.95$, CrI: [−5.2, −2.72], $\Pr(\beta > 0) \approx 0$), but a positive effect in its presence ($\hat{\Delta} = 1.76$, CrI: [0.88, 2.62], $\Pr(\beta > 0) \approx 1$).

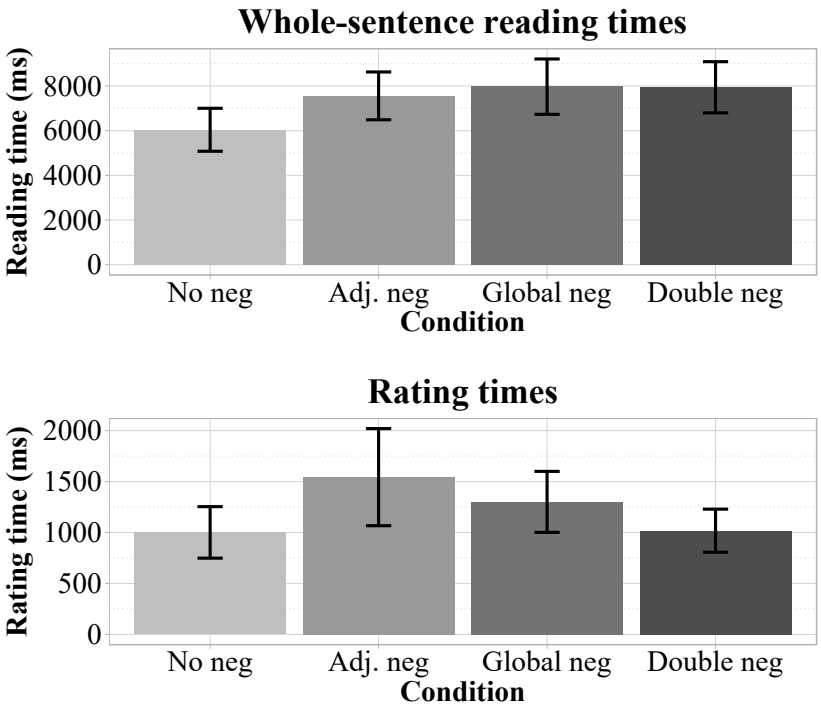


Figure 1: **Experiment 1 – Condition means for whole-sentence reading times and rating times.** Error bars show 95% confidence intervals.

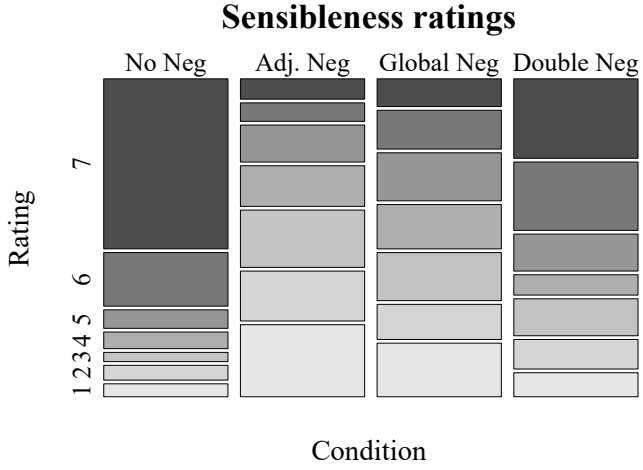


Figure 2: **Experiment 1 – Distribution of sensibleness ratings by condition (1 = not sensible, 7 = perfectly sensible).**

2.4 Discussion

In Experiment 1, increasing the number of negations did not lead to a monotonic increase in processing difficulty: Reading times in the double negation (that is, depth charge) condition did not differ from those in the conditions with only global negation or only adjectival negation. We speculate that participants ran up against a capacity or motivation limit even in the latter two

conditions and aborted further deliberation. In principle, such a scenario would be compatible with the overloading account. In the absence of global negation, rating times increased in the presence of adjectival negation, which disrupts the internal consistency of the evoked scale. The absence of this effect when global negation is present suggests a partial masking of the semantic incongruity. Finally, the pattern of ratings indicates that meaning inversion occurred: Despite being the semantically and pragmatically most anomalous condition under a compositional interpretation, the double negation condition received the second-highest sensibleness ratings among the four conditions, second only to the no negation condition.

The ambiguity hypothesis offers an entirely different account for the observed pattern: As soon as readers notice that they are faced with an instance of the *No X is too Y to Z* construction, they may switch from compositional processing to a more idiomatic analysis.[9] That ratings in the double negation condition are not strongly affected by the observed processing difficulty may be explained by assuming that the difficulty is not due to negation overload but rather reflects participants' reasoning about whether the intended meaning is "negative" or "positive". As soon as the (presumably) intended meaning has been identified, the subjective experience would be one of "success" rather than "failure". Under Cook & Stevenson's (2010) version of the ambiguity hypothesis, the point at which the ultimate interpretation is decided would be the "negative" verb.

One potentially problematic aspect of the results for the ambiguity hypothesis is the fact that the double negation condition received lower ratings than the no negation condition, as can be seen in Figure 2. This pattern is unexpected if instances of the negative *No X is too Y to Z* construction – that is, double negation sentences – are seen as completely sensible and acceptable, and thus no different in their status from no negation sentences. However, proponents of the ambiguity hypothesis could plausibly argue that lower ratings were given in the double negation condition because there is no context that licenses the use of the construction and disambiguates the intended meaning, so that participants may have recognized the construction but sometimes failed to interpret it.

Given that Experiment 1 showed evidence of the depth charge effect occurring in German, and beyond that revealed what could be a "ceiling effect" in processing time due to an inherent capacity limit or strategic time-outs, we turned to eye tracking in order to investigate the on-line processing of depth charge sentences in more detail. For the eye tracking study, we divided the sentence into three regions of interest, which allows inferences as to which part of the sentence is most problematic for readers.

3 Experiment 2A

Our second experiment is concerned with two empirical predictions derived from previous work on the depth charge effect. The first prediction is that any measurable effects of non-compositional processing and eventual inversion of meaning should first become visible at the final verb. The final verb has been claimed to be the main source of the depth charge effect both by the original proponents of the overloading hypothesis (Wason & Reich, 1979), as well as by proponents of the ambiguity hypothesis (Cook & Stevenson, 2010). The second prediction

[9] As Fortuin (2014, p. 279) notes, this by no means necessitates that compositional processing be fully "switched off", but that certain well-formedness constraints may be disregarded.

concerns the possible influence of readers' working memory capacity on the depth charge effect. If meaning inversion is due to linguistic complexity overloading the reader's working memory capacity, individuals with higher capacity should have partial immunity to it, given that they may occasionally be able to process depth charge sentences compositionally and realize that their meaning is incongruous. Meanwhile, the ambiguity hypothesis does not assume overloading and therefore no influence of working memory capacity on the depth charge effect is predicted.

3.1 Method

Participants Sixty-one native speakers of German were recruited from the local student population. Subjects either received credit points or were paid €10 as compensation.

Materials The same materials as in Experiment 1 were used. For the statistical analysis, three regions of interest were defined: the initial noun phrase, the region from the copula to the comma and the final *to*-phrase, as indicated by the diamonds below.

(2) **Global negation present, adjectival negation present**
(DOUBLE NEGATION)

a. Keine Kopfverletzung ◇ ist zu ungefährlich, ◇ um ignoriert zu werden.
No head injury is too un-dangerous to ignored to get

"No head injury is too innocuous to be ignored."

This partitioning allows us to pinpoint the approximate location of the trigger of meaning inversion within the sentence (see predictions below) as well as look at the distribution of reading time across the different regions during possible attempts at reinterpretation.

Procedure Prior to the main experiment, participants completed an operation span test as a measure of working memory capacity, as previously used by Nicenboim et al. (2016) and von der Malsburg & Vasishth (2013), following the recommendations of Conway et al. (2005).[10] As in Nicenboim et al. (2016), single letters as opposed to words were used as recall targets to minimize lexical influences, assuming that working memory is largely a domain-general resource (Kane et al., 2004).

In the main experiment, participants were instructed to read the sentences at their own pace while their eye movements were recorded. We report results for first-pass reading times, regression-path durations (also called go-past times) and total reading times. As in Experiment 1, participants were asked to rate each sentence's sensibleness on a scale from 1–7. A more detailed description of the experimental setup and procedure is given in Appendix B.

3.2 Data analysis

Factor coding, prior specification, sampling and interpretation were carried out analogously to Experiment 1, but working memory capacity as well as all possible interactions with the experimental factors were added to the model. The predictor reflecting working memory capacity, as measured in partial credit units (PCU) (Conway et al., 2005), was centered and scaled prior to

[10]The python script used to carry out the task is available at `https://github.com/tmalsburg/py-span-task`.

being entered into the model, so that the associated parameter estimates reflect the expected effect of increasing working memory capacity by one standard deviation on the PCU scale. PCU scoring assigns partial credit to trials for which one or more items were incorrectly recalled, and does not assign higher weight to trials with higher memory load. A detailed description of the data analysis procedure is given in Appendix B.

3.3 Predictions

The overloading hypothesis predicts that low-capacity participants should run up against their limit sooner than high-capacity participants and thus show shorter processing times across the negation conditions. High-capacity participants may show a difference between the double negation condition and the global negation condition: Their capacity-driven or strategic limit may not be exhausted in the double negation condition, thus allowing further compositional processing, so that more processing difficulty is visible in this group in the double negation condition compared to the global negation condition. Low-capacity readers, on the other hand, are not expected to show such a difference, given that their lower limit should may be exhausted in all of the negation conditions.[11] Given the intuition of Wason & Reich (1979), these effects should occur in the final region of the sentence, where the verb *ignore* appears. Furthermore, if meaning inversion is caused by overloading of working memory, high-capacity participants should show an overall weaker inversion effect, that is, lower ratings in the double negation condition.

The ambiguity hypothesis does not predict any effect of working memory capacity on the interpretation of the depth charge construction. However, as the account of Cook & Stevenson (2010) assumes that the final verb signals the intended semantics of the ambiguous *No X is too Y to Z construction*, any signs of non-compositional processing should first become visible at the final region of the sentence, where the "negative" verb is encountered. Meanwhile, the account of Fortuin (2014) does not make such a prediction, but instead predicts that non-compositional processing could already become visible at *too*, that is, in the pre-final region of the sentence. Fortuin (2014, p. 278f.) claims that the combination of global negation with the element *too* results in a presupposition that there is no excessive degree of trivialness of head injuries, which arguably triggers meaning inversion.

3.4 Results

Figure 3 shows the means of the three eye tracking measures of interest by region. Figure 4 shows mean rating times by condition. Figure 5 shows the distribution of sensibleness ratings by condition.

First-pass reading times

[11] It may seem counterintuitive that high-capacity participants are predicted to show more as opposed to less processing difficulty. However, the prediction follows directly from the assumption that non-compositional processing is faster than compositional processing, and that high-capacity readers are more likely to engage in the latter for a longer time. Such a tendency would also match findings by von der Malsburg & Vasishth (2013) and Nicenboim et al. (2016) suggesting that high-capacity readers experience slowdowns in computationally demanding linguistic environments because they carry out more parsing operations than low-capacity readers.

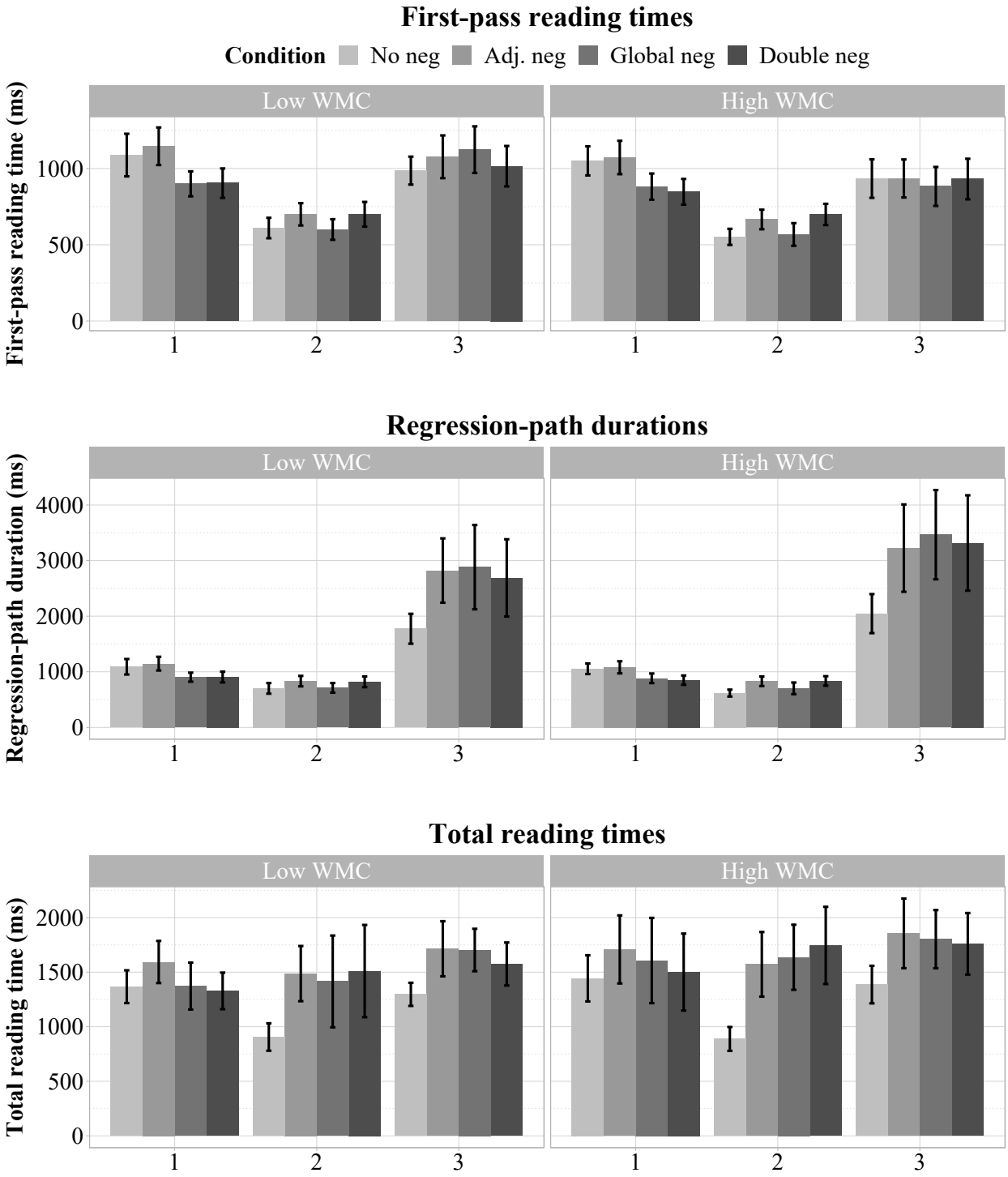


Figure 3: **Experiment 2A – Reading measures by region and condition for low- and high-WMC groups (median split).** Error bars show 95% confidence intervals.

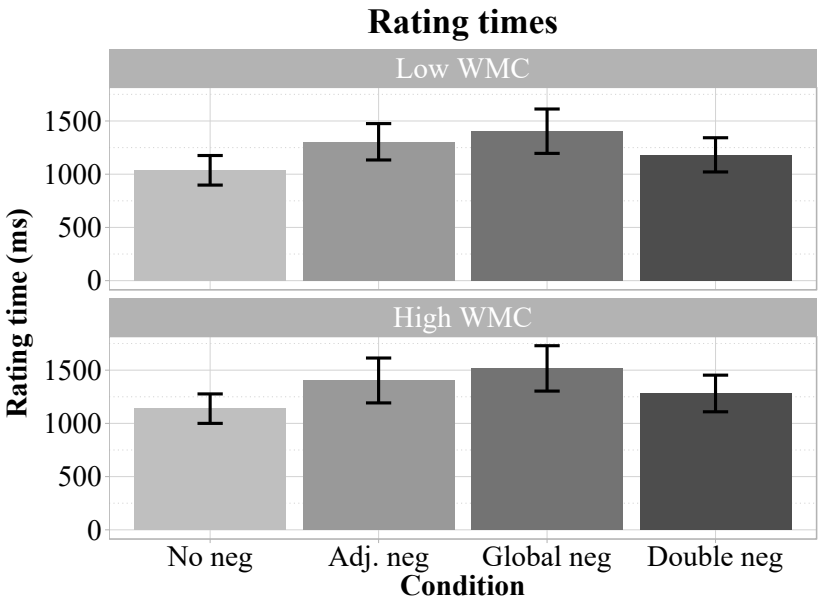


Figure 4: **Experiment 2A – Condition means for rating times for low- and high-WMC groups (median split).** Error bars show 95% confidence intervals.

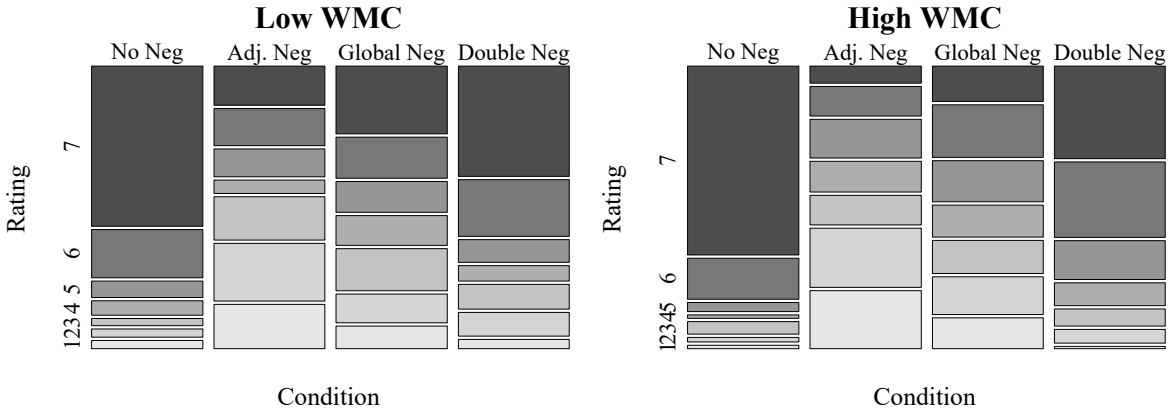


Figure 5: **Experiment 2A – Distribution of sensibleness ratings by condition for low- and high-WMC groups (median split; 1 = not sensible, 7 = perfectly sensible).**

Region 1 (*No head injury*) There was an effect of global negation on first-pass reading times in region 1, such that globally negated sentences were read faster than non-globally negated sentences ($\hat{\Delta} = -199$ ms, CrI: $[-249$ ms, -152 ms$]$, $\Pr(\beta > 0) \approx 0$).

Region 2 (*is too (un)dangerous*) In region 2, there was an effect of adjectival negation, such that sentences with negated adjectives had longer reading times than sentences with non-negated adjectives ($\hat{\Delta} = 95$ ms, CrI: $[50$ ms, 139 ms$]$, $\Pr(\beta > 0) \approx 1$). There was also an interaction between adjectival negation and working memory capacity ($\hat{\Delta} = 26$ ms, CrI: $[1$ ms, 51 ms$]$, $\Pr(\beta > 0) = 0.98$), such that high-capacity participants spent more time reading negated adjectives.

Region 3 (*to be ignored*) No effects in evidence.

Regression-path durations

Region 1 (*No head injury*) See first-pass reading times.

Region 2 (*is too (un)dangerous*) Regression-path durations at region 2 were longer for sentences with adjectival negation ($\hat{\Delta} = 138\,\text{ms}$, CrI: [87 ms, 189 ms], $\Pr(\beta > 0) \approx 1$) , and there was again an interaction with working memory capacity ($\hat{\Delta} = 29\,\text{ms}$, CrI: [5 ms, 54 ms], $\Pr(\beta > 0) = 0.99$), such that regression-path durations were longer for high-capacity participants in the presence of a negated adjective.

Region 3 (*to be ignored*) In region 3, regression-path durations were longer in the presence of global negation ($\hat{\Delta} = 416\,\text{ms}$, CrI: [220 ms, 623 ms], $\Pr(\beta > 0) \approx 1$) and adjectival negation ($\hat{\Delta} = 359\,\text{ms}$, CrI: [249 ms, 470 ms], $\Pr(\beta > 0) \approx 1$) . Regression paths in this region were longer for participants with high working memory capacity ($\hat{\Delta} = 213\,\text{ms}$, CrI: [−32 ms, 487 ms], $\Pr(\beta > 0) = 0.96$). There was also an interaction between global negation and adjectival negation $\hat{\Delta} = -475\,\text{ms}$, CrI: [−630 ms, −324 ms], $\Pr(\beta > 0) \approx 0$), driven by a slowdown due to adjectival negation in the absence of global negation only ($\hat{\Delta} = 835\,\text{ms}$, CrI: [637 ms, 1035 ms], $\Pr(\beta > 0) \approx 1$).

Total reading times

Region 1 (*No head injury*) Global negation led to a decrease in total reading times in region 1 ($\hat{\Delta} = -129\,\text{ms}$, CrI: [−206 ms, −49 ms], $\Pr(\beta > 0) \approx 0$) while adjectival negation led to an increase ($\hat{\Delta} = 69\,\text{ms}$, CrI: [22 ms, 114 ms], $\Pr(\beta > 0) \approx 1$). An interaction between the two negation types was also present ($\hat{\Delta} = -112\,\text{ms}$, CrI: [−166 ms, −57 ms], $\Pr(\beta > 0) \approx 0$), due to adjectival negation leading to longer reading times only in the absence of global negation ($\hat{\Delta} = 178\,\text{ms}$, CrI: [108 ms, 248 ms], $\Pr(\beta > 0) \approx 1$).

Region 2 (*is too (un)dangerous*) In region 2, total reading times were longer in the presence of global negation ($\hat{\Delta} = 231\,\text{ms}$, CrI: [142 ms, 321 ms], $\Pr(\beta > 0) \approx 1$) and also in the presence of adjectival negation ($\hat{\Delta} = 314\,\text{ms}$, CrI: [216 ms, 413 ms], $\Pr(\beta > 0) \approx 1$). An interaction between global and adjectival negation was also present ($\hat{\Delta} = -204\,\text{ms}$, CrI: [−276 ms, −133 ms], $\Pr(\beta > 0) \approx 0$), such that the slowdown due to adjectival negation in the absence of global negation ($\hat{\Delta} = 518\,\text{ms}$, CrI: [401 ms, 638 ms], $\Pr(\beta > 0) \approx 1$) was larger than in the presence of global negation ($\hat{\Delta} = 110\,\text{ms}$, CrI: [−19 ms, 233 ms], $\Pr(\beta > 0) = 0.96$). Finally, there was an interaction between global negation and working memory capacity ($\hat{\Delta} = 59\,\text{ms}$, CrI: [−10 ms, 130 ms], $\Pr(\beta > 0) = 0.95$), such that high-capacity readers had longer reading times in the globally negated conditions.

Region 3 (*to be ignored*) In region 3, total reading times were longer in the presence of global negation ($\hat{\Delta} = 136\,\text{ms}$, CrI: [47 ms, 225 ms], $\Pr(\beta > 0) \approx 1$) as well as in the presence of adjectival negation ($\hat{\Delta} = 132\,\text{ms}$, CrI: [75 ms, 190 ms], $\Pr(\beta > 0) \approx 1$). There was also an interaction between global and adjectival negation ($\hat{\Delta} = -203\,\text{ms}$, CrI: [−278 ms, −130 ms], $\Pr(\beta > 0) \approx 0$), due to adjectival negation causing a slowdown in the absence of global negation

($\hat{\Delta} = 335$ ms, CrI: [232 ms, 439 ms], $\Pr(\beta > 0) \approx 1$) but a speedup in its presence ($\hat{\Delta} = -71$ ms, CrI: [−155 ms, 11 ms], $\Pr(\beta > 0) = 0.04$). Finally, high-capacity readers had longer total reading times in this region ($\hat{\Delta} = 94$ ms, CrI: [−17 ms, 221 ms], $\Pr(\beta > 0) = 0.95$).

Rating times Rating times were elevated in the presence of global negation ($\hat{\Delta} = 87$ ms, CrI: [33 ms, 141 ms], $\Pr(\beta > 0) \approx 1$). There was also an interaction with adjectival negation ($\hat{\Delta} = -166$ ms, CrI: [−235 ms, −96 ms], $\Pr(\beta > 0) \approx 0$), revealed by nested contrasts to be due to faster rating times due to adjectival negation in the presence of global negation ($\hat{\Delta} = -149$ ms, CrI: [−235 ms, −68 ms], $\Pr(\beta > 0) \approx 0$) but slower rating times in its absence ($\hat{\Delta} = 188$ ms, CrI: [104 ms, 271 ms], $\Pr(\beta > 0) \approx 1$).

Sensibleness ratings In sensibleness ratings, there was an effect of adjectival negation ($\hat{\Delta} = -1.34$, CrI: [−1.65, −1.01], $\Pr(\beta > 0) \approx 0$) as well as a two-way interaction with global negation ($\hat{\Delta} = 2.94$, CrI: [2.28, 3.55], $\Pr(\beta > 0) \approx 1$) and a three-way interaction with global negation and working memory capacity ($\hat{\Delta} = 0.44$, CrI: [0.08, 0.84], $\Pr(\beta > 0) = 0.99$). The two-way interaction was due to adjectival negation leading to lower ratings in the absence of global negation ($\hat{\Delta} = -4.19$, CrI: [−4.71, −3.52], $\Pr(\beta > 0) \approx 0$) but to higher ratings in its presence ($\hat{\Delta} = 1.41$, CrI: [0.91, 1.91], $\Pr(\beta > 0) \approx 1$), while the three-way interaction was due to both of the aforementioned differences being larger for participants with high working memory capacity (see Figure 5).

3.5 Discussion

The main findings of Experiment 1 were replicated in Experiment 2A: Comparable amounts of processing difficulty were observed in regression-path durations in the final region of the sentence across the negated conditions, with the no negation condition being easier to process, along with a crossover interaction in ratings signaling the occurrence of the depth charge effect. As opposed to Experiment 1, there was a crossover interaction in rating times as well, such that the no negation and double negation conditions showed shorter mean rating times compared to the adjectival negation and global negation conditions. In the eye tracking measures, we found evidence of interactions in regression-path durations at the final region and in total reading times across the whole sentence. The data indicate a partial suppression of the expected slowdown due to adjectival negation when global negation is present, which suggests that readers omit some steps necessary for the compositional interpretation of the double negation – that is, depth charge – sentences. With regard to the locus of the effect, the critical interaction between negation types first appeared in regression-path durations in region 3, which contained the verb, consistent with the predictions of Wason & Reich (1979) and Cook & Stevenson (2010).

We found no evidence that high-capacity participants are more resistant to the depth charge effect. While they spent more time in region 2 on the first pass compared to low-capacity readers when it contained a negated adjective, and showed overall increased regression-path durations in the final region, there was no three-way interaction of the shape predicted by the overloading hypothesis of Wason & Reich (1979) in any measure. If anything, high-capacity readers even showed a tendency toward higher ratings in the double negation condition compared to low-capacity readers, which is not compatible with partial immunity to meaning inversion. However, apart from the statistical results for the three-way interaction in the eye tracking measures being inconclusive, it may be that even high-capacity readers already reach their limit in the single negation conditions, so that there is no visible increase in processing difficulty in the

double negation condition. While such an explanation of the null result is possible, it is made somewhat unlikely by the fact that the single negation sentences received lower ratings than the double negation sentences: If capacity overload occurs in all the negation conditions, it would appear to have different effects on the final rating across conditions.

Again, the findings are compatible with the ambiguity hypothesis, which does not predict an effect of working memory capacity on the probability of non-compositional processing, and thus on meaning inversion. The fact that the critical interaction first became visible in the sentence-final region is also consistent with the claim of Cook & Stevenson (2010) that it is the verb that signals the semantics of the ambiguous *No X is too Y to Z* construction. Meanwhile, the data yield no evidence in favor of the claim of Fortuin (2014) that meaning inversion is already triggered at *too*.

In order to see whether world knowledge further contributes to meaning inversion, as assumed by Wason & Reich (1979) and Fortuin (2014), we conducted a follow-up experiment. Experiment 2B tests whether there is a direct connection between world knowledge and the strength of the depth charge effect.

4 Experiment 2B

Experiment 2B is an ancillary study to Experiment 2A and aims to investigate the influence of participants' world knowledge on the depth charge effect. Experiment 2A yielded no reliable evidence in favor of memory overload. It is possible that rather than running out of working memory, readers run up against a motivational limit to compositional interpretation in depth charge sentences. They may then turn the contents of their mental buffer into a "bag of words" and combine the contents in a way that appears plausible according to their world knowledge. An influence of world knowledge on meaning inversion had already been hypothesized by Wason & Reich (1979) and was investigated in previous experimental work, with mixed results (Natsopoulos, 1985; O'Connor, 2015). Here, we are specifically interested in whether world knowledge can be shown to affect the on-line processing of depth charge sentences, as would be assumed under the overloading hypothesis as well as under one version of the ambiguity hypothesis (Fortuin, 2014). Under both accounts, depth charge sentences should become easier to process and receive higher ratings if the inverted reading is consistent with world knowledge.

There is an influential stream in language processing research which claims that under certain conditions, comprehenders make use of "fast and frugal" heuristics or a low-level "pseudo-grammar" to derive sentence meaning, as opposed to computing syntax and meaning compositionally (e.g. Townsend & Bever, 2001; Ferreira et al., 2002; Sanford & Sturt, 2002; Ferreira, 2003; Dwivedi, 2013; Karimi & Ferreira, 2016; Christianson, 2016). Such "good enough" representations are arguably more likely to be adopted when compositional processing difficulty is high and/or task demands do not require detailed representations of sentence meaning (e.g. Swets et al., 2008). It has also been argued that factors such as the real-world plausibility of events may be recruited during the creation of "good enough" meaning representations. For instance, Ferreira (2003) reports that implausible passive sentences (*The dog was bitten by the man*) are often misinterpreted, and attributes the finding to the use of a frequency-based heuristic that assigns the agent role to the first noun phrase in the sentence, in addition to the use of general world knowledge. Although the classic overloading account of the depth charge effect would predict that heuristics are used only after compositional processing has failed, some variants

of the "good enough" approach assume that heuristics are used first, and that compositional processing is used as a second step to check the derived interpretation. Under this account, when readers reach their limit, they would abort the compositional checking procedure and adopt the heuristic meaning that is already in place.

Meanwhile, the particular version of the ambiguity hypothesis put forward by Cook & Stevenson (2010) explicitly does not consider the pragmatic dimension of depth charge sentences, but limits itself to lexical semantic features of the component words, thus not predicting effects of world knowledge on interpretation. In their corpus study, the model of Cook & Stevenson was reportedly able to correctly identify the intended meaning of 170 depth charge sentences in 88% of cases without using information beyond the lexical semantic level. On the other hand, Fortuin (2014, p. 276) claims that "language users [...] use [...] their general knowledge of their language (and their general background knowledge) to process and make sense of a particular instance [of the depth charge construction]", which would be consistent with world knowledge affecting interpretation.

We are interested not only in an effect of world knowledge on sensibleness ratings, but also on on-line processing, that is, on the eye tracking measures collected in Experiment 2A. Here, we limited our focus to regression-path durations in the sentence-final region, because this was the region where the numerically largest effect occurred.

In order to get an approximate measure of an average person's world knowledge about the content of the stimulus sentences, we had a new set of participants indicate how strongly they agreed with the sensible – that is, negation-free – version of the stimuli from Experiment 2A (*Some head injuries are too dangerous to be ignored*). We assume that participants only form strong opinions about topics which they subjectively feel they know a lot about. The rationale behind choosing the no negation version was that under meaning reversal, the double negation condition is assigned a meaning that is close to that of the no negation sentence (*Treat even seemingly trivial head injuries*). Thus, if the double negation sentence is transformed to have approximately the same meaning as the no negation sentence, and if approval of the proposition expressed by the no negation sentence is high, participants should be more convinced that they have interpreted the double negation sentence correctly.

4.1 Method

Participants Thirty-five native speakers of German were recruited through social media. They did not receive any compensation for their participation.

Materials The no negation versions of the 32 sentences from the previous experiments – for instance, *Some head injuries are too dangerous to be ignored* – were used as stimuli. There were also 32 fillers, which consisted mainly of political statements and philosophical quotes.

Procedure The experiment was run on-line on Ibex farm (Drummond, 2018). Presentation order was randomized at runtime. Participants were instructed to read the sentences at their own pace and to indicate on a scale from 1 to 5 whether they agreed with the statement (1 = do not agree at all, 5 = agree completely) and, also on a scale from 1 to 5, how easy they found the sentence to understand (1 = impossible to understand, 5 = easy to understand). We chose a five-point scale for this experiment, as opposed to the seven-point scale used in the previous

rating studies, mainly because "ease of understanding" is likely somewhat difficult to evaluate introspectively, and more possible rating categories may have made the task more difficult.

Data analysis As comprehensibility was not of primary interest in the reanalysis of the data from Experiment 2A, approval was residualized against comprehensibility in hopes of getting a clearer estimate of participants' world knowledge. Residuals were extracted from a linear mixed-effects model fitted to approval with comprehensibility as a fixed effect, as well as random intercepts and random slopes by participant, in lme4 (Bates et al., 2015). The resulting measure was. Mean residual approval was computed for each experimental item, and the scores were centered and scaled before being entered into the analysis as a continuous predictor, such that parameter estimates from the model indicate the effect of increasing approval by one standard deviation.

We reanalyzed regression-path durations in region 3 (*to be ignored*), rating times and sensibleness ratings from Experiment 2A by fitting maximal models to each measure analogously to the previous experiments. As we were primarily interested in the relationship between the double negation – that is, depth charge – condition and the no negation condition, we dropped the global negation condition from the analysis. The adjectival negation condition, meanwhile, was kept as a control (see predictions below). Two treatment contrasts (one per negation condition) were defined that used the no negation condition as the baseline. All two-way interactions between approval and condition were entered into the model. Working memory was also entered into the model, along with its two-way interactions with condition.

A more detailed description of the data analysis procedure is given in Appendix B.

4.2 Predictions

The overloading account plausibly predicts an effect of world knowledge on the adopted non-compositional meaning, given that "good enough" processing has been argued to rely, inter alia, on real-world plausibility (e.g. Ferreira, 2003; Dwivedi, 2013). While Cook & Stevenson's (2010) version of the ambiguity hypothesis explicitly predicts no influence of world knowledge on meaning inversion, as lexical semantic information should be sufficient, the account of Fortuin (2014) does assume that world knowledge is recruited during the processing depth charge sentences.

Assuming that world knowledge is recruited, sensibleness ratings in Experiment 2A should be higher in the no negation and double negation conditions for items whose no negation version received higher approval ratings in Experiment 2B, as the inverted meaning of the double negation sentence is similar to that of the no negation sentence. Compared to the no negation condition as the baseline, there should be a larger negative interaction with approval for the adjectival negation sentences compared to the double negation sentences: As the meaning of the double negation sentence is closer to that of the no negation sentence under inversion, approval of the no negation sentence should have more of a positive effect on ratings for the double negation than for the adjectival negation sentence. By the same logic, the effect of approval in the adjectival negation condition should be negative, given that its meaning is nonsensical, especially when compared to a sensible belief about the way that things should normally be.

Under both the overloading account of Wason & Reich (1979) and the ambiguity hypothesis of Fortuin (2014), rating times and regression-path durations in region 3 of Experiment 2A should show a pattern that mirrors the effect on ratings, that is, both measures should show facilitation in the no negation and double negation conditions for items with high approval ratings. If readers compute the inverted interpretation while reading the sentence, this should be easier when it matches their world knowledge. Given this assumption, processing difficulty should increase along with approval in the adjectival negation condition, given the clash between sentence content and world knowledge. Meanwhile, the account of Cook & Stevenson (2010) does not predict any effects of world knowledge on the on-line measures.

4.3 Results

Figures 6 and 7 show regression-path durations in region 3 (*to be ignored*), rating times and rating distributions for low- and high-approval items (median split).

Regression-path durations in region 3 (*to be ignored*) In the baseline no negation condition, regression-path durations were shorter for items with high approval ($\hat{\Delta} = -133$ ms, CrI: $[-224$ ms, -34 ms$]$, $\Pr(\beta > 0) \approx 0$). There was also an effect of working memory capacity, such that participants with high working memory capacity showed longer regression paths ($\hat{\Delta} = 152$ ms, CrI: $[1$ ms, 321 ms$]$, $\Pr(\beta > 0) = 0.98$). As in the original analysis, regression-path durations were longer compared to the baseline for both double negation ($\hat{\Delta} = 694$ ms, CrI: $[452$ ms, 962 ms$]$, $\Pr(\beta > 0) \approx 1$) and adjectival negation sentences ($\hat{\Delta} = 757$ ms, CrI: $[559$ ms, 975 ms$]$, $\Pr(\beta > 0) \approx 1$).

Rating times Rating times were increased compared to the baseline for the double negation condition ($\hat{\Delta} = 98$ ms, CrI: $[29$ ms, 168 ms$]$, $\Pr(\beta > 0) \approx 1$) as well as for the adjectival negation condition ($\hat{\Delta} = 177$ ms, CrI: $[97$ ms, 268 ms$]$, $\Pr(\beta > 0) \approx 1$).

Sensibleness ratings In sensibleness ratings, there was a positive effect of approval in the baseline no negation condition ($\hat{\Delta} = 0.21$, CrI: $[0.05, 0.35]$, $\Pr(\beta > 0) \approx 1$). Compared to the baseline, ratings were lower in the double negation condition ($\hat{\Delta} = -0.83$, CrI: $[-1.15, -0.4]$, $\Pr(\beta > 0) \approx 0$) as well as in the adjectival negation condition ($\hat{\Delta} = -3.86$, CrI: $[-4.55, -2.85]$, $\Pr(\beta > 0) \approx 0$). For the latter, there was also an interaction with approval ($\hat{\Delta} = -0.5$, CrI: $[-1.12, -0.12]$, $\Pr(\beta > 0) = 0.01$), due to the effect of approval showing weak evidence of being negative in the adjectival negation condition ($\hat{\Delta} = -0.15$, CrI: $[-0.41, 0.12]$, $\Pr(\beta > 0) = 0.13$). Lastly, the two-way interaction between adjectival negation and working memory was reliable ($\hat{\Delta} = -0.42$, CrI: $[-0.9, -0.09]$, $\Pr(\beta > 0) \approx 0$), reflecting a tendency of high-capacity readers to give more extreme ratings.

4.4 Discussion

If world knowledge contributes to the depth charge effect, stronger world knowledge should make the double negation easier to process and increase the strength of the depth charge effect. This prediction was partly borne out in the data, consistent with both the overloading hypothesis and the ambiguity hypothesis as proposed by Fortuin (2014). The results suggest that human readers *do* use pragmatic information to resolve the meaning of depth charge sentences, calling

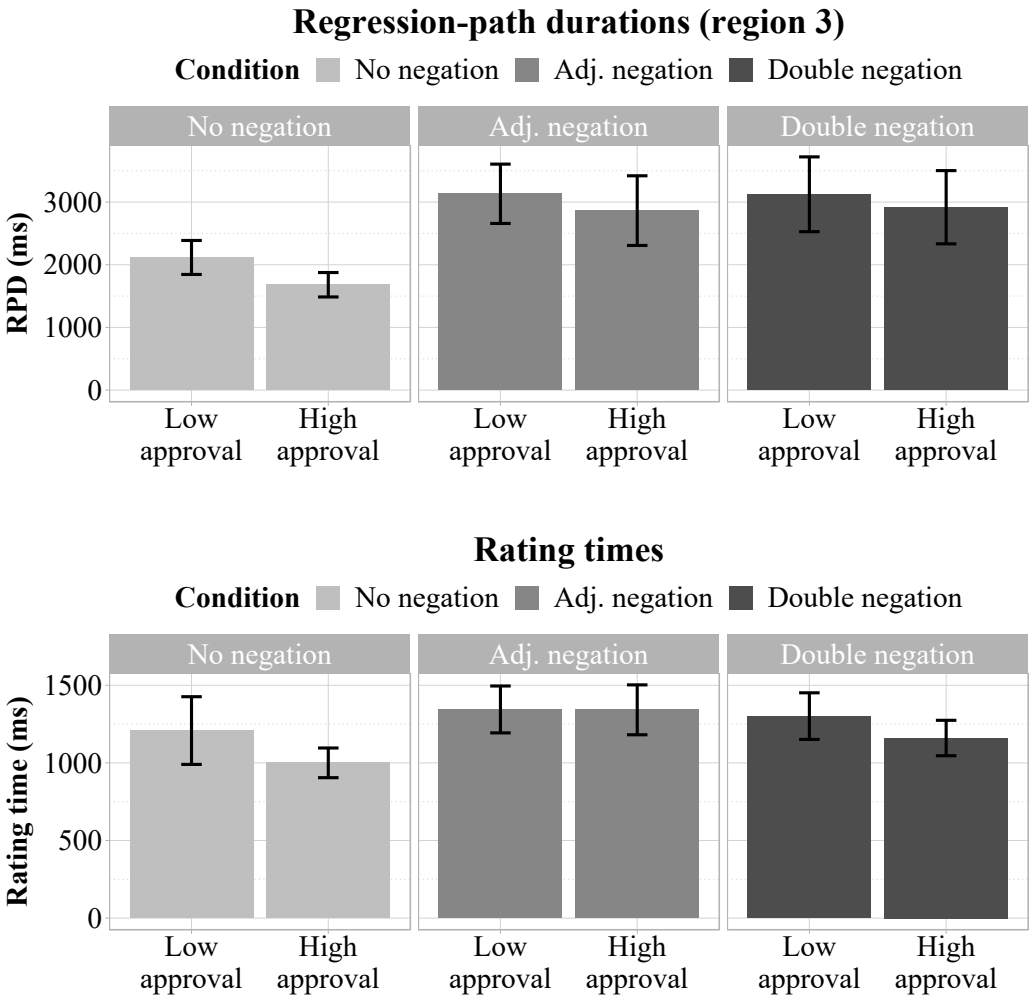


Figure 6: **Experiment 2A – Regression-path durations in region 3 (*to be ignored*) and rating times for high- and low-approval items (median split).** Error bars show 95% confidence intervals.

into question the conclusion of Cook & Stevenson (2010) that lexical information from the verb is sufficient. Ratings of sensibleness showed an effect of approval ratings from Experiment 2B on sensibleness ratings from Experiment 2A, such that higher sensibleness ratings were given in the double negation condition for high-approval items, just like in the baseline no negation condition. There was also evidence that the adjectival negation condition received lower ratings than the baseline for high-approval items.

High-approval items also showed shorter regression paths from the region containing the lexical verb, suggesting that world knowledge reduced processing difficulty, but there was no interaction with condition. The absence of such an interaction is somewhat unexpected, given that the semantics of the adjectival negation condition show a transparent mismatch with the semantics of the no negation condition, which our measure of world knowledge was based on. There is, however, weak evidence for both interaction terms for the non-baseline conditions to be positive. There is thus evidence that world knowledge does play a role in determining the final

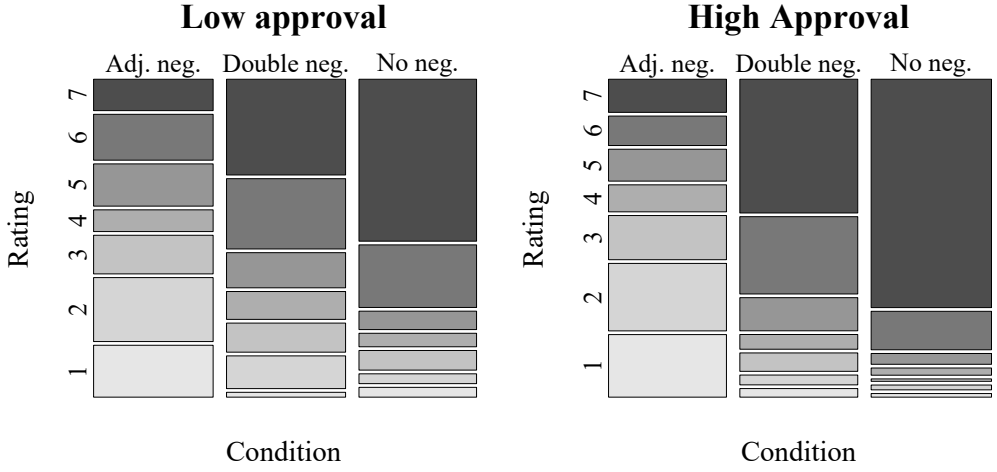


Figure 7: **Experiment 2A – Distribution of sensibleness ratings by approval (median split) and condition (1 = not sensible, 7 = perfectly sensible).**

interpretation of depth charge sentences, as had been hypothesized in earlier work (Wason & Reich, 1979; Natsopoulos, 1985; Fortuin, 2014; O'Connor, 2015; Kizach et al., 2015).

5 Experiment 3

Our next experiment investigates the central point of disagreement between the overloading hypothesis and the ambiguity hypothesis, the question of whether the inverted interpretation of depth charge sentences is the result of one or multiple processing errors or a feature of grammar. The overloading hypothesis predicts that meaning inversion should occur in other environments that are of comparable linguistic complexity to the original depth charge construction, given that participants' processing capacity should be exceeded in the same way. On the other hand, the ambiguity hypothesis would not necessarily predict the effect to generalize to other linguistic environments, unless additional assumptions are made regarding mutual similarity of constructions in the grammar.

If the depth charge effect is intimately tied to the hypothesized *No X is too Y to Z* construction that is stored as a grammatical template – which is one interpretation of the ambiguity hypothesis endorsed by Cook & Stevenson (2010) and Fortuin (2014) – one would not necessarily assume comparable effects in sentences that are closely matched to the construction in terms of compositional semantics. Fortuin notes that there are such constructions, which may or may not license "negative" readings (*No sport is too marginal as to be ignored*, p. 282), but does not offer a formal account of the relationships between constructions, nor of the effects that such relationships have on on-line processing. Indeed, Fortuin states that "one could speak about a typology or perhaps network of constructions, as long as one keeps in mind that the different constructions exist independently of one another, perhaps in different linguistic systems, even though general (i.e. non language dependent) semantic-pragmatic and perhaps cognitive factors may explain their occurrence" (p. 286). Given this qualification, it remains an open question under the ambiguity hypothesis whether the depth charge effect should be thought of as a more general phenomenon. The overloading account, on the other hand, naturally predicts that the

effect should generalize to other constructions that share (some of) the problematic aspects of classic depth charge sentences.

In order to investigate how potential depth charge configurations behave in different linguistic constructions, and thus to find out whether the effect generalizes beyond the classic *No X is too Y to Z* schema, we identified two alternative ways of expressing the same meaning. The first alternative construction keeps the particle *zu*, 'too', but substitutes the *to*-infinitive for a finite clause introduced by *als dass*, 'as that'. Apart from being a different type of potentially conventionalized form-meaning pair, this particular construction contains neither the final *zu*-infinitive nor the passive construction in the infinitival clause, which may lighten the overall processing load, given that passives are known to be troublesome (Ferreira, 2003). The second construction replaces the particle *zu*, 'too' with *so*, 'so', thereby eliminating the implicit negation carried by the former.

5.1 Method

Participants Sixty native speakers of German from the local student population participated in the experiment. They were paid either €7 or received credit points as compensation.

Materials The experiment employed a 2 × 3 design with the factors negation (adjectival negation versus double negation) and construction (*too . . . to* versus *too . . . as that* versus *so . . . that*). An example item in all six conditions is shown in (3). In order to derive the same meaning as in the original sentence, an overt negation appears in the final clause of the *so . . . that* construction. Both alternative constructions were compared against the *zu . . . um* construction as a baseline. To assure balanced presentation of the six conditions, only thirty out of the original 32 items were used. The modal verbs appearing in the *too . . . as that* and *so . . . that* constructions varied between items, and sometimes between conditions of the same item as well. We used either *könnte*, 'could', or *sollte*, 'should', according to our own judgment of which of the two sounded more acceptable in a given context.

(3) **Global negation absent, adjectival negation present**
(ADJECTIVAL NEGATION)

b. Manch eine Kopfverletzung ist zu ungefährlich,
Some a head injury is too un-dangerous

"Some head injuries are too innocuous . . . "

Global negation present, adjectival negation present
(DOUBLE NEGATION)

d. Keine Kopfverletzung ist zu ungefährlich,
No head injury is too un-dangerous

"No head injury is too innocuous . . . "

too . . . to construction

um ignoriert zu werden.
to ignored to get

". . . to be ignored."

too . . . as that construction

als dass man sie ignorieren könnte.
as that one it ignore could

". . . that one could ignore it (/them)."

so . . . that construction

Global negation absent, adjectival negation present
(ADJECTIVAL NEGATION)

b'. Manch eine Kopfverletzung ist so ungefährlich,
Some a head injury is so un-dangerous

"Some head injuries are so innocuous . . . "

Global negation present, adjectival negation present
(DOUBLE NEGATION)

d'. Keine Kopfverletzung ist so ungefährlich,
No head injury is so un-dangerous

"No head injury is so innocuous . . . "

dass sie nicht ignoriert werden sollte.
that it not ignored get should

". . . that it (/they) should not be ignored".

Procedure The procedure was the same as in Experiment 1.

Data analysis Data analysis was carried out analogously to Experiment 1. For all models, the factor negation was sum-coded with the double negation condition being coded as 1 and the adjectival negation condition being coded as −1. For the factor construction, a treatment contrast with the *um . . . zu* construction as the baseline was coded.

5.2 Predictions

Under the overloading hypothesis, given that the *too . . . as that* construction features an active verb, an overt subject and an overt modal verb, it should possibly cause fewer inversions than the *too . . . to* construction, as the reader needs to make fewer inferences (*Should/could/must be ignored by whom?*) and processing difficulty should thus be reduced. The *so . . . that* construction keeps the passive, but removes the implicit negation carried by *too*. As this negation is one out of a total of three that appear prior to inversion being triggered, and as implicit negation may be even more difficult to process than overt negation, the *so . . . that* construction may show the depth charge effect in a weakened form or not at all. Generally, if the presence of too many negations (and possibly other processing factors) makes subjects resort to "good enough" processing strategies, meaning inversion should occur across the entire spectrum of constructions.

A strong version of the ambiguity hypothesis would predict that only the *too . . . to* construction should show an illusion, given that it is represented lexically as a holistic unit with two prespecified meanings. The account of Cook & Stevenson (2010) makes no predictions as to whether meaning inversion generalizes to other constructions. With nothing else said, it may be assumed that the effect should not generalize. On the other hand, Fortuin (2014, p. 279) discusses systematic commonalities and differences between constructions that would predict the depth charge effect to occur in the *too . . . as that* construction but not in the *so . . . that* construction. According to Fortuin's analysis, both *too . . . to* and *too . . . as that* share the expressed semantics of an "excessive" degree introduced by *too*, compared to constructions with *so* that arguably do not express an excessive degree. Specifically, Fortuin (2014, p. 281) claims that the negative *too* construction "easily suggests [...] an excessive degree [on a scale] such that some situation is blocked, due to which the situation cannot be realized", which creates the need to express an additional, compositionally unlicensed negation. Furthermore, "[t]his inherent modality is absent in the resultative degree construction [with *so . . . that*]" (ibid.), hence no additional negation needs to be expressed. The prediction of Fortuin's account is thus that the double negation condition should receive higher ratings and be easier to process than the adjectival negation condition for the *too . . . to* and the *too . . . as that* construction, but not for the *so . . . that* construction, which should be processed compositionally and recognized as not being sensible in both conditions.

5.3 Results

Whole-sentence reading times and rating times by construction and condition are shown in Figure 8. Figure 9 shows the distribution of sensibleness ratings across constructions and conditions.

Reading times Compared to the baseline *too . . . to* construction, whole-sentence reading times were increased for both the *too . . . as that* construction ($\hat{\Delta}$ = 676 ms, CrI: [273 ms, 1104 ms], $\Pr(\beta > 0) \approx 1$) and the *so . . . that* construction ($\hat{\Delta} = 1108$ ms, CrI: [657 ms, 1605 ms], $\Pr(\beta > 0) \approx 1$). There was also a reliable interaction between negation and the *so . . . that* construction ($\hat{\Delta} = 990$ ms, CrI: [206 ms, 1744 ms], $\Pr(\beta > 0) = 0.99$), which was driven by a slowdown in the presence of double negation ($\hat{\Delta} = 1340$ ms, CrI: [467 ms, 2243 ms], $\Pr(\beta > 0) \approx 1$) that was absent in the baseline condition.

Rating times No effects in evidence.

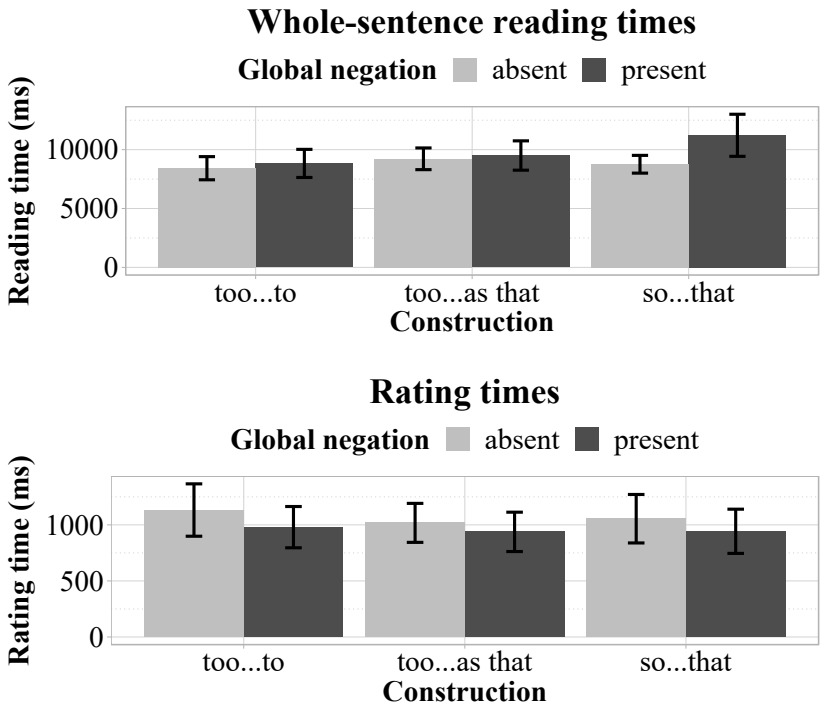


Figure 8: **Experiment 3 – Whole-sentence reading times and rating times by construction and condition.** Error bars show 95% confidence intervals.

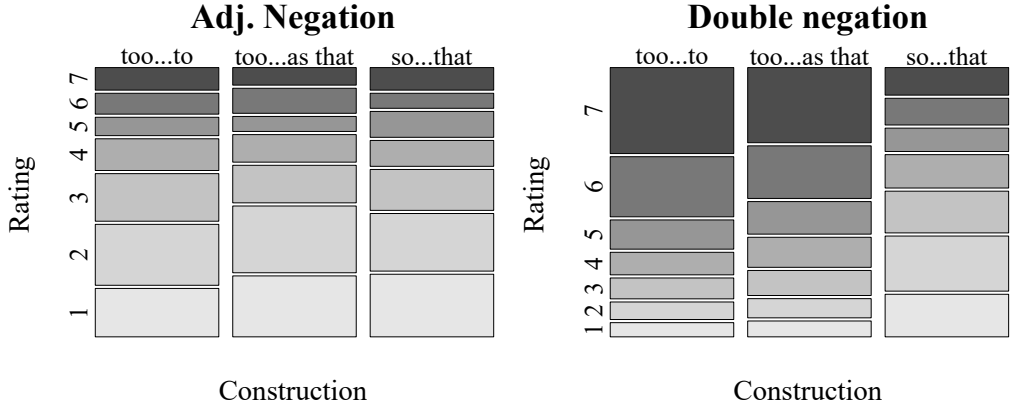


Figure 9: **Experiment 3 – Distribution of sensibleness ratings by construction and condition (1 = not sensible, 7 = perfectly sensible).**

Sensibleness ratings Sensibleness ratings were higher in the double negation compared to the adjectival negation condition in the baseline *too . . . to* construction ($\hat{\Delta} = 3.12$, CrI: [2.57, 3.58], $\Pr(\beta > 0) \approx 1$). Compared to the baseline, ratings were lower for the *too . . . as that* construction ($\hat{\Delta} = -0.36$, CrI: [−0.67, −0.03], $\Pr(\beta > 0) = 0.01$) as well as for the *so . . . that* construction ($\hat{\Delta} = -1.44$, CrI: [−1.76, −1.07], $\Pr(\beta > 0) \approx 0$). There was also an interaction between negation and the *so . . . that* construction ($\hat{\Delta} = -2.54$, CrI: [−3.14, −1.87], $\Pr(\beta > 0) \approx 0$), as the positive effect of double negation on ratings was weaker in the *so . . . that* construction

compared to that for the baseline ($\hat{\Delta} = 0.57$, CrI: [0.15, 0.96], $\Pr(\beta > 0) = 0.99$).

5.4 Discussion

Whole-sentence reading times showed no difference regarding the effect of double negation between the baseline *too . . . to* and the similar *too . . . as that* construction, but did show a difference between the *too . . . to* and the *so . . . that* construction, such that the latter showed increased reading times in the double negation condition while the former did not. Ratings of sensibleness also did not show a difference in the effect of double negation between the *too . . . to* and the *too . . . as that* construction, both of which exhibited an inversion effect of equal strength. The *so . . . that* construction on the other hand showed a significantly smaller but nevertheless non-zero increase in sensibleness ratings in the double negation condition. Taken together, the results suggest that the depth charge effect is not limited to the construction originally investigated by Wason & Reich (1979) and used in later empirical studies (Natsopoulos, 1985; Kizach et al., 2015; O'Connor, 2015). Furthermore, the observed pattern suggests that the *too . . . to* and the *too . . . as that* construction are essentially processed in the same way while the *so . . . that* construction shows a marked difference both with regard to visible processing difficulty as well as to the strength of the effect in sensibleness ratings.

Our result suggests that a strong version of the ambiguity hypothesis where *No X is too Y to Z* is the only construction that shows the depth charge effect is not tenable. Nevertheless, it is possible that the depth charge effect generalizes due to similarities between the different stored constructions. Fortuin's (2014) account predicts that constructions featuring *too* should show inversion while constructions with *so* should not. The prediction can be argued to have been borne out in our study, given that the additional negation only created processing difficulty in the *so . . . that* construction. However, Fortuin's account does not explain why higher sensibleness ratings were given in the double negation condition compared to the adjectival negation condition for this construction. To our minds, this result may either indicate a weak but nevertheless non-zero inversion effect or simply confusion on part of the participants. The latter explanation strikes us as less convincing because confusion should result in lower rather than higher ratings, unless participants gravitate towards the middle of the rating scale in such situations. The overloading account, meanwhile, would attribute the increase in processing difficulty in the *so . . . that* construction to the presence of an explicit negation, which is less likely to surreptitiously license the negative verb and cause meaning inversion.

All in all, the results of our experiments so far do not decisively favor either the overloading or the ambiguity-based account. However, as we believe the overloading account to be more in line with the broader empirical literature (see general discussion), we will propose a novel approach to the genesis of the depth charge effect that assumes a specific type of composition failure. In the next section, we sketch an account of how heuristic processing of multiple negations, presumably triggered by the implicit negation of *too*, may lead to the depth charge effect. Crucially, our proposed account assumes that the depth charge effect is triggered before the final verb. In order to test this prediction, we conduct a sentence completion study in which the stimulus sentences are truncated before the verb, similarly to O'Connor (2015; 2017).

6 Experiment 4

The findings of Experiment 2A appear to allow for the conclusion that the lexical verb is the source of the depth charge effect, consistent with the version of the ambiguity hypothesis proposed by Cook & Stevenson (2010) and the intuition of Wason & Reich (1979). However, the evidence from Experiment 2A is less conclusive than it appears at first glance, given that there may be partial processing spillover from previous regions: Linguistic processes triggered by an input word may become visible in on-line measures after readers have already advanced to the next word or even beyond, as outstanding integration steps may be carried over. At the end of the sentence, the spillover buffer is cleared in a "wrap-up" process (Just & Carpenter, 1980; Rayner et al., 2000), so that effects whose origins could potentially lie anywhere in the sentence will become visible at the final region. Given the high number of negations in the stimulus sentences, participants may have been forced to delay certain aspects of the compositional computation to the end of the sentence. Any leftover processing may then cause regressions to earlier regions for verification or reanalysis purposes, though currently there exists no precise theory as to how these processes operate (von der Malsburg & Vasishth, 2011, 2013).

Given that the evidence so far is inconclusive, further investigations are in order. The accounts of Cook & Stevenson (2010) and Wason & Reich (1979) predict that the depth charge effect should disappear when the lexical verb is removed from the sentence. This can be achieved by presenting participants with only the part of the sentence that leads up to the verb, and having them choose an appropriate continuation by themselves. If participants provide continuations that match a non-compositional as opposed to a compositional interpretation of the preamble, this would cast severe doubt on the assumption that the depth charge effect is triggered by the lexical verb. Recall that the account of Fortuin (2014) makes the prediction that meaning inversion should occur before the lexical verb, namely when the presupposition of "not acting" is triggered at *too*, setting it apart from the other proposed accounts. Wason & Reich (1979, p. 592) anecdotally report that if the verb *ignored* is substituted with *noticed* in the original example, the resulting sentence *No head injury is too trivial to be noticed* is often claimed to be nonsensical, even though it is compositionally sensible. A possible implication is that readers expect the verb *ignore* – or something semantically similar to it – to appear at the end of the sentence, as opposed to a verb that would be sensible under a compositional reading. This, along with the previous results from O'Connor (2015; 2017), would suggest that semantic inversion already occurs before the final verb appears.

It is not necessary to subscribe to the ambiguity account to derive the prediction that compositional processing is suspended prior to the appearance of the lexical verb. As a variant of the overloading approach of Wason & Reich (1979), one can assume that semantic composition fails before the lexical verb is encountered. Taking inspiration from the "good enough" approach to language processing, we suggest two plausible candidate heuristics that readers may apply. We assume that the processing of depth charge sentences is compositional at first, but that readers reach a motivation limit at some point where they consider a heuristic analysis to be a "good enough" approximation of the sentence meaning.

The first heuristic is *negation cancellation*: It assumes that adding another negation to a negated sentence always nullifies the effect of both negations.

Negation cancellation: Assume that two negations in a clause will cancel each other out. (*duplex negatio affirmat*)

This is not entirely unreasonable as a rule of thumb. For instance, the sentence *It's not like you didn't cheat on me* means *You cheated on me*, given that written American and British English, like German, do not exhibit negative concord.[12] However, in depth charge sentences, the configuration is such that the *duplex negatio affirmat* rule does not hold. The copula clause assigns a property to the subject, and negation scoping over the copula does not change the content of the property. Asserting that no head injury has the property of being too trivial to be ignored should thus leave intact the absurd meaning of the property *too trivial to be ignored*. The intuition behind the heuristic is that readers are generalizing a rule that holds in some double negation contexts to a context in which it should not be applied: If global and adjectival negation are assumed to cancel out, the sentence *No head injury is too trivial to be ignored* is transformed into *A least one head injury is too dangerous to be ignored*, which appears sensible.

Yet another possibility is that readers do not reason about how to combine the negations at all when faced with (seemingly) insurmountable processing difficulty; they just know from experience that it is usually the lexical verb of the sentence that negation is applied to.

Negate the verb: When in doubt, negate the lexical verb.

Note that the lexical verb is, in fact, *not* negated in depth charge sentences: Even though there is global negation, compositionally, the sentence *does* state that head injuries *should* be ignored, just like the sentence *No missile is too small to be banned* is usually correctly interpreted to mean that all missiles should be banned (Wason & Reich, 1979). The intuition behind *negate the verb* is that it may not be immediately obvious under processing pressure that the global negation does not negate the lexical verb, because *too* is not necessarily registered as introducing implicit negation. The negation on the verb can be seen as an "echo" of the global negation that is generated because the processor has lost track of how many negations it has encountered, presumably when *too* is read. The difference in the *missile* sentence is that the correct meaning is compatible with most people's views about the world and the scale is internally consistent (*smaller missile* $\rightarrow$ *less reason to ban*), so the comprehension system is never under enough pressure to have to resort to heuristics. In a depth charge configuration, then, semantic and pragmatic factors possibly conspire and derail compositional interpretation, so the processor has to heuristically infer that the sentence should, minimally, be taken to mean that something should *not be ignored*.

Having described two possible mechanisms by which meaning inversion may be triggered, both of which may apply before encountering the lexical verb, we now turn to our sentence completion study.

6.1 Method

Participants Sixty native speakers of German from the local student population participated in the experiment. They were either paid €7 or received credit points as compensation.

Materials The preambles used to elicit the sentence completions consisted of sentences from the double and adjectival negation conditions of the previous experiments that were pruned

[12] Negative concord refers to the semantically redundant use of negation, as in the French sentence *Personne ne boit rien* (lit. "Nobody drinks nothing"), meaning "Nobody drinks anything". Corblin (1995) analyses the second instance of negation to be "parasitic" on the first and argues that negative concord is enforced during processing by a constraint that serves to limit derivational complexity (see also Larrivée, 2016).

after *um*, 'too'. We chose the adjectival negation condition as our control condition because it received the lowest ratings in the previous experiments, indicating that no meaning inversion is to be expected.

(4) **Global negation absent, adjectival negation present**
(ADJECTIVAL NEGATION)

b. Manch eine Kopfverletzung ist zu ungefährlich, um ...
Some a head injury is too un-dangerous to

"Some head injuries are too innocuous to ..."

Global negation present, adjectival negation present
(DOUBLE NEGATION)

d. Keine Kopfverletzung ist zu ungefährlich, um ...
No head injury is too un-dangerous to

"No head injury is too innocuous to ..."

The rationale behind the design is that if inversion occurs in (4d), participants should volunteer completions like *be ignored*, whereas under a compositional reading of the preamble completions like *... be noticed* or *... be treated* should be given. For (4b), the latter two completions are also sensible under a compositional reading, and inversion is not expected to occur.

We opted for a sentence completion as opposed to a forced-choice design – in which one would have forced participants to choose either *... be ignored* or *... to be treated* as the continuation – because we did not want to bias subjects by explicitly offering written alternatives, which may trigger readings that would not otherwise be available. Given that participants were free to produce any kind of continuation, we had coders who were blind to experimental manipulation group the completions into binary categories (inversion versus no inversion, see below).

Procedure Participants were asked to complete the sentences in the way they found most plausible. Both the time taken to read the preamble and the time taken to finish typing in the response were recorded.

We created two coding schemes that allowed grouping the completions into binary categories ('inversion'/'no inversion'): Scheme A had coders decide whether completions signaled that the subject of the sentence was of low importance or interest (*head injury – ignore*), under the assumption that potential low importance is a hallmark of meaning inversion (see Appendix B for discussion).[13] Scheme B tested whether the completion fit with a sensible, negation-free sentence (*This head injury is too trivial to be ignored*), based on the observation that the inverted meaning is "normalized" to fit into a sensible template. The coding schemes are described in detail in Appendix B.

Data analysis For both coding schemes, data points for which a coder could not decide on a category were removed from the respective data set. Data for one item were completely removed from further analysis as an incomplete preamble had been presented by mistake. For the remaining items, inter-coder agreement was higher for coding scheme A (Fleiss' $\kappa = 0.77$,

[13]Note that potentially low importance also appears to summarize many, though not all, of the predicate classes identified by Fortuin (2014) based on corpus evidence.

"substantial agreement")[14] than for coding scheme B (Fleiss' $\kappa = 0.49$, "moderate agreement"). Completion types according to the different coding schemes were correlated at the observation level ($r = 0.52$, 95% confidence interval: [0.51, 0.53]).

The coded completions (inversion/no inversion) were analyzed using hierarchical logistic regression in brms with random intercepts and slopes for items, subjects and coders, as well as random intercepts for all coder-item and coder-subject pairs, given that each coder encountered the same item as well as responses from the same subject more than once. For the fixed effect of condition, the double negation condition was coded as 1 and the adjectival negation condition as -1. Reading times for the preamble as well as the time taken to produce the completion were analyzed analogously to previous experiments. For completion times, the length of the produced completion in characters was entered into the analysis as a centered and scaled predictor. Further details of the statistical analysis are described in Appendix B.

6.2 Predictions

The version of the ambiguity hypothesis proposed by Cook & Stevenson (2010) assumes that the lexical verb decides the ultimate meaning of the *No X is too Y to Z* construction. Given that subjects do not have access to the verb in the present design, they should thus be unsure as to which is the intended meaning, and plausibly resort to guessing, or default to the same meaning in both conditions. If readers experience confusion and resort to guessing, sometimes selecting the compositional meaning and sometimes the inverted meaning, we expect about 50% answers from both categories. Moreover, increased confusion in the double negation condition should lead to increased reading and/or completion times if subjects are having trouble deciding which continuation best fits the preamble. Wason & Reich (1979) hypothesized that encountering the lexical verb triggers memory overload and non-compositional interpretation, so that no depth charge effect is predicted in the absence of the verb and interpretation may always proceed compositionally.

Meanwhile, if the origin of the depth charge effect lies before the verb, as argued by Fortuin (2014), more inversion-signaling continuations are expected in the double negation condition compared to the adjectival negation condition. The account of Fortuin (2014) claims that a preamble such as *No head injury is too trivial ...* "presupposes" the use of a "negative" verb (p. 278). The gist of the proposal is that "negativity" in the preamble leads the reader to assume that a "negative" interpretation is intended, which may include a presupposition to the effect of "not acting" (p. 264).

Alternatively, under our proposed version of the overloading account, it is possible that doubly negated preambles lead to heuristic processing strategies such as the *negation cancellation* and *negate the verb* being used to predict a verb that matches the inverted meaning. If heuristic interpretation strategies are reliably applied, or if the "negative" preamble reliably allows for an intended "negative" meaning of the stored construction to be inferred, we expect to see a proportion of inversion-signaling continuations above 50% in the double negation condition.

[14] See Landis & Koch (1977) for a guide on how to interpret κ values. Also, note that the κ values were computed for the three coders assigned to each trial, without taking into account that different coders were assigned different parts of each list, which may have pushed down agreement values.

6.3 Results

Figure 10 shows preamble reading times, completion times and inversion proportions by condition.

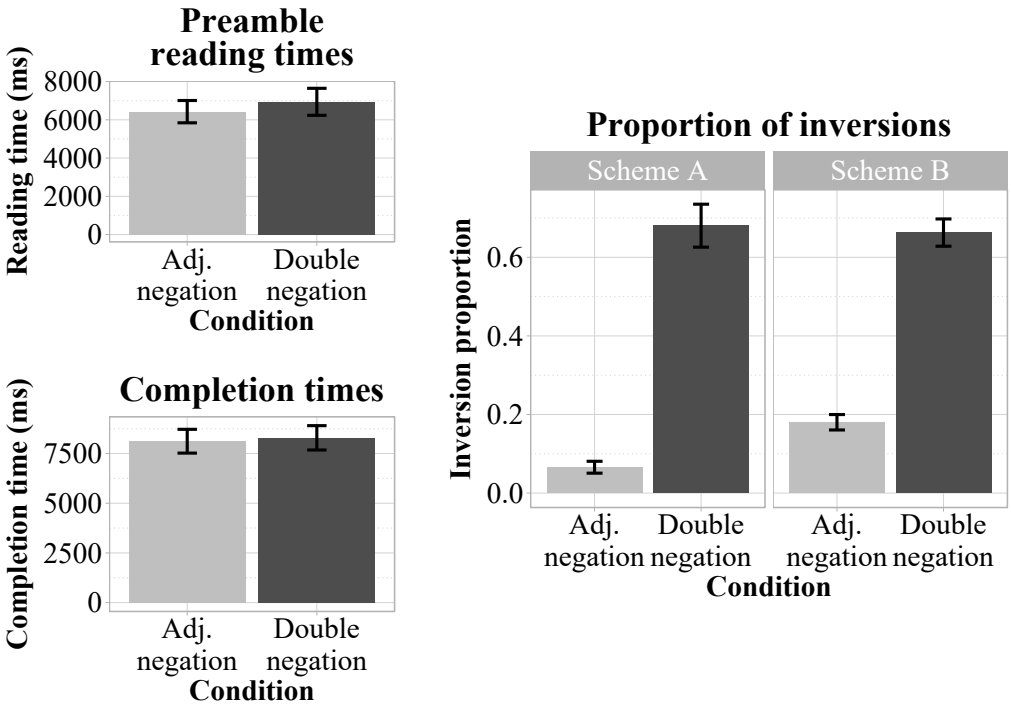


Figure 10: **Experiment 4 – Preamble reading times, completion times and proportion of inversions by condition.** Error bars show 95% confidence intervals.

Sentence completions Under coding scheme A, inversion occurred more often in the double negation than in the adjectival negation condition ($\hat{\Delta} = 0.67$, CrI: [0.57, 0.75], $\Pr(\beta > 0) \approx 1$). The same was true for coding scheme B ($\hat{\Delta} = 0.41$, CrI: [0.29, 0.5], $\Pr(\beta > 0) \approx 1$). Furthermore, the posterior mean of the proportion of inversion-signaling completions for the double negation condition lay above the chance expectation of 0.5 for both coding scheme A (posterior mean: 0.81, CrI: [0.71, 0.88]) and coding scheme B (posterior mean: 0.73, CrI: [0.61, 0.82]).

Reading and sentence completion times Neither reading nor sentence completion times showed any evidence of a difference between the conditions. Longer completions did, however, take longer to produce ($\hat{\Delta} = 2430$ ms, CrI: [2357 ms, 2506 ms], $\Pr(\beta > 0) \approx 1$).

6.4 Discussion

The findings suggest that the depth charge effect is not mainly caused by the sentence-final verb, but that the anomalous verb is selected because the preamble reliably "keys" the non-compositional meaning in double negation sentences (Fortuin, 2014; see also O'Connor, 2015). This conclusion does not change depending on whether an abstract semantic dimension of subjective importance or the fit with a matched negation-free sentence is used as the basis for coding.

Detailed results for the two coding schemes, as well as a discussion of potentially problematic data points, can be found in Appendix B.

The results are compatible with the particle *too* rather than the lexical verb being the main culprit behind the depth charge effect, as assumed by Fortuin (2014). Inversion-signaling continuations were produced more often than would be expected if readers were choosing a completion by chance, suggesting that non-compositional processing prior to encountering the lexical verb is the norm in depth charge sentences, and that the inverted meaning is reliably computed. As in previous experiments, there was no indication that the double negation condition was more difficult to process than the adjectival negation condition, suggesting that some aspect of the compositional analysis is left out. This can be interpreted either as evidence that readers are accessing the "negative" version of the *No X is too Y to Z* construction, as predicted by Fortuin's (2014) version of the ambiguity hypothesis, or that they are switching to heuristic processing – using *negation cancellation* or *negate the verb* – and thus predicting the compositionally unlicensed verb.

7 General discussion

In a series of five experiments on German, we have shed new light onto the classic meaning inversion effect in depth charge sentences, such as *No head injury is too trivial to be ignored*. Using a design which varied the presence versus absence of negation at the beginning of the sentence and at the adjective within the copula phrase, Experiments 1 and 2A showed that only when both negations were present, the depth charge effect became visible in ratings of sensibleness. Both experiments also yielded evidence that in the presence of global negation, adjectival negation, despite causing the sentence to be internally inconsistent, does not cause an increase in processing difficulty, as indicated by both whole-sentence reading times and eye tracking measures across the sentence. Both findings are, in principle, compatible both with the assumption of composition failure (Wason & Reich, 1979; Kizach et al., 2015) and with the ambiguity hypothesis (Cook & Stevenson, 2010; Fortuin, 2014). However, as Experiment 2A yielded no reliable evidence that the depth charge effect is influenced by working memory capacity, the data do not provide evidence for the overloading account of Wason & Reich (1979) as a specific instance of the overloading hypothesis. This is not to say that we have found evidence *against* the overloading account; the results are simply inconclusive. Nevertheless, we have suggested that as opposed to a limit of working memory capacity, readers may run into a limit of motivation to further process the sentence, at which point they switch to a "good enough" interpretation strategy.

Based on the assumption of Fortuin (2014) and Cook & Stevenson (2010) that *No X is too Y to be Z* is interpreted as a holistic unit, Experiment 3 tested whether the depth charge effect is also detectable in two related constructions in German. We found that meaning inversion does indeed occur in these constructions, but that it appears to be strongly related to the appearance of the particle *too*, which introduces an implicit negation: When *too* was absent, the depth charge effect still appeared, but was of a much smaller magnitude compared to when it was present. While the results are compatible with the ambiguity hypothesis under the assumption that different constructions may exhibit the same behavior, they provide evidence against an account in which the *No X is too Y to be Z* construction is the only configuration in which the depth charge effect occurs.

We also investigated the point of origin of the depth charge effect within the sentence. Ex-

periment 4 showed that meaning inversion occurs prior to the appearance of the lexical verb using sentence completions, as had been previously observed by O'Connor (2015; 2017). This suggests that both Wason & Reich's speculation that the compositional derivation is derailed by the verb *ignore* as well as the assumption by Cook & Stevenson (2010) that the verb is the key to the intended meaning of the construction are incorrect.[15] Under the overloading account, one needs to assume that compositional processing fails at an earlier point in the sentence and that non-compositional processes are then used to predict the incorrect verb. We have argued that our proposed heuristics *negation cancellation* and *negate the verb* are plausible candidates for mechanisms that ultimately yield the inverted reading when readers exceed their motivation limit. Alternatively, the proposal by Fortuin (2014) that the use of two negative elements (*no* and *too*) creates and then cancels a presupposition of "not acting" predicts the licensing of a "negative" verb in depth charge contexts. Finally, our result matches Wason & Reich's (1979) observation that the sentence *No head injury is too trivial to be noticed* is sometimes judged not to be sensible despite having a sensible compositional meaning: Readers apparently expect a continuation with the approximate semantics of *ignored* and are surprised when they encounter a continuation that has the opposite meaning.

We suggest that readers may make use of the *negation cancellation* and *negate the verb* heuristics as a last resort when faced with a sentence that is otherwise impossible to parse. When *negation cancellation* is applied to the sentence *No head injury is too trivial to be ignored*, global and adjectival negation nullify each other, which leaves *At least one (≈ some) head injury is too dangerous to ignore*, the meaning of the sensible no negation sentence. Subjects applying *negation cancellation* would be wrongly using the "conversion method", where a doubly negated sentence is converted into its non-negated counterpart in order to be more easily interpretable (e.g. Clark, 1976). Indeed, we have anecdotal evidence from two subjects who reported using this method to interpret depth charge sentences. Furthermore, anecdotal evidence reported by Wason (1961) suggests that subjects may use the conversion method even in single-negation sentences.[16]

One fact that is not accounted for by *negation cancellation* is the existence of examples of depth charge sentences without adjectival negation, such as *No challenge is too big to stop us from saving our children from polio* (Fortuin, 2014). Fortuin (p. 253) gives a list of such examples and argues that they show the "four negations" generalization made by Wason & Reich (1979) to be incorrect. However, as we have noted before, the results of Kizach et al. (2015) show that adjectival negation *does* contribute significantly to the depth charge effect, even though it is not a necessary prerequisite. Thus, while *negation cancellation* cannot be the sole explanation for the depth charge effect, it nevertheless potentially accounts for a large subset of meaning inversions.

[15]Recall, however, that the computational model implemented by Cook & Stevenson (2010) was highly accurate at identifying the "correct" meaning using only information from the verb. Nevertheless, it does not seem to be the case that the verb causally "gives rise to the meaning of the target construction" (p. 68).

[16]Note that *duplex negatio affirmat* generally does not apply in multiclausal sentences such as *Because he didn't want to insult her, he did not make a comment.* The fact that the rule cannot be lawfully applied in depth charge sentence thus fits well with Schwarzschild's (2008) semantic analysis of *too*, which is argued to contain an implicit embedded clause headed by *because* (*X is too young to smoke* → X should not smoke because X is too young). The double-negation rule does also not apply to cases of double negative quantification, as in *None of the girls met none of the boys.* Furthermore, as noted by Horn (2001) and pointed out by an anonymous reviewer, two negations routinely do not cancel out if one of them is expressed by a bound morpheme (e.g. *The king is not unkind to his enemies*), as in classic depth charge sentences. Given these observations, it would be highly dubious to claim that *negation cancellation* is widely applied as a heuristic during normal sentence processing, as the number of exceptions would be too high.

A formalization of the *negation cancellation* and *negate the verb* heuristics, along with a third possibility, namely the conversion of *too* into the semantic equivalent of *enough*, is given in Appendix C.

7.1 The depth charge illusion in the broader empirical context

The notion of a strategic "time-out" that stops compositional processing after a fixed period of time is compatible with the idea of partially "good enough" or "shallow" linguistic processing (e.g. Ferreira et al., 2002; Sanford & Sturt, 2002; Karimi & Ferreira, 2016; Christianson, 2016), and with the idea of a "stop rule" that terminates processing when the current output is deemed satisfactory (e.g. Simon, 1972). The "good enough" approach to sentence comprehension maintains that readers do not necessarily construct a fully specified representation of the input, but are in pursuit of an interpretation that is deemed sufficient given current task demands (e.g. Swets et al., 2008). It is by no means clear whether such "shallow" representations are usually the result of strategies that are consciously applied by participants or whether resource limitations force subjects to adopt incomplete representations. For instance, von der Malsburg & Vasishth (2013) found evidence suggesting that participants with low working memory capacity leave syntactic attachments underspecified more often. If low-capacity participants are forced into underspecification due to their processing system's inherent limitations, "good enough" is something of a misnomer: It implies a conscious decision to abort processing when one can be reasonably sure that the task at hand can be solved given the current representation. On the other hand, if the system simply runs out of resources at some point, there is no reasonable expectation that the current output structure will be sufficient.

The ambiguity hypothesis avoids this question by assuming that readers are mainly trying to infer the communicative (or rhetorical) intention behind the utterance to decide whether a "positive" or a "negative" reading should be derived. Interestingly, when reading depth charge sentences, one neither gets the subjective impression of having been exposed to a rhetorical device nor of having failed to grasp the correct meaning due to complexity overload. Under the overloading account, "failure" apparently does not entail "awareness of failure", unlike in particularly difficult garden-path sentences (*The horse raced past the barn fell*; see also O'Connor, 2015, p. 226/7).

It is remarkable that our depth charge stimuli showed relatively high ratings across studies despite the amount of processing difficulty they cause in comparison to negation-free sentences. Normally, one would assume acceptability to suffer more noticeably than it did in the present study when there is processing difficulty (e.g. Warren & Gibson, 2002; Fanselow & Frisch, 2006; Hofmeister et al., 2013). We speculate that our instructions prompted subjects to not take into account how easy or difficult the sentences were to process when assigning their ratings, but focus on the end result of their effort. Recall that the instructions were to indicate whether the sentences "made clear sense and contained no grammatical mistakes". As depth charge sentences appear to make sense at first glance, they would probably not arouse any suspicion in a connected discourse, where the prior expectation that utterances are sensible would likely cause an immediate switch to non-compositional processing. In such a setting, the subjective impression of fluent processing would likely also preempt any inclination to second-guess the adopted interpretation. However, in the context of an explicit judgment task, checking for errors and meaning incongruity likely causes processing to be experienced as more disfluent, which may in turn lead to more analytic processing of the stimulus (Alter et al., 2007).

With regard to the apparent mismatch between the assumption of processing failure and perceived acceptability, depth charge patterns with a number of other phenomena where normal processing fails or is suspended but no conscious failure is registered, and even processing facilitation may be observed:

- **Agreement attraction**: Sentences that are ungrammatical due to number mismatch between subject and verb sometimes appear grammatical, and are processed faster, when a noun phrase with a matching number feature appears in a structurally inaccessible position (**The key to the cabinets are on the table*; e.g. Kimball & Aissen, 1971; Wagers et al., 2009; Dillon et al., 2013; Lago et al., 2015; Jäger et al., 2017).
- **Intrusive NPI licensing**: Similarly to agreement attraction, the negative polarity item *ever* can sometimes be erroneously licensed by a negative element in a non-c-commanding position, which results in some processing disruption and positive grammaticality judgments (**A pirate who had no beard was ever thrifty*; Drenhaus et al., 2005; Vasishth et al., 2008; Xiang et al., 2013; Parker & Phillips, 2016).
- **Structural forgetting**: Sentences containing complex center embeddings are read faster in English when a required verb is missing (**The apartment that the maid who the service had sent over was well decorated*; Gibson & Thomas, 1999; Vasishth et al., 2010; Frank et al., 2016).
- **Underspecification**: In the presence of syntactic ambiguity, processing time is shorter if the reader does not commit to an analysis (Swets et al., 2008; von der Malsburg & Vasishth, 2013; Logačev & Vasishth, 2016; Nicenboim et al., 2016).
- **Comparative illusions**: Participants often judge sentences such as *More people have been to Russia than I have* as well-formed even though they are not (O'Connor, 2015; Wellwood et al., 2018).

Looking at the depth charge effect in this context, the claim made by the ambiguity hypothesis that the compositionally incorrect reading is licensed by the grammar becomes less convincing: The existence of systematic patterns of positive acceptability judgments in the absence of a word-by-word compositional derivation does not necessarily constitute evidence that these judgments are licensed by grammar.

7.2 A possible synthesis of the accounts, and the role of world knowledge

In light of the results of Experiment 3, where the depth charge effect was observed for different constructions, one could argue that the overloading account is the more parsimonious approach to explaining the depth charge effect, as it naturally predicts that the effect should generalize. However, setting up the overloading account and the ambiguity account as mutually exclusive may be misguided. The model of Kuperberg (2007) assumes that there are two linguistic processing streams that work in parallel during comprehension: A semantic stream and a combinatorial syntactic stream. The semantic stream is sensitive to meaning relationships between content words while the combinatorial stream keeps track of the syntactic structure of the input. It is entirely plausible that the intuition behind the ambiguity hypothesis partly maps onto the workings of the semantic processing stream, which is sensitive to lexical meaning, associative relationships between words and world knowledge. The claim of Fortuin (2014) that the "negative" reading of depth charge sentences serves a rhetorical function could potentially be subsumed under this aspect of processing. Kuperberg assumes that the syntactic-combinatorial stream

usually overrules the semantic stream in case of a mismatch between the respective representations of sentence meaning. However, if compositional processing is aborted due to complexity overload or strategic suspension, the output of the semantic stream may determine sentence interpretation.[17] Such a combined account would possibly obviate the need for negation-related processing heuristics such as the ones we have proposed.

Recall that Experiment 2B yielded evidence that world knowledge was recruited even in sentences with no overt negations. The predicted interaction between world knowledge and the number of negations appeared only in sensibleness ratings, but not in the on-line measures, which may suggest that the depth charge effect is partly due to world knowledge affecting "post-interpretive" processing (Caplan & Waters, 1999). This supports the proposal of Kuperberg (2007, p. 37) that "the meanings [of the words] are first combined through pragmatic or inferential heuristic mechanisms into tentative propositions (a 'quick and dirty' means of deriving the gist of a proposition) and that it is the plausibility of this proposition as a whole that is then evaluated against real-world knowledge [...]". Our results suggest that world knowledge serves both as a guide during on-line processing and as the basis for a final check of the derived semantics, where it contributes to the depth charge effect.[18]

World knowledge may affect the "mental model" that readers construct of the sentence meaning during processing (e.g. Glenberg et al., 1987; Kaup et al., 2006, 2007). Mental models are simulations that rely on experience, so it is not unlikely that they would resist conforming to the nonsensical input, and that readers may use a "pragmatically normalized" (Fillenbaum, 1974; Garrod & Sanford, 1995) representation of the proposition as a basis (such as the treatment of a minor head injury). Speculatively, readers may first construct a basic "setting" including the concepts that are mentioned in the sentence, computing their relations only as a second step during off-line processing.[19].

7.3 The role of incrementality and prediction

A further point that can be made in support of the overloading account is that it is, in principle, possible to make readers aware of the incorrectness of the inverted reading. One strategy is to instruct readers to not start reading the depth charge sentence from the beginning, but to only look at the phrase *too trivial to be ignored* first before trying to combine it with the global negation, which anecdotally often results in them noticing the error. The success of this strategy highlights the role of incrementality in the genesis of the depth charge effect: Only when global negation is processed before encountering the *too* phrase does meaning inversion

[17] See Townsend & Bever (2001) and Ferreira (2003) for related proposals.

[18] O'Connor (2015) found no evidence for an effect of world knowledge, but it is possible that this is a false negative result, or that the experimental design employed was not optimal for detecting an effect. The latter may be partly because *all* was taken to be the semantic opposite of *no*, whereas we have argued that *some* is better suited. In addition, O'Connor (2015) asked participants to rate "the extent to which [the sentence] describes a realistic scenario" (p. 191), which may lead to quite different inferences compared to our approval ratings in Experiment 2B, given that depth charge sentences do not describe concrete scenarios but rather express an opinion (*Head injuries should be treated*).

[19] This idea is similar to the account proposed by Fischler et al. (1983), in which negated sentences such as *A robin is not a truck* are first evaluated without the negation before the actual proposition is computed (but see Nieuwland & Kuperberg, 2008 for a contrasting view)

become irreversible.[20] One can also present minimal pairs of *too*- and *enough*-sentences or *ignore*- and *treat*-sentences side-by-side and point out that they cannot (or at least should not) mean the same thing. Or one can paraphrase the sentence by putting all content except the global negation into a subordinate clause, as in *It is not the case that a head injury can be too trivial to be ignored*, though some readers will insist on the inverted reading even then.

Experiment 4 suggests that readers usually *expect* to see a non-compositionally licensed verb in depth charge sentences. The influential surprisal theory (Hale, 2001; Levy, 2008a) claims that low predictability of a word in a context increases processing effort, so it is not surprising that compositionally sensible versions of depth charge sentences (*No head injury is too trivial to be noticed*) are often difficult to understand. However, it might also be argued that readers start out *not* expecting the anomalous verb, but instead retroactively change their mental representation of the preceding input because they are uncertain as to what they have read, as predicted by the noisy-channel model of language processing (Levy, 2008b).[21] The noisy- or lossy-context surprisal account proposed by Futrell & Levy (2017) and Futrell et al. (2020) would instead assume that by the time readers reach the verb, they have partially forgotten the previous input and thus fail to make the correct prediction. We believe that neither of these accounts offers a good explanation of the depth charge effect: The noisy-channel model assumes that mentally revising the previous input to conform with the unexpected current input is computationally costly, yet our data show no evidence of increased processing cost in the depth charge condition. The noisy-context surprisal model assumes that previous input is simply erased from memory. Given that rereading was the norm in Experiment 2A, it would be surprising if readers remained convinced of the correctness of their misinterpretation even when having had the opportunity to refresh their memory of the input. Regressions do not appear to lead to low ratings for depth charge sentences: In the double negation condition, 62% of trials had at least one regression, and across these, there was a mean of six regressions per trial, indicating several passes over the material.[22] Still, the mean sensibleness rating across trials with six or more regressions was a little over 5 (95% confidence interval: [4.53, 5.59]), and therefore on the positive side of the scale. For comparison, the mean rating for the more transparently incoherent adjectival negation condition was 3.5.

One possible takeaway under the overloading account is that the depth charge effect is like a fishing weir: Once the inverted interpretation has been even tentatively adopted, there is no going back; that is, reanalysis is impossible or near impossible. The alternative view is that once readers have identified the intended meaning of the *No X is too Y to Z* construction, they may regress to check if their interpretation is correct, but will mostly stick with the initially assigned semantics.

[20] The same point is also made by (Fortuin, 2014, p. 278), who does not see the role of incrementality as standing in opposition to a construction-based account of the depth charge effect.

[21] A different account would claim that readers' expectation for the correct verb is so strong that it simply overrides the aberrant input in depth charge sentences (Pickering & Garrod, 2007). Our results as well as those of O'Connor (2015; 2017) provide evidence against this account.

[22] Note that as we divided sentences into three multi-word regions of interest, it is likely that additional inter-word regressions occurred.

7.4 Outlook

There is no shortage of contexts in which the interpretation of multiple negative elements does not work as expected (e.g. Horn, 2009; Krifka, 2011; de Dios-Flores, 2019). It is an open question whether the negation-related heuristics we have described are generally applied to sentences with multiple negations, and whether influences of world knowledge on interpretation can be found in different contexts as well. It appears that "negation" or "negative polarity" needs to be understood in a wider sense than just referring to items such as *not* and *no*: It also encompasses negative affixes on adjectives, "negative" verbs like *ignore*, and possibly nouns, as shown by examples such as *Loss of virtue is irretrievable*, taken from Jane Austen's *Pride and Prejudice* (Beck et al., 2008).

Yet another aspect to the processing of depth charge sentences that we have not touched upon in detail is the processing of embedded entailments. The negative quantifier *no* is downward-entailing, that is, if no head injury has property X, then it follows that any one head injury does not have property X. The *too*-phrase is also downward-entailing: *The tea is too hot to hold and drink* does not entail *The tea is too hot to drink* (inference to a superset), but does entail *The tea is too hot to hold and drink quickly* (inference to a subset). Furthermore, *too*-phrases license negative polarity items (*The tea is too hot to ever be held*), which is usually taken as an indicator of downward-entailingness (Ladusaw, 1980). It is possible that "implicit negation" does not correctly describe the meaning contribution of *too*, but that the operator's crucial property is that of inverting entailment relations.[23] While we cannot rule out that the depth charge effect is at least partly due to the difficulty induced by having two downward-entailing operators in the sentence (Geurts & van Der Slik, 2005), we refrain from offering an account based on entailment processing here, leaving the issue to future work.

In conclusion, we have demonstrated that even multi-faceted phenomena like the depth charge effect can be disassembled in ways that put them within the scope of detailed experimental evaluation of hypotheses, which yields valuable information about the mechanisms involved. In the future, we hope to further hone our empirical and theoretical tools in order to be able to tackle other negation-related phenomena in the literature that have previously been noted as curious anomalies but not been subjected to large-scale experimental investigation, such as the many examples of hypo- and hypernegation given by Horn (2009) and O'Connor (2015). To echo the conclusion of Horn (p. 419), "negation is the un-wizzywig of grammatical categories, where all too often what you see is what you don't get — and vice versa".

Authors' address

Universität Potsdam
Department Linguistik
Karl-Liebknecht-Straße 24-25
D-14476 Potsdam
Germany

[23] We do not make the claim that *too* should necessarily be grouped with elements that introduce "proper" negation, but maintain that "implicit negation" is a fitting term, seeing that the negative inference *X should not Z* does occur in the presence of the operator. In this context, also compare the claim of Schwarzschild (2008) that *too* introduces an implicit negated modal verb (*can't*, *shouldn't*) and licenses the use of *let alone* (*This car is too expensive to buy, let alone drive!*), as well as Horn's (2009) classification of *too* as introducing negation.

Acknowledgments

The authors wish to thank the Vasishth Lab team, the audiences at CUNY 2017 at MIT, AMLaP 2018 in Berlin and the 2018 Workshop on Psycholinguistic and Computational Perspectives on Non-Compositional Meaning in Phrases in Tübingen, as well as three anonymous reviewers for criticism and helpful comments. We also wish to thank Johanna Thieke for assistance with subject recruitment and data collection. All experiments were funded by the University of Potsdam.

References

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, *136*(4), 569.

Anderson, S. R. (2006). *Doctor Dolittle's delusion: Animals and the uniqueness of human language*. New Haven: Yale University Press.

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.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Beck, S., Crnic, L., & Götz, T. (2008). Ruin and restitution. *Natural Language Semantics*, *16*(2), 111–114.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Cacciari, C., & Corradini, P. (2015). Literal analysis and idiom retrieval in ambiguous idioms processing: A reading-time study. *Journal of Cognitive Psychology*, *27*(7), 797-811. doi: 10.1080/20445911.2015.1049178

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, *27*(6), 668–683.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(1), 77–94.

Chomsky, N. (1964). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Christianson, K. (2016). When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *The Quarterly Journal of Experimental Psychology*, *69*(5), 817–828.

Clark, H. H. (1976). *Semantics and Comprehension*. The Hague: Mouton.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

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* (pp. 61–69).

Corblin, F. (1995). Compositionality and complexity in multiple negation. *Logic Journal of the IGPL*, *3*(2–3), 449–471.

de Dios-Flores, I. (2019). Processing sentences with multiple negations: Grammatical structures that are perceived as unacceptable. *Frontiers in Psychology*, *10*, 2346.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*(2), 85–103.

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 145–165.

Drummond, A. (2018). Ibex farm [Computer software manual]. Retrieved from `http://spellout.net/ibexfarm/`

Dwivedi, V. D. (2013). Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, *8*(11), e81461.

Fanselow, G., & Frisch, S. (2006). Effects of processing difficulty on judgments of acceptability. In G. Fanselow, C. Féry, M. Schlesewsky, & R. Vogel (Eds.), (pp. 291–316). Oxford: Oxford University Press.

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

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*(1), 11–15.

Fillenbaum, S. (1974). Pragmatic normalization: Further results for some conjunctive and disjunctive sentences. *Journal of Experimental Psychology*, *102*(4), 574.

Fillmore, C. J. (1985). Syntactic intrusions and the notion of grammatical construction. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* (Vol. 11, pp. 73–86).

Fischler, I., Bloom, P. A., Childers, D. G., Roucos, S. E., & Perry Jr, N. W. (1983). Brain potentials related to stages of sentence verification. *Psychophysiology*, *20*(4), 400–409.

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), 249–292.

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*(3), 554–578.

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814.

Futrell, R., & Levy, R. (2017). Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (p. 688–698).

Garrod, S., & Sanford, A. (1995). Incrementality in discourse understanding. In D. Milward & P. Sturt (Eds.), *Incremental interpretation* (pp. 99–122).

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Geurts, B., & van Der Slik, F. (2005). Monotonicity and processing load. *Journal of Semantics*, *22*(1), 97–117.

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

Glenberg, A. M., Meyer, M., & Lindem, K. (1987). Mental models contribute to foregrounding during text comprehension. *Journal of Memory and Language*, *26*(1), 69–83.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8).

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., & Snider, N. (2013). The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes*, *28*(1-2), 48–87.

Horn, L. R. (2001). Flaubert triggers, squatitive negation and other quirks of grammar. *Perspectives on negation and polarity items*, 173–200.

Horn, L. R. (2009). Hypernegation, hyponegation, and parole violations. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* (Vol. 35, pp. 403–423).

Jackendoff, R., & Wittenberg, E. (2014). What you can say without syntax: A hierarchy of grammatical complexity. In F. J. Newmeyer & L. B. Preston (Eds.), *Measuring Grammatical Complexity* (pp. 65–82). Oxford: Oxford University Press.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217.

Karimi, H., & Ferreira, F. (2016). Good-enough linguistic representations and online cognitive equilibrium in language processing. *The Quarterly Journal of Experimental Psychology*, *69*(5), 1013–1040.

Kaup, B., Lüdtke, J., & Zwaan, R. A. (2006). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, *38*(7), 1033–1050.

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R. A., & Lüdtke, J. (2007). Experiential simulations of negated text information. *The Quarterly Journal of Experimental Psychology*, *60*(7), 976–990.

Kimball, J., & Aissen, J. (1971). I think, you think, he think. *Linguistic Inquiry*, *2*(2), 241–246.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*(5), 580–602.

Kizach, J., Christensen, K. R., & Weed, E. (2015). A verbal illusion: Now in three languages. *Journal of Psycholinguistic Research*, 1–16.

Krifka, M. (2011). How to interpret "expletive" negation under *bevor* in *German*. In T. Hanneforth & G. Fanselow (Eds.), *Language and logos. Studies in theoretical and computational linguistics* (pp. 214–236). Berlin: Akademie Verlag.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49.

Ladusaw, W. A. (1980). *Polarity sensitivity as inherent scope relations*. New York: Garland.

Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement attraction in spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

Lambrecht, K. (1988). There was a farmer had a dog: Syntactic amalgams revisited. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 14, pp. 319–339).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Larrivée, P. (2016). The markedness of double negation. In P. Larrivée & C. Lee (Eds.), *Negation and polarity: Experimental perspectives* (pp. 177–198). Cham: Springer.

Levy, R. (2008a). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R. (2008b). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 234–243).

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.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

Logačev, P., & Vasishth, S. (2016). Understanding underspecification: A comparison of two computational implementations. *The Quarterly Journal of Experimental Psychology*, *69*(5), 996–1012.

McClelland, J. L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and cognitive processes*, *4*(3-4), 287–335.

Natsopoulos, D. (1985). A verbal illusion in two languages. *Journal of Psycholinguistic Research*, *14*(4), 385–397.

Ng, V., Dasgupta, S., & Arifin, S. M. N. (2006). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions* (pp. 611–618). Stroudsburg, PA, USA: Association for Computational Linguistics.

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.

Nieuwland, M. S., & Kuperberg, G. R. (2008). When the truth is not too hard to handle: An event-related potential study on the pragmatics of negation. *Psychological Science*, *19*(12), 1213–1218.

O'Connor, E. (2015). Comparative illusions at the syntax-semantics interface. *Los Angeles, CA: University of Southern California dissertation.*

O'Connor, E. (2017). The accidental ambiguity of inversion illusions. In A. Lamont & K. Tetzloff (Eds.), *Proceedings of NELS 47* (p. 329-342).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing, Volume 10* (pp. 79–86).

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321–339.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Retrieved from `https://www.R-project.org/` (Version 3.4.4)

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705.

Rayner, K., Kambe, G., & Duffy, S. A. (2000). The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology Section A*, *53*(4), 1061–1080.

Rohde, D. (2003). *Linger.* Retrieved from `http://tedlab.mit.edu/ dr/Linger/` (Version 2.94)

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377–381.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences*, *6*(9), 382–386.

Schielzeth, H., & Forstmeier, W. (2008). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology*, *20*(2), 416–420.

Schwarzschild, R. (2008). The semantics of comparatives and other degree constructions. *Language and Linguistics Compass*, *2*(2), 308–331.

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, *15*(2), 143–157.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and Organization*, *1*(1), 161–176.

Stan Development Team. (2018). Stan modeling language users guide and reference manual, Version 2.18.0 [Computer software manual]. Retrieved from `http://mc-stan.org`

Swets, B., Desmet, T., Clifton, C., & Ferreira, F. (2008). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, *36*(1), 201–216.

Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. MIT Press.

Trousdale, G. (2012). Grammaticalization, constructions and the grammaticalization of constructions. In K. Davidse, T. Breban, L. Brems, & T. Mortelmans (Eds.), (pp. 167–198). Amsterdam: John Benjamins.

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4), 685–712.

Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*(4), 533–567.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth ed.). New York: Springer. Retrieved from `http://www.stats.ox.ac.uk/pub/MASS4` (ISBN 0-387-95457-0)

von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*(2), 109–127.

von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*(2), 206–237.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, *85*(1), 79–112.

Wason, P. C. (1961). Response to affirmative and negative binary statements. *British Journal of Psychology*, *52*(2), 133–142.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*(4), 591–597.

Wellwood, A., Pancheva, R., Hacquard, V., & Phillips, C. (2018). The anatomy of a comparative illusion. *Journal of Semantics*, *35*(3), 543–583.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing* (pp. 60–68).

Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*.

Xiang, M., Grove, J., & Giannakidou, A. (2013). Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Frontiers in Psychology*, *4*, 708.

Appendix A: Model calls and fixed-effects output

```
packageVersion("brms")
```

```
## [1] '2.10.0'
```

```
library(data.table)
library(tidyverse)
library(brms)

# Priors for models on log scale (reading times/measures) and logit scale (ratings, proportions)

prior_log <- c(set_prior("normal(0,5)",class="b"),
               set_prior("normal(0,5)",class="Intercept"),
               set_prior("lkj(2)", class = "cor"))
prior_logit <- c(set_prior("normal(0,2)",class="b"),
                 set_prior("normal(0,2)",class="Intercept"),
                 set_prior("lkj(2)", class = "cor"))

# Note: brms default priors were used for the remaining parameters, see
# https://www.rdocumentation.org/packages/brms/versions/2.10.0/topics/set_prior
# https://www.rdocumentation.org/packages/brms/versions/2.10.0/topics/brmsformula

# Note that condition labels differ from those used in the text:
# a: Double negation
# b: Global negation
# c: Adjectival negation
# d: No negation

load("exp1_data.Rda")

exp1_data <- filter(exp1_data,rat_time<10000&rat_time>150)

# Experiment 1: Reading times

exp1_rt <- brm(bf(read_time ~ gneg*aneg+(gneg*aneg|subj)+(gneg*aneg|item)),
               exp1_data, family = shifted_lognormal(), chains = 4, cores = 4,
               iter=2000,prior = prior_log)
save(exp1_rt,file="exp1_rt.Rda")
summary(exp1_rt)

# ab: Adj. negation within global negation; cd: Adj. negation within no global negation

exp1_rt_nested <- brm(bf(read_time ~ ab+cd+gneg+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                      exp1_data, family = shifted_lognormal(), chains = 4, cores = 4,
                      iter=2000,prior = prior_log)
save(exp1_rt_nested,file="exp1_rt_nested.Rda")
summary(exp1_rt_nested)

exp1_rat <- brm(bf(rat_time ~ gneg*aneg+(gneg*aneg|subj)+(gneg*aneg|item)),
                exp1_data, family = shifted_lognormal(), chains = 4, cores = 4,
                iter=2000,prior = prior_log)
save(exp1_rat,file="exp1_rat.Rda")
summary(exp1_rat)
```

```
exp1_rat_nested <- brm(bf(rat_time ~ ab+cd+gneg+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                      exp1_data, family = shifted_lognormal(), chains = 4, cores = 4,
                      iter=2000,prior = prior_log)
save(exp1_rat_nested,file="exp1_rat_nested.Rda")
summary(exp1_rat_nested)

exp1_rating <- brm(bf(rating ~ gneg*aneg+(gneg*aneg|subj)+(gneg*aneg|item)),
                   exp1_data, family = cumulative(link="logit", threshold="flexible"),
                   chains=4,cores=4,iter=2000,prior = prior_logit)
save(exp1_rating,file="exp1_rating.Rda")
summary(exp1_rating)

exp1_rating_nested <- brm(bf(rating ~ ab+cd+gneg+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                          exp1_data, family = cumulative(link="logit", threshold="flexible"),
                          chains=4,cores=4,iter=2000,prior = prior_logit)
save(exp1_rating_nested,file="exp1_rating_nested.Rda")
summary(exp1_rating_nested)

load("exp2a_data.Rda")

trials_2a <- data.frame(exp2a_data %>% filter(rating_rt<10000&rating_rt>150) %>%
                          group_by(subj,item,gneg,aneg,pcu,condition,pcu_hl,triali,resa,ab,cd) %>%
                          dplyr::summarise(rating=mean(rating),rating_rt=mean(rating_rt),strt=sum(strt),
                                           stfp=mean(stfp),strr=mean(strr)))

curr <- filter(exp2a_data,id==1 & firsfp==1 & firrdt>80)
fp_id1 <- brm(bf(firrdt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(fp_id1,file="fp_id1.Rda")
summary(fp_id1)

curr <- filter(exp2a_data,id==2 & firsfp==1 & firrdt>80)
fp_id2 <- brm(bf(firrdt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(fp_id2,file="fp_id2.Rda")
summary(fp_id2)

curr <- filter(exp2a_data,id==3 & firsfp==1 & firrdt>80)
fp_id3 <- brm(bf(firrdt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(fp_id3,file="fp_id3.Rda")
summary(fp_id3)

curr <- filter(exp2a_data,id==2 & firsfp==1 & regrpd>80)
rp_id2 <- brm(bf(regrpd ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(rp_id2,file="rp_id2.Rda")
summary(rp_id2)

curr <- filter(exp2a_data,id==3 & firsfp==1 & regrpd>80)
rp_id3 <- brm(bf(regrpd ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
```

```
              iter=2000,prior = prior_log)
save(rp_id3,file="rp_id3.Rda")
summary(rp_id3)

curr <- filter(exp2a_data,id==1 & dwellt>80)
tt_id1 <- brm(bf(dwellt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(tt_id1,file="tt_id1.Rda")
summary(tt_id1)

curr <- filter(exp2a_data,id==2 & dwellt>80)
tt_id2 <- brm(bf(dwellt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(tt_id2,file="tt_id2.Rda")
summary(tt_id2)

curr <- filter(exp2a_data,id==3 & dwellt>80)
tt_id3 <- brm(bf(dwellt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              curr, family = lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(tt_id3,file="tt_id3.Rda")
summary(tt_id3)

et_rat_nested <- brm(bf(rating_rt ~ (ab+cd+gneg)*pcu+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                     trials_2a, family = shifted_lognormal(), chains = 4, cores = 4,
                     iter=2000,prior = prior_log)
save(et_rat_nested,file="et_rat_nested.Rda")
summary(et_rat_nested)

et_rat <- brm(bf(rating_rt ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
              trials_2a, family = shifted_lognormal(), chains = 4, cores = 4,
              iter=2000,prior = prior_log)
save(et_rat,file="et_rat.Rda")
summary(et_rat)

et_rating_nested <- brm(bf(rating ~ (ab+cd+gneg)*pcu+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                        trials_2a, family = cumulative(link="logit", threshold="flexible"),
                        chains=4,cores=4,iter=2000,prior = prior_logit)
save(et_rating_nested,file="et_rating_nested.Rda")
summary(et_rating_nested)

et_rating <- brm(bf(rating ~ gneg*aneg*pcu+(gneg*aneg|subj)+(gneg*aneg|item)),
                 trials_2a, family = cumulative(link="logit", threshold="flexible"),
                 chains=4,cores=4,iter=2000,prior = prior_logit)
save(et_rating,file="et_rating.Rda")
summary(et_rating)

curr<-subset(trials_2a,condition%in%c("a","c","d"))
curr$condition<-factor(curr$condition)
(contrasts(curr$condition)<-contr.treatment(3,base=3))

et_rat_resa <- brm(bf(rating_rt ~ (resa+pcu)*condition+(condition*resa|subj)+(condition|item)),
                   curr, family = shifted_lognormal(), chains=4,cores=4,
                   iter=2000,prior = prior_log)
```

```
save(et_rat_resa,file="et_rat_resa.Rda")
summary(et_rat_resa)

et_rating_resa <- brm(bf(rating ~ (resa+pcu)*condition+(condition*resa|subj)+(condition|item)),
                      curr, family = cumulative(link="logit", threshold="flexible"),
                      chains=4,cores=4,iter=2000,prior = prior_logit)
save(et_rating_resa,file="et_rating_resa.Rda")
summary(et_rating_resa)

curr <- filter(trials_2a,condition=="d")
et_rating_resa_nested1 <- brm(bf(rating ~ resa+pcu+(resa|subj)+(1|item)),
                              curr, family = cumulative(link="logit", threshold="flexible"),
                              chains=4,cores=4,iter=2000,prior = prior_logit)
save(et_rating_resa_nested1,file="et_rating_resa_nested1.Rda")
summary(et_rating_resa_nested1)

curr <- filter(trials_2a,condition=="c")
et_rating_resa_nested2 <- brm(bf(rating ~ resa+pcu+(resa|subj)+(1|item)),
                              curr, family = cumulative(link="logit", threshold="flexible"),
                              chains=4,cores=4,iter=2000,prior = prior_logit)
save(et_rating_resa_nested2,file="et_rating_resa_nested2.Rda")
summary(et_rating_resa_nested2)

curr <- filter(exp2a_data,id==3 & firsfp==1 & regrpd>80 & condition%in%c("a","c","d"))
curr$condition<-factor(curr$condition)
contrasts(curr$condition)<-contr.treatment(3,base=3)

et_rpd_resa <- brm(bf(regrpd ~ (resa+pcu)*condition+(condition*resa|subj)+(condition|item)),
                   curr, family = lognormal(), chains = 4, cores = 4,
                   iter=2000,prior = prior_log)
save(et_rpd_resa,file="et_rpd_resa.Rda")
summary(et_rpd_resa)

curr <- filter(exp2a_data,id==3 & regrpd>80)
rp_id3_nested <- brm(bf(regrpd ~ (ab+cd+gneg)*pcu+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                     curr, family = lognormal(), chains = 4, cores = 4,
                     iter=2000,prior = prior_log)
save(rp_id3_nested,file="rp_id3_nested.Rda")
summary(rp_id3_nested)

curr <- filter(exp2a_data,id==1 & dwellt>80)
tt_id1_nested <- brm(bf(dwellt ~ (ab+cd+gneg)*pcu+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                     curr, family = lognormal(), chains = 4, cores = 4,
                     iter=2000,prior = prior_log)
save(tt_id1_nested,file="tt_id1_nested.Rda")
summary(tt_id1_nested)

curr <- filter(exp2a_data,id==2 & dwellt>80)
tt_id2_nested <- brm(bf(dwellt ~ (ab+cd+gneg)*pcu+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
                     curr, family = lognormal(), chains = 4, cores = 4,
                     iter=2000,prior = prior_log)
save(tt_id2_nested,file="tt_id2_nested.Rda")
summary(tt_id2_nested)

curr <- filter(exp2a_data,id==3 & dwellt>80)
tt_id3_nested <- brm(bf(dwellt ~ (ab+cd+gneg)*pcu+(ab+cd+gneg|subj)+(ab+cd+gneg|item)),
```

```
                    curr, family = lognormal(), chains = 4, cores = 4,
                    iter=2000,prior = prior_log)
save(tt_id3_nested,file="tt_id3_nested.Rda")
summary(tt_id3_nested)

load("exp3_data.Rda")

exp3_data <- filter(exp3_data,rat_t<10000&rat_t>150)

(contrasts(exp3_data$constr)<-contr.treatment(3,base=3))

dc2_rat <- brm(bf(rating ~ neg*constr+(neg*constr|subj)+(neg*constr|item)),
               exp3_data, family = cumulative(link="logit", threshold="flexible"),
               chains=4,cores=4,iter=2000,prior = prior_logit)
save(dc2_rat,file="dc2_rat.Rda")
summary(dc2_rat)

dc2_read <- brm(bf(read_t ~ neg*constr+(neg*constr|subj)+(neg*constr|item)),
                exp3_data, family = shifted_lognormal(), chains=4,cores=4,
                iter=2000,prior = prior_log)
save(dc2_read,file="dc2_read.Rda")
summary(dc2_read)

dc2_ratime <- brm(bf(rat_t ~ neg*constr+(neg*constr|subj)+(neg*constr|item)),
                  exp3_data, family = shifted_lognormal(), chains=4,cores=4,
                  iter=2000,prior = prior_log)
save(dc2_ratime,file="dc2_ratime.Rda")
summary(dc2_ratime)

curr <- filter(exp3_data,constr=="sothat")
dc2_read_nested1 <- brm(bf(read_t ~ neg+(neg|subj)+(neg|item)),
                        curr, family = shifted_lognormal(), chains=4,cores=1,
                        iter=2000,prior = prior_log)
save(dc2_read_nested1,file="dc2_read_nested1.Rda")

dc2_rat_nested1 <- brm(bf(rating ~ neg+(neg|subj)+(neg|item)),
                       curr, family = cumulative(link="logit", threshold="flexible"),
                       chains=4,cores=1,iter=2000,prior = prior_logit)
save(dc2_rat_nested1,file="dc2_rat_nested1.Rda")

curr <- filter(exp3_data,constr=="tooto")
dc2_read_nested2 <- brm(bf(read_t ~ neg+(neg|subj)+(neg|item)),
                        curr, family = shifted_lognormal(), chains=4,cores=1,
                        iter=2000,prior = prior_log)
save(dc2_read_nested2,file="dc2_read_nested2.Rda")

dc2_rat_nested2 <- brm(bf(rating ~ neg+(neg|subj)+(neg|item)),
                       curr, family = cumulative(link="logit", threshold="flexible"),
                       chains=4,cores=1,iter=2000,prior = prior_logit)
save(dc2_rat_nested2,file="dc2_rat_nested2.Rda")

load("exp4_data.Rda")

comp_read <- brm(bf(read_t ~ cond+(cond|subj)+(cond|item)),
                 exp4_data, family = shifted_lognormal(), chains=4,cores=4,
                 iter=2000,prior = prior_log)
```

```
save(comp_read,file="comp_read.Rda")
summary(comp_read)

exp4_data$len<-scale(exp4_data$len)
comp_compt <- brm(bf(comp_t ~ cond+len+(cond|subj)+(cond|item)),
                  exp4_data, family = shifted_lognormal(), chains=4,cores=4,
                  iter=2000,prior = prior_log)
save(comp_compt,file="comp_compt.Rda")
summary(comp_compt)

comp_s1 <- brm(bf(judgment1~cond+(cond|subj)+(cond|item)+(cond|coder1)
                  +(1|coder1:item)+(1|coder1:subj)),
               family=bernoulli(),data=exp4_data, chains=4,cores=4,
               iter=2000,prior = prior_log)
save(comp_s1,file="comp_s1.Rda")
summary(comp_s1)

comp_s2 <- brm(bf(judgment2~cond+(cond|subj)+(cond|item)+(cond|coder2)
                  +(1|coder2:item)+(1|coder2:subj)),
               family=bernoulli(),data=exp4_data, chains=4,cores=4,
               iter=2000,prior = prior_log)
save(comp_s2,file="comp_s2.Rda")
summary(comp_s2)
```

Table 1: **Experiment 1: Reading times**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	8.66	0.10	8.46	8.85	≈1.00
gneg	0.08	0.02	0.04	0.14	≈1.00
aneg	0.07	0.02	0.03	0.11	≈1.00
gneg:aneg	−0.06	0.02	−0.11	−0.02	0.01

gneg: Global negation, aneg: Adj. negation

Table 2: **Experiment 1: Reading times (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	8.66	0.10	8.45	8.86	≈1.00
ab	0.00	0.03	−0.06	0.06	0.54
cd	0.13	0.03	0.07	0.20	≈1.00
gneg	0.08	0.02	0.03	0.13	≈1.00

ab: a vs b, cd: c vs d, gneg: Global negation

Table 3: **Experiment 1: Rating times**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.38	0.20	5.99	6.76	≈1.00
gneg	0.01	0.04	−0.07	0.10	0.60
aneg	0.06	0.04	−0.02	0.13	0.93
gneg:aneg	−0.09	0.05	−0.19	0.00	0.03

gneg: Global negation, aneg: Adj. negation

Table 4: **Experiment 1: Rating times (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	6.37	0.19	6.00	6.76	≈1.00
ab	−0.04	0.06	−0.16	0.08	0.25
cd	0.15	0.06	0.03	0.27	0.99
gneg	0.01	0.04	−0.07	0.10	0.60

ab: a vs b, cd: c vs d, gneg: Global negation

Table 5: **Experiment 1: Ratings**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept[1]	−3.03	0.33	−3.67	−2.39	≈0.00
Intercept[2]	−1.90	0.32	−2.53	−1.29	≈0.00
Intercept[3]	−0.90	0.31	−1.51	−0.27	≈0.00
Intercept[4]	−0.15	0.31	−0.76	0.47	0.31
Intercept[5]	0.74	0.31	0.14	1.37	0.99
Intercept[6]	2.01	0.32	1.40	2.68	≈1.00
gneg	−0.24	0.12	−0.48	0.00	0.03
aneg	−0.60	0.17	−0.94	−0.28	≈0.00
gneg:aneg	1.34	0.24	0.88	1.82	≈1.00

gneg: Global negation, aneg: Adj. negation

Table 6: **Experiment 1: Ratings (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept[1]	−3.02	0.34	−3.68	−2.35	≈0.00
Intercept[2]	−1.89	0.32	−2.52	−1.24	≈0.00
Intercept[3]	−0.89	0.31	−1.51	−0.28	≈0.00
Intercept[4]	−0.15	0.31	−0.76	0.48	0.31
Intercept[5]	0.74	0.31	0.14	1.37	0.99
Intercept[6]	2.01	0.32	1.38	2.67	≈1.00
ab	0.72	0.19	0.36	1.10	≈1.00
cd	−1.90	0.38	−2.64	−1.15	≈0.00
gneg	−0.24	0.13	−0.49	0.02	0.03

ab: a vs b, cd: c vs d, gneg: Global negation

Table 7: **Experiment 2A: First-pass reading times, region 1**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	6.76	0.07	6.62	6.89	≈1.00
gneg	−0.12	0.01	−0.14	−0.09	≈0.00
aneg	0.00	0.01	−0.02	0.02	0.55
pcu	0.02	0.04	−0.06	0.08	0.67
gneg:aneg	−0.01	0.01	−0.03	0.00	0.07
gneg:pcu	0.00	0.01	−0.02	0.02	0.35
aneg:pcu	−0.01	0.01	−0.02	0.01	0.16
gneg:aneg:pcu	0.01	0.01	−0.01	0.02	0.80

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 8: **Experiment 2A: First-pass reading times, region 2**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.33	0.04	6.24	6.41	≈1.00
gneg	0.00	0.01	−0.03	0.02	0.38
aneg	0.08	0.02	0.04	0.12	≈1.00
pcu	0.01	0.04	−0.06	0.08	0.66
gneg:aneg	0.01	0.01	−0.01	0.03	0.86
gneg:pcu	0.00	0.01	−0.02	0.02	0.39
aneg:pcu	0.02	0.01	0.00	0.05	0.98
gneg:aneg:pcu	0.00	0.01	−0.01	0.03	0.69

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 9: **Experiment 2A: First-pass reading times, region 3**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.72	0.05	6.63	6.82	≈1.00
gneg	−0.01	0.01	−0.03	0.02	0.27
aneg	−0.01	0.01	−0.03	0.02	0.32
pcu	0.00	0.04	−0.08	0.09	0.53
gneg:aneg	−0.01	0.01	−0.04	0.02	0.21
gneg:pcu	0.00	0.01	−0.02	0.03	0.55
aneg:pcu	0.01	0.01	−0.02	0.03	0.69
gneg:aneg:pcu	0.00	0.01	−0.02	0.03	0.62

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 10: **Experiment 2A: Regression-path durations, region 2**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.49	0.05	6.40	6.58	≈1.00
gneg	0.01	0.01	−0.01	0.03	0.82
aneg	0.10	0.02	0.07	0.14	≈1.00
pcu	0.02	0.04	−0.05	0.09	0.72
gneg:aneg	−0.01	0.01	−0.03	0.01	0.11
gneg:pcu	0.00	0.01	−0.02	0.02	0.62
aneg:pcu	0.02	0.01	0.00	0.04	0.99
gneg:aneg:pcu	−0.01	0.01	−0.02	0.01	0.26

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 11: **Experiment 2A: Regression-path durations, region 3**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	7.65	0.06	7.53	7.77	≈1.00
gneg	0.10	0.02	0.05	0.15	≈1.00
aneg	0.09	0.01	0.06	0.11	≈1.00
pcu	0.10	0.06	−0.02	0.21	0.96
gneg:aneg	−0.11	0.02	−0.15	−0.08	≈0.00
gneg:pcu	0.01	0.02	−0.03	0.04	0.64
aneg:pcu	−0.01	0.01	−0.04	0.01	0.21
gneg:aneg:pcu	−0.01	0.02	−0.04	0.02	0.29

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 12: **Experiment 2A: Regression-path durations, region 3 (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	7.64	0.06	7.53	7.77	≈1.00
ab	−0.03	0.02	−0.07	0.01	0.09
cd	0.20	0.02	0.15	0.24	≈1.00
gneg	0.10	0.02	0.05	0.15	≈1.00
pcu	0.10	0.06	−0.02	0.21	0.95
ab:pcu	−0.02	0.02	−0.06	0.02	0.17
cd:pcu	0.00	0.02	−0.04	0.04	0.46
gneg:pcu	0.01	0.02	−0.03	0.04	0.64

ab: a vs b, cd: c vs d, gneg: Global negation
pcu: Working memory capacity

Table 13: **Experiment 2A: Total reading times, region 1**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	7.11	0.08	6.97	7.26	≈1.00
gneg	−0.05	0.02	−0.08	−0.02	≈0.00
aneg	0.03	0.01	0.01	0.05	≈1.00
pcu	0.05	0.04	−0.03	0.14	0.88
gneg:aneg	−0.05	0.01	−0.07	−0.02	≈0.00
gneg:pcu	0.00	0.01	−0.03	0.03	0.56
aneg:pcu	0.00	0.01	−0.02	0.01	0.36
gneg:aneg:pcu	0.00	0.01	−0.03	0.02	0.33

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 14: **Experiment 2A: Total reading times, region 1 (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	7.10	0.08	6.95	7.25	≈1.00
ab	−0.02	0.02	−0.05	0.01	0.13
cd	0.07	0.01	0.04	0.10	≈1.00
gneg	−0.05	0.02	−0.08	−0.02	≈0.00
pcu	0.05	0.04	−0.03	0.14	0.89
ab:pcu	−0.01	0.01	−0.04	0.02	0.30
cd:pcu	0.00	0.01	−0.02	0.03	0.55
gneg:pcu	0.00	0.01	−0.03	0.03	0.57

ab: a vs b, cd: c vs d, gneg: Global negation
pcu: Working memory capacity

Table 15: **Experiment 2A: Total reading times, region 2**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	6.99	0.06	6.88	7.11	≈1.00
gneg	0.11	0.02	0.06	0.15	≈1.00
aneg	0.14	0.02	0.10	0.19	≈1.00
pcu	0.07	0.05	−0.03	0.18	0.92
gneg:aneg	−0.09	0.02	−0.13	−0.06	≈0.00
gneg:pcu	0.03	0.02	0.00	0.06	0.95
aneg:pcu	0.00	0.01	−0.02	0.03	0.60
gneg:aneg:pcu	−0.01	0.01	−0.04	0.02	0.19

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 16: **Experiment 2A: Total reading times, region 2 (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.99	0.06	6.88	7.10	≈1.00
ab	0.05	0.03	−0.01	0.11	0.96
cd	0.24	0.03	0.18	0.29	≈1.00
gneg	0.11	0.02	0.06	0.15	≈1.00
pcu	0.07	0.05	−0.02	0.18	0.93
ab:pcu	−0.01	0.02	−0.05	0.03	0.28
cd:pcu	0.02	0.02	−0.02	0.06	0.79
gneg:pcu	0.03	0.02	0.00	0.06	0.95

ab: a vs b, cd: c vs d, gneg: Global negation
pcu: Working memory capacity

Table 17: **Experiment 2A: Total reading times, region 3**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	7.26	0.04	7.17	7.35	≈1.00
gneg	0.05	0.02	0.02	0.08	≈1.00
aneg	0.05	0.01	0.03	0.07	≈1.00
pcu	0.06	0.04	−0.01	0.14	0.95
gneg:aneg	−0.07	0.01	−0.10	−0.05	≈0.00
gneg:pcu	0.00	0.01	−0.02	0.02	0.54
aneg:pcu	−0.01	0.01	−0.03	0.01	0.24
gneg:aneg:pcu	−0.01	0.01	−0.03	0.02	0.30

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 18: **Experiment 2A: Total reading times, region 3 (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	7.26	0.05	7.17	7.35	≈1.00
ab	−0.02	0.01	−0.06	0.00	0.04
cd	0.12	0.02	0.08	0.15	≈1.00
gneg	0.05	0.02	0.02	0.08	≈1.00
pcu	0.06	0.04	−0.02	0.14	0.94
ab:pcu	−0.01	0.01	−0.04	0.01	0.17
cd:pcu	0.00	0.02	−0.04	0.03	0.48
gneg:pcu	0.00	0.01	−0.02	0.02	0.55

ab: a vs b, cd: c vs d, gneg: Global negation
pcu: Working memory capacity

Table 19: **Experiment 2A: Rating times**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.90	0.05	6.79	7.00	≈1.00
gneg	0.04	0.01	0.02	0.07	≈1.00
aneg	0.01	0.01	−0.01	0.04	0.79
pcu	0.06	0.05	−0.04	0.16	0.90
gneg:aneg	−0.08	0.02	−0.12	−0.05	≈0.00
gneg:pcu	0.01	0.01	−0.01	0.04	0.82
aneg:pcu	0.01	0.01	−0.02	0.03	0.70
gneg:aneg:pcu	0.01	0.01	−0.02	0.04	0.74

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 20: **Experiment 2A: Rating times (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	6.90	0.05	6.79	7.01	≈1.00
ab	−0.07	0.02	−0.12	−0.03	≈0.00
cd	0.10	0.02	0.05	0.14	≈1.00
gneg	0.04	0.01	0.02	0.07	≈1.00
pcu	0.06	0.05	−0.03	0.16	0.91
ab:pcu	0.02	0.02	−0.02	0.05	0.82
cd:pcu	0.00	0.02	−0.04	0.03	0.46
gneg:pcu	0.01	0.01	−0.01	0.04	0.81

ab: a vs b, cd: c vs d, gneg: Global negation
pcu: Working memory capacity

Table 21: **Experiment 2A: Ratings**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept[1]	−3.66	0.20	−4.05	−3.27	≈0.00
Intercept[2]	−2.28	0.18	−2.63	−1.92	≈0.00
Intercept[3]	−1.43	0.18	−1.77	−1.08	≈0.00
Intercept[4]	−0.87	0.17	−1.21	−0.52	≈0.00
Intercept[5]	−0.12	0.17	−0.46	0.23	0.24
Intercept[6]	1.18	0.17	0.84	1.53	≈1.00
gneg	−0.11	0.08	−0.26	0.05	0.08
aneg	−0.64	0.08	−0.79	−0.48	≈0.00
pcu	−0.22	0.15	−0.52	0.06	0.06
gneg:aneg	1.32	0.14	1.05	1.61	≈1.00
gneg:pcu	−0.04	0.06	−0.15	0.08	0.25
aneg:pcu	−0.06	0.05	−0.16	0.04	0.14
gneg:aneg:pcu	0.23	0.09	0.05	0.40	0.99

gneg: Global negation, aneg: Adj. negation
pcu: Working memory capacity

Table 22: **Experiment 2A: Ratings (nested contrasts)**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept[1]	−3.68	0.20	−4.10	−3.28	≈0.00
Intercept[2]	−2.29	0.18	−2.65	−1.93	≈0.00
Intercept[3]	−1.45	0.18	−1.80	−1.09	≈0.00
Intercept[4]	−0.88	0.18	−1.23	−0.53	≈0.00
Intercept[5]	−0.14	0.18	−0.48	0.21	0.22
Intercept[6]	1.16	0.18	0.82	1.52	≈1.00
ab	0.68	0.12	0.44	0.92	≈1.00
cd	−1.98	0.19	−2.36	−1.60	≈0.00
gneg	−0.12	0.08	−0.28	0.04	0.07
pcu	−0.22	0.15	−0.52	0.08	0.08
ab:pcu	0.16	0.09	−0.02	0.34	0.96
cd:pcu	−0.29	0.12	−0.53	−0.06	0.01
gneg:pcu	−0.04	0.06	−0.15	0.08	0.25

ab: a vs b, cd: c vs d, gneg: Global negation
pcu: Working memory capacity

Table 23: **Experiment 2A: RPDs ID3, approval**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	7.35	0.06	7.24	7.46	≈1.00
resa	−0.09	0.03	−0.16	−0.02	≈0.00
pcu	0.09	0.05	0.00	0.19	0.98
dneg	0.37	0.06	0.26	0.48	≈1.00
adjneg	0.40	0.05	0.31	0.49	≈1.00
resa:dneg	0.04	0.05	−0.05	0.13	0.83
resa:adjneg	0.03	0.04	−0.04	0.10	0.78
pcu:dneg	−0.01	0.05	−0.11	0.08	0.40
pcu:adjneg	0.00	0.04	−0.09	0.08	0.49

resa: Residual approval, pcu: Working memory capacity
dneg: Double negation, adjneg: Adjectival negation

Table 24: **Experiment 2A: Rating times, approval**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	6.77	0.06	6.65	6.88	≈1.00
resa	−0.04	0.03	−0.09	0.01	0.07
pcu	0.05	0.05	−0.05	0.14	0.82
dneg	0.11	0.04	0.03	0.18	≈1.00
adjneg	0.19	0.04	0.11	0.27	≈1.00
resa:dneg	−0.01	0.04	−0.08	0.05	0.33
resa:adjneg	0.05	0.04	−0.03	0.14	0.89
pcu:dneg	0.03	0.04	−0.04	0.10	0.83
pcu:adjneg	0.00	0.04	−0.07	0.06	0.45

resa: Residual approval, pcu: Working memory capacity
dneg: Double negation, adjneg: Adjectival negation

Table 25: **Experiment 2A: Ratings, approval**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept[1]	−5.77	0.31	−6.38	−5.16	≈0.00
Intercept[2]	−4.25	0.29	−4.83	−3.70	≈0.00
Intercept[3]	−3.42	0.28	−4.00	−2.88	≈0.00
Intercept[4]	−2.90	0.28	−3.46	−2.35	≈0.00
Intercept[5]	−2.15	0.28	−2.70	−1.63	≈0.00
Intercept[6]	−0.80	0.27	−1.34	−0.28	≈0.00
resa	0.58	0.23	0.14	1.04	≈1.00
pcu	0.11	0.19	−0.27	0.48	0.72
dneg	−1.41	0.26	−1.91	−0.92	≈0.00
adjneg	−3.83	0.34	−4.49	−3.17	≈0.00
resa:dneg	−0.17	0.24	−0.66	0.31	0.24
resa:adjneg	−0.71	0.28	−1.26	−0.16	0.01
pcu:dneg	−0.20	0.16	−0.53	0.12	0.11
pcu:adjneg	−0.59	0.24	−1.08	−0.12	≈0.00

resa: Residual approval, pcu: Working memory capacity
dneg: Double negation, adjneg: Adjectival negation

Table 26: **Experiment 2A: Ratings, approval, Adj. negation condition**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept[1]	−2.00	0.25	−2.51	−1.53	≈0.00
Intercept[2]	−0.43	0.22	−0.87	0.01	0.03
Intercept[3]	0.38	0.22	−0.06	0.82	0.95
Intercept[4]	0.88	0.23	0.43	1.31	≈1.00
Intercept[5]	1.70	0.24	1.22	2.16	≈1.00
Intercept[6]	2.85	0.26	2.34	3.39	≈1.00
resa	−0.12	0.11	−0.33	0.10	0.13
pcu	−0.46	0.21	−0.87	−0.05	0.01

resa: Residual approval, pcu: Working memory capacity

Table 27: **Experiment 3: Reading times**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	8.67	0.06	8.54	8.79	≈1.00
neg	0.02	0.03	−0.04	0.07	0.73
tooas	0.11	0.03	0.05	0.17	≈1.00
sothat	0.17	0.03	0.11	0.24	≈1.00
neg:tooas	−0.02	0.03	−0.08	0.04	0.24
neg:sothat	0.08	0.03	0.02	0.15	0.99

neg: Negation, tooas: too...as construction
sothat: so...that construction

Table 28: **Experiment 3: Reading times, so...that**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	8.74	0.07	8.59	8.89	≈1.00
neg	0.11	0.04	0.04	0.18	≈1.00

neg: Negation

Table 29: **Experiment 3: Rating times**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept	6.23	0.12	5.98	6.46	≈1.00
neg	−0.04	0.03	−0.10	0.02	0.14
tooas	−0.06	0.05	−0.15	0.04	0.13
sothat	−0.05	0.04	−0.14	0.03	0.11
neg:tooas	0.00	0.04	−0.09	0.08	0.45
neg:sothat	−0.02	0.04	−0.10	0.06	0.33

neg: Negation, tooas: too...as construction
sothat: so...that construction

Table 30: **Experiment 3: Ratings**

	Estimate	Est.Error	Q2.5	Q97.5	Pr($\beta > 0$)
Intercept[1]	−3.13	0.20	−3.53	−2.75	≈0.00
Intercept[2]	−1.54	0.18	−1.90	−1.19	≈0.00
Intercept[3]	−0.58	0.18	−0.94	−0.23	≈0.00
Intercept[4]	0.17	0.18	−0.19	0.52	0.84
Intercept[5]	0.86	0.18	0.51	1.22	≈1.00
Intercept[6]	2.06	0.19	1.68	2.43	≈1.00
neg	1.26	0.14	0.98	1.54	≈1.00
tooas	−0.28	0.13	−0.52	−0.03	0.02
sothat	−1.19	0.18	−1.55	−0.86	≈0.00
neg:tooas	−0.01	0.13	−0.26	0.23	0.47
neg:sothat	−0.95	0.14	−1.24	−0.68	≈0.00

neg: Negation, tooas: too...as construction
sothat: so...that construction

Table 31: **Experiment 3: Ratings, so...that**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept[1]	−2.18	0.31	−2.80	−1.60	≈0.00
Intercept[2]	−0.41	0.29	−0.99	0.16	0.07
Intercept[3]	0.64	0.29	0.08	1.23	0.99
Intercept[4]	1.46	0.30	0.88	2.05	≈1.00
Intercept[5]	2.31	0.30	1.72	2.92	≈1.00
Intercept[6]	3.34	0.33	2.70	3.98	≈1.00
neg	0.29	0.11	0.08	0.51	0.99

neg: Negation

Table 32: **Experiment 4: Reading times**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	8.31	0.09	8.13	8.48	≈1.00
dneg	0.02	0.03	−0.04	0.08	0.73

dneg: Double negation

Table 33: **Experiment 4: Completion times**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	8.69	0.04	8.63	8.76	≈1.00
dneg	−0.01	0.01	−0.04	0.01	0.17
len	0.34	0.00	0.33	0.35	≈1.00

dneg: Double negation, len: Length

Table 34: **Experiment 4: Completions, Scheme A**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	−1.81	0.35	−2.50	−1.15	≈0.00
dneg	3.25	0.29	2.71	3.87	≈1.00

dneg: Double negation

Table 35: **Experiment 4: Completions, Scheme B**

	Estimate	Est.Error	Q2.5	Q97.5	$\Pr(\beta > 0)$
Intercept	−0.74	0.27	−1.28	−0.21	0.01
dneg	1.73	0.28	1.19	2.28	≈1.00

dneg: Double negation

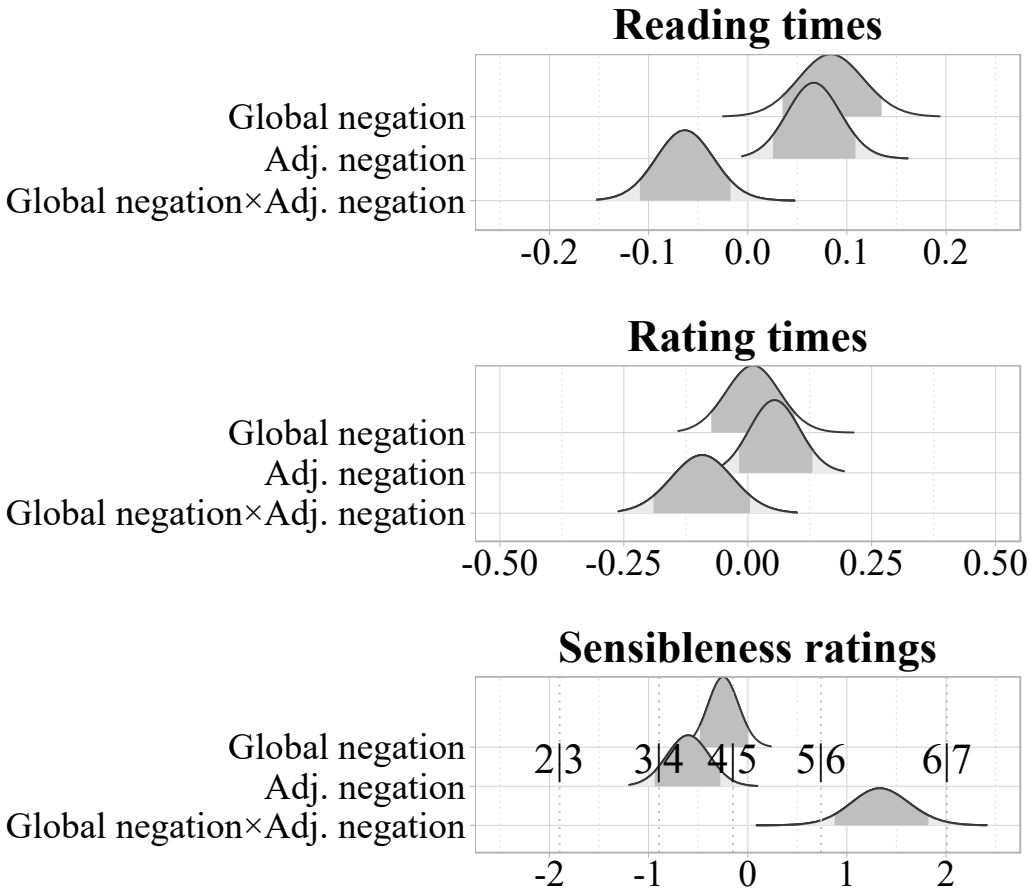


Figure 1: : Experiment 1 – Posterior distributions of the parameters for whole-sentence reading times (log scale), rating times (log scale) and sensibleness ratings (logit scale). Cutpoints also shown for ratings.

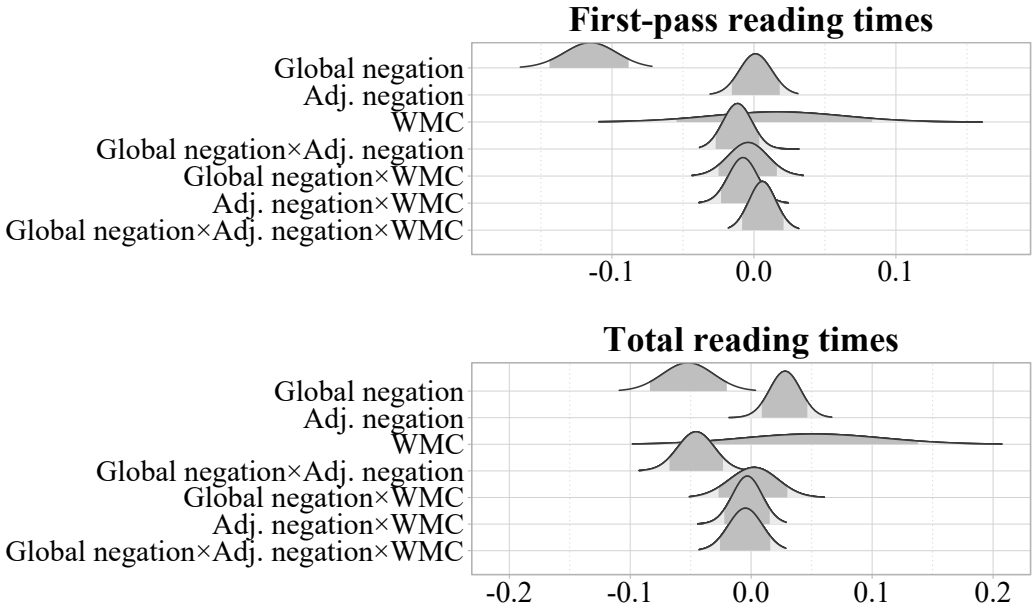


Figure 2: : Experiment 2A – Posterior distributions of the parameters for all reading measures in region 1, extitNo/some head injury ... (log scale). Note that regression-path durations in region 1 are equal to first-pass reading times.

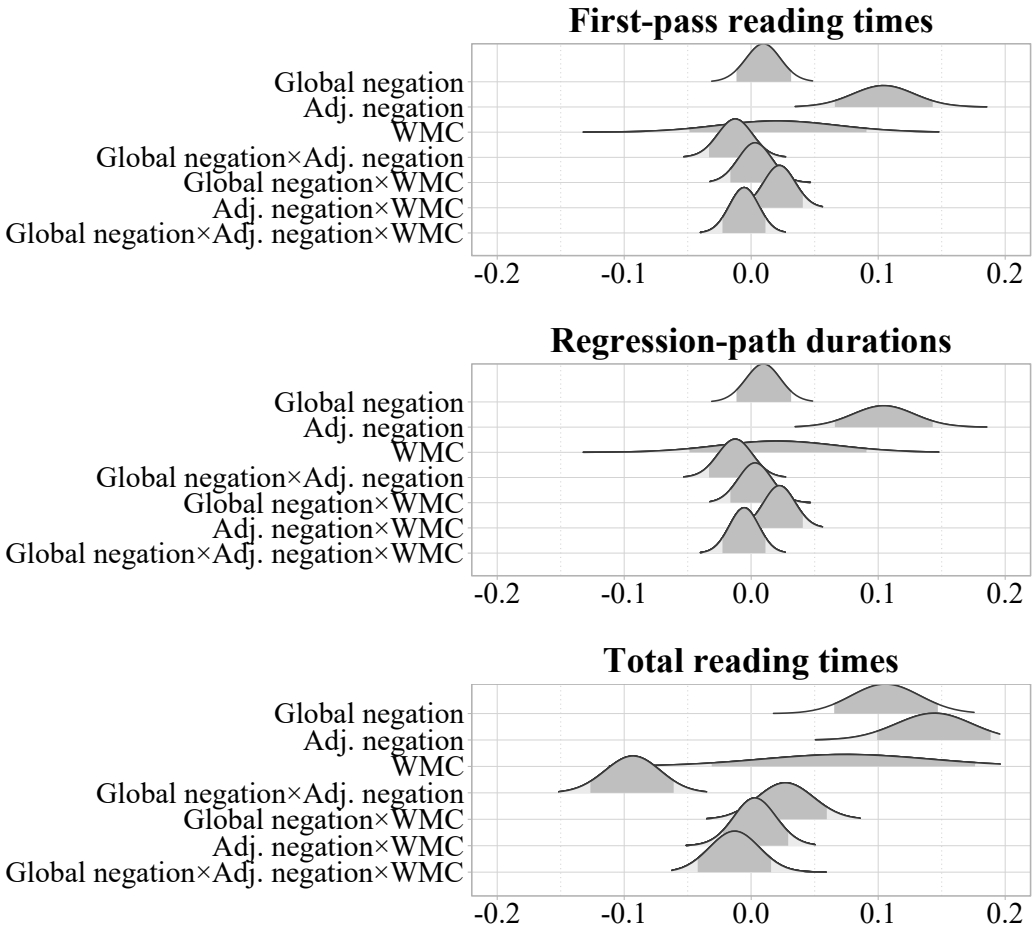


Figure 3: : Experiment 2A – Posterior distributions of the parameters for all reading measures in region 2, extit... is too (un-)dangerous ... (log scale).

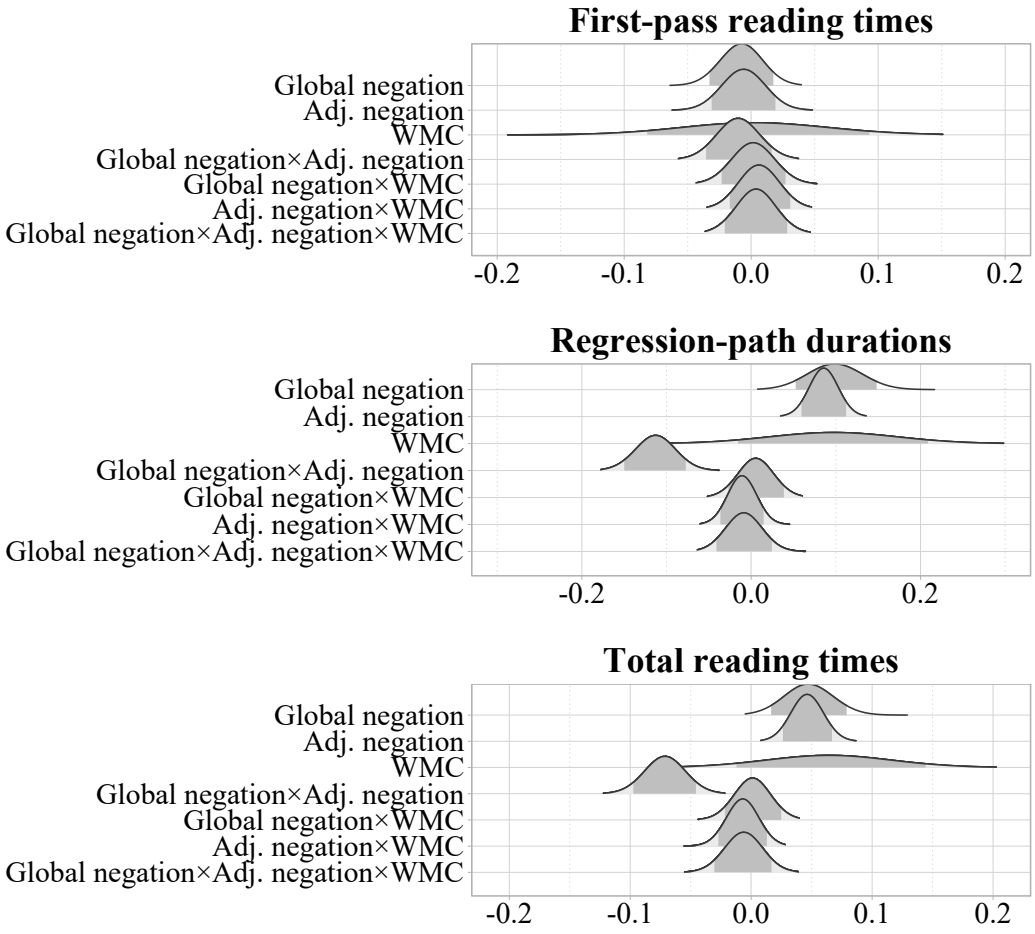


Figure 4: : Experiment 2A – Posterior distributions of the parameters for all reading measures in region 3, extit... to be ignored (log scale). Cutpoints also shown for ratings.

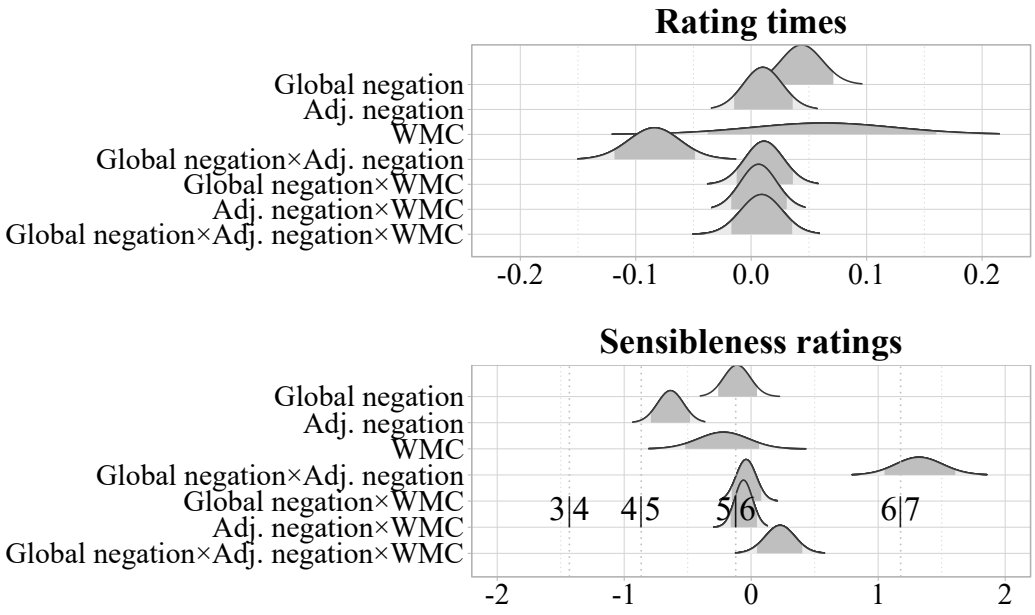


Figure 5: : Experiment 2A – Posterior distributions of the parameters for rating times (log scale) and sensibleness ratings (logit scale). Cutpoints also shown for ratings.

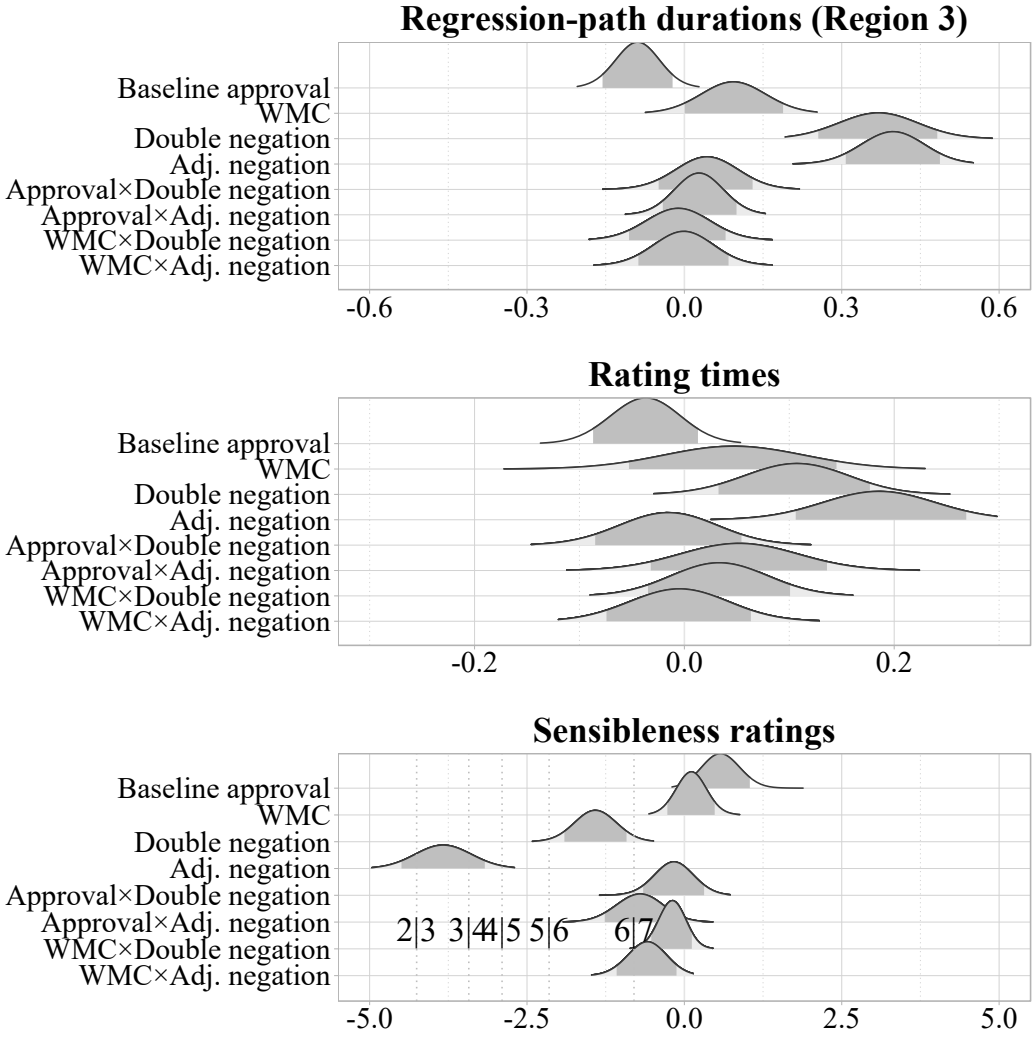


Figure 6: : Experiment 2A – Posterior distributions of the parameters for regression-path durations in region 3 (log scale), rating times (log scale) and sensibleness ratings (logit scale), with item-wise approval measure from Experiment 2B as predictor.

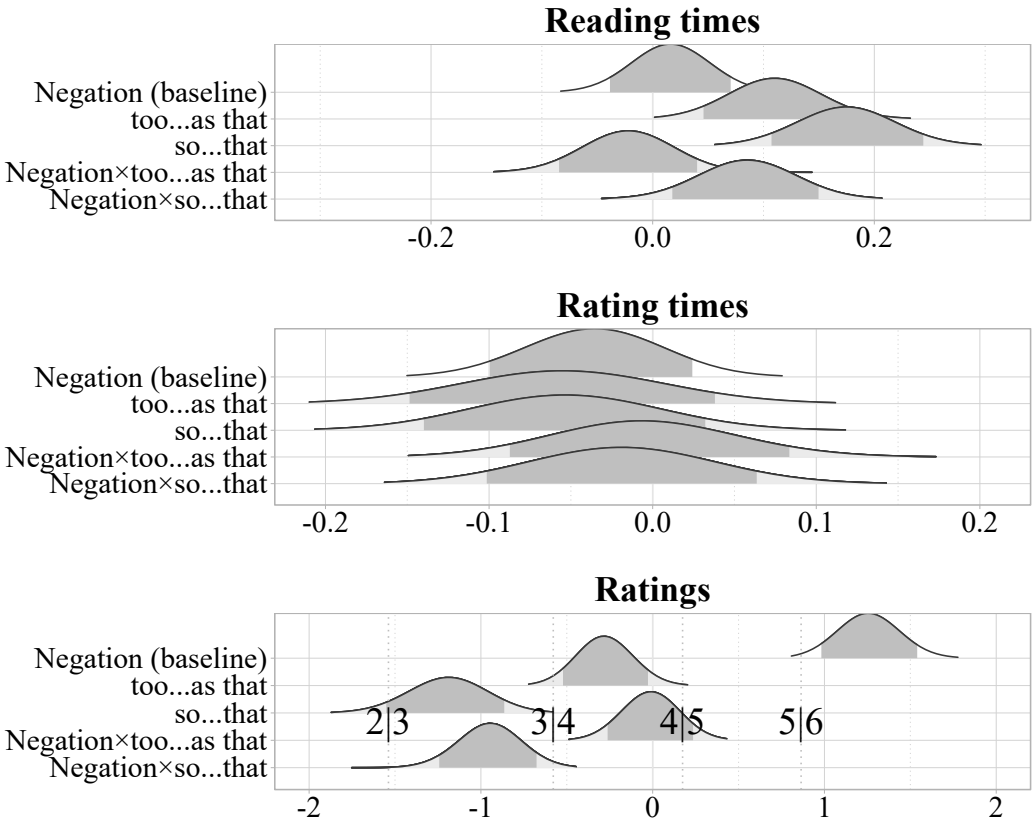


Figure 7: : Experiment 3 – Posterior distributions of the parameters for whole-sentence reading times (log scale), rating times (log scale) and sensibleness ratings (logit scale).

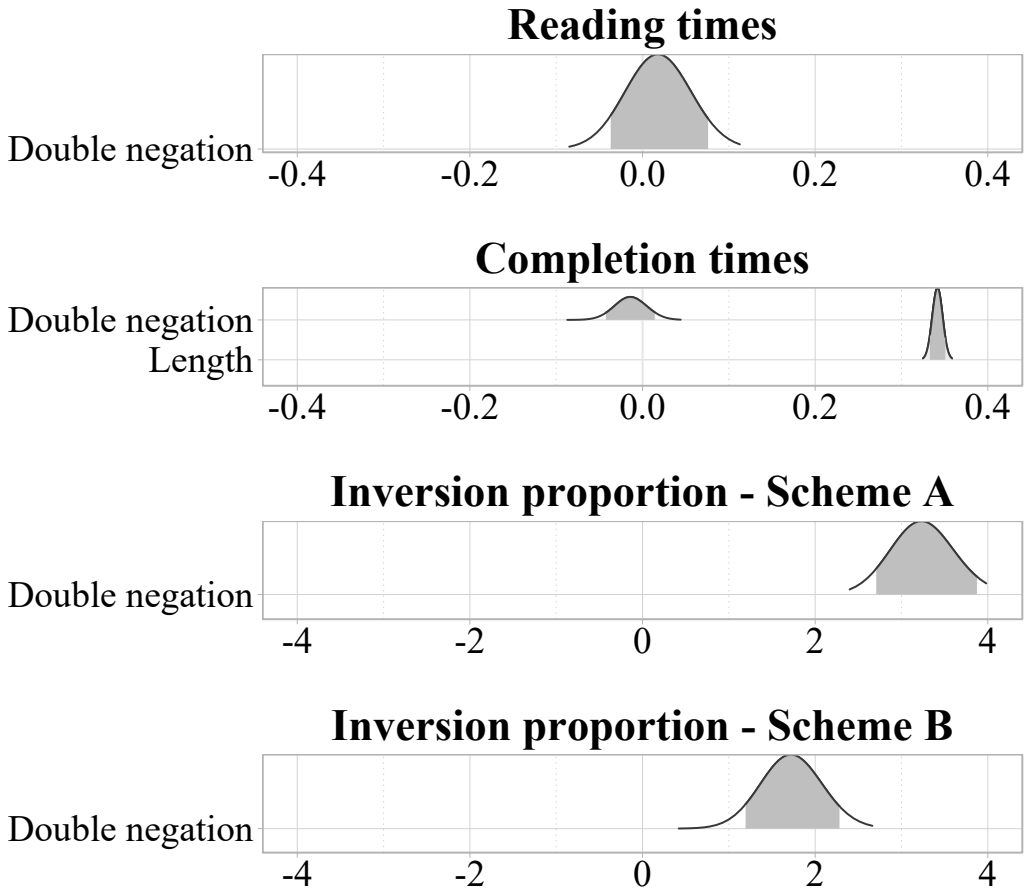


Figure 8: : Experiment 4 – Posterior distributions of the parameters for reading times (log scale), completion times (log scale) and inversion proportion (logit scale).

Appendix B – Detailed procedure and data analysis for Experiments 2A/2B/4

1 Experiment 2A – Procedure details

Eye movements were recorded by an SR Research Eyelink 1000 tracker with a desktop-mounted camera setup at a sampling rate of 1000 Hz. Sentences were presented in 20 pt Times New Roman font on a computer screen set to a resolution of 1680 × 1050 pixels, with subjects sitting at a distance of 60 cm from the screen. A chin rest was used to keep participants' head position stable during tracking. At the beginning of each experimental session the eye tracker was calibrated using a nine-point grid. Each trial started with the presentation of a dot on the monitor which participants had to fixate in order to start the presentation of the stimulus sentence, whose first character then appeared in the same location as the dot. Participants signaled that they were done reading by looking in the lower right corner of the screen for one second.

Each session began with the presentation of two practice items to familiarize subjects with the procedure. The presentation order of the remaining sentences was randomized at runtime. There were three obligatory breaks during the experiment, and subjects were told that they could take additional breaks at any time. The eye-tracker was recalibrated using a nine-point grid after each break and when tracking accuracy fell below acceptable levels. Recording sessions lasted 45 minutes on average.

2 Experiment 2A – Data analysis details

We limited our analysis to three measures out of the many available options because each additional measure analyzed would have increased the risk of a false-positive result (von der Malsburg & Angele, 2017). Furthermore, the three measures we chose should be sufficient to get a window into relatively early processing (first-pass reading times) as well as later processing (total reading times), including regressions to earlier material (regression-path durations), assuming that the timing of the cognitive processes involved maps onto the eye tracking measures in a systematic way (Vasishth et al., 2013).

Bayesian linear mixed-effects models with full random effects structures were fitted to first-pass reading times, regression-path durations, total reading times and rating times. All data points below 80 ms were removed prior to analysis. For rating times, data points above 10 s were also removed. Lognormal distributions were assumed for the reading measures while a shifted lognormal distribution was assumed for the rating times. Sensibleness ratings were analyzed using a cumulative logit model.

Bayesian data analysis requires the specification of priors that represent pre-existing beliefs about likely values for each of the model parameters. Using Bayes' rule, the prior and the data (also called the likelihood) are combined to yield the posterior distribution for each parameter

based on which inferences can be drawn. As deriving the overall posterior distribution is usually computationally intractable (Cooper, 1990), Bayesian inference typically makes use of Markov chain Monte Carlo methods to generate samples from the posterior distribution that eventually yield a close-enough approximation of its shape.

3 Experiment 2B – Data analysis details

Both approval and comprehensibility ratings were transformed from the 1 to 5 scale to a -2 to 2 scale for easier interpretability. Comprehensibility and approval were correlated at the individual observation level ($\hat{r} = 0.36$, 95% confidence interval: [0.30, 0.41]). Most experimental items showed positive mean approval values on the transformed scale (min: -0.06, median: 1.04, max: 1.97), reflecting the fact that the items had been designed to be sensible in the no negation condition. All items also showed positive mean comprehensibility values (min: 0.1, median: 1.73, max: 1.97). Interestingly, among all of our items, the famous *No head injury ...* sentence of Wason & Reich (1979) showed both the highest approval as well as the highest comprehensibility value. The lowest mean approval was assigned to the item *Some special taxes are too profitable to be waived* while the lowest comprehensibility value was assigned to the item *Some target returns are too modest to be missed*.

4 Experiment 4 – Details on coding schemes

Coding Scheme A was based on the original intuition of Wason & Reich (1979) that the verb *ignore* contains an implicit negation, and that this negation triggers meaning inversion. Similarly, Cook & Stevenson (2010) use the notion of "negative" and "positive" verbs. By extension, there should be a class of verbs whose appearance in a completion would signal that inversion has been triggered. The concept of a "negative" verb is by no means well-defined. Kizach et al. ran a pretest to establish whether the verbs in question indicated "deliberate absence of an action" (p. 755; see also Wason & Reich, 1979, p. 595f.), whereas the verbs used in our study were selected on intuitive grounds. Kizach et al. (2015) formalized the original intuition as the verb communicating "deliberate absence of action"; however, our own original items as well as a preliminary look at participants' completions suggest that this may not be the only relevant dimension to inversion-signaling verbs. For instance, in a sentence like *No heirloom is too worthless to be thrown away*, the verb *throw away* does not signal absence of action, but the sentence nevertheless appears to be sensible despite its compositional semantics being incoherent. What seems to be expressed here is that the object in question is of no or little importance and/or interest. We should note that subjective importance has, to our knowledge, not been established as a semantic feature in the linguistic literature; we rely on our own intuitions at this point.

We formalized the notion of subjective importance by instructing a group of nine coders (3 per trial, 3 per randomized list) to judge whether, for instance, the pairing *A heirloom – throw it away* indicated that the thing in question was considered to be of low importance/interest, which would signal the occurrence of inversion. Coders were blind to the experimental manipulation, given that they were not presented with the original preambles.

Coding Scheme B used slightly changed versions of the no negation sentences from Experiment 2A. Specifically, the quantifier phrase *manch eine*, 'some a', was replaced by a demonstrative, as in (1). This change was intended to remove the presence of quantification as a possible source of between-coder variation. Coders in both groups received the same compensation as the subjects for their participation.

(1) **No global negation, no adjectival negation** (NO NEGATION)

d'.	Diese	Kopfverletzung	ist	zu	gefährlich,	um	...
	This	head injury	is	too	dangerous	to	

The completions volunteered by participants for the double negation and adjectival negation conditions were pasted into this template and given to 12 new coders (3 per trial, 4 per randomized list). The coders were instructed to indicate whether the sentences seemed sensible. As before, coders were blind to the experimental manipulation, given that they always saw the same preamble. The reasoning behind this coding scheme was that the double negation sentence attains a meaning similar to that of the no negation sentence under inversion, so that the same continuation should result in a sensible meaning for both. For instance, if the preamble *No head injury is too un-dangerous to ...* produces the continuation *... be ignored*, and this continuation is judged by coders to yield a sensible meaning when combined with (1) above (*This head injury is too dangerous to be ignored*), this would signal the occurrence of inversion. If the produced continuation is the compositionally correct one, *... be treated*, however, the resulting sentence for (1) is not sensible (*This head injury is too dangerous to be treated*). For the adjectival negation condition (*Some head injuries are too un-dangerous ...*), where no inversion should occur, only *... be treated* – or something semantically similar – is ever expected a continuation, so that (1) should always come out as not sensible.

5 Experiment 4 – Individual differences and problematic data points

While the results of the statistical analysis allow the conclusion that the depth charge effect generalizes beyond the particular subjects and items tested in the current study, it is nevertheless informative to look at how each individual subject responded to the manipulation, and see whether there are any experimental items that do not exhibit the depth charge effect. Figures 1 and 2 were generated by computing the posterior difference in inversion proportions between the double and adjectival negation conditions for each subject and item, respectively, based on the random effects structure of the hierarchical logistic regression. Figure 1 shows that the inversion effect is non-negative across participants, meaning that almost everybody in the sample group showed more inversions in the double than in the adjectival negation condition. There is a subset of subjects for whom the difference between condition is almost 1, which would mean no inversions in the adjectival negation condition and 100% inversions in the double negation condition. There is, however, also a subset of subjects (IDs 37, 40 and 51) who show almost zero difference between the conditions, especially under Coding Scheme A, meaning that these participants are largely immune to the depth charge effect.

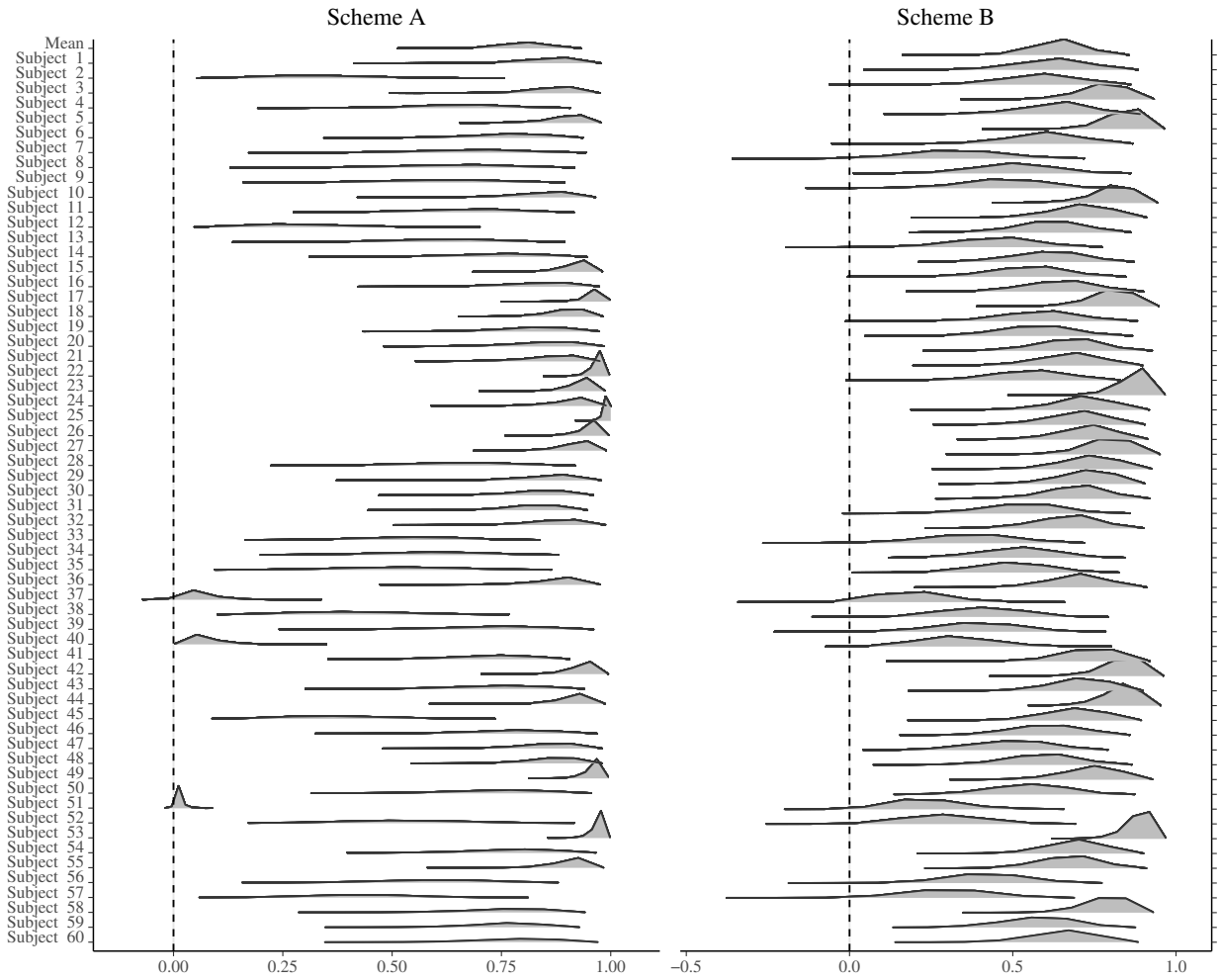


Figure 1: **Experiment 4 – Posterior distributions of the mean and subject-wise condition differences (probability scale). Scheme A: Coding based on subjective importance; Scheme B: Coding based on splicing response to no negation sentence.**

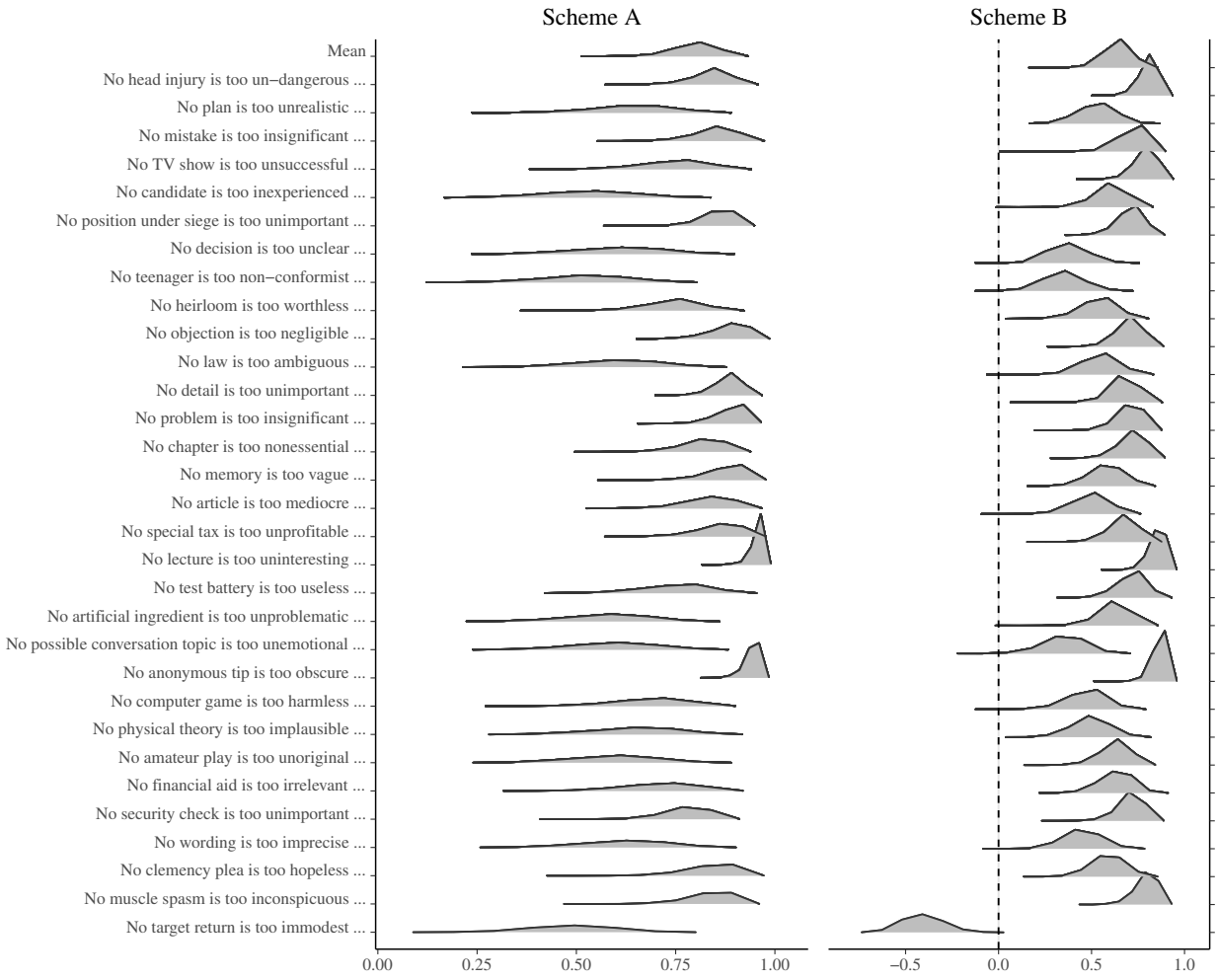


Figure 2: **Experiment 4 – Posterior distributions of the mean and item-wise condition differences (probability scale). Scheme A: Coding based on subjective importance; Scheme B: Coding based on splicing response to no negation sentence.**

As can be seen in Figure 2, the inversion effect is reliably present across different experimental items, with one notable exception under coding scheme B. Incidentally, the only item that showed a reverse effect in the present experiment – at least under the second scheme – was the item that was rated lowest on the comprehensibility scale in Experiment 2B: *No/Some target returns are too immodest to be missed*.

It is somewhat mysterious to us why this particular sentence should behave differently from the rest; in fact, the non-negated and non-quantified version shows precisely the pattern that matches the intuition behind the scheme: the depth charge version yields a sensible semantics while the compositional continuation does not (*This target return is too modest to be missed/?reached*). It is, however, possible that the perceived complexity of financial topics in tandem with negation overload somehow causes even deeper confusion on part of the reader, yielding this novel and unexpected pattern of results. Still, it should be noted that if scheme A should turn out to be the more appropriate way of formalizing meaning inversion,

the item in question would *not* be an outlier, as shown by the posterior distribution.

Another remark concerns a subclass of continuations involving overt negation and, in many cases, the adverb *doch*, 'after all', as in (2).

(2) Kein Bewerber ist zu unerfahren, um ihn nicht doch anzuhören.
No job candidate is too inexperienced to him not after all listen to
'No job candidate is too inexperienced to not listen to him after all.'

Here, the final part of the sentence has an overall positive polarity, that is, the recommendation would be to listen to the candidate, which is sensible under a compositional interpretation. It thus appears that the elements *not* and *after all* essentially cancel each other out; but then why produce them in the first place? Here, negation appears to mainly be used to convey uncertainty that is eventually overcome. However, the coder on this particular trial indicated that inversion had occurred under coding scheme B. There thus seem to be continuations for which coding scheme B does not necessarily yield the expected result, but such cases seem to be comparatively rare: in our data, they amount to about 2% of all supplied continuations. We thus remain confident that coding scheme B, despite being imperfect, is a fundamentally sensible way of categorizing responses into inversion versus no-inversion cases.

References

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* (pp. 61–69).

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, *42*(2–3), 393–405.

Kizach, J., Christensen, K. R., & Weed, E. (2015). A verbal illusion: Now in three languages. *Journal of Psycholinguistic Research*, 1–16.

Vasishth, S., von der Malsburg, T., & Engelmann, F. (2013). What eye movements can tell us about sentence comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(2), 125–134.

von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, *94*, 119–133.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*(4), 591–597.

Appendix C – Formalization of the proposed heuristics

Attempts at providing an explicit formal account of the mechanisms that replace proper compositional interpretation in depth charge sentences have not been made by previous proponents of the overloading account (Wason & Reich, 1979; Kizach et al., 2015), thus leaving the arguably most central part of the explanation underspecified. The proposal of a formal mechanism by itself does not entail that the mechanism is licensed by grammar: Even though they operate on grammatical representations, the proposed heuristics themselves are thought of as extragrammatical. Despite having observed a weak depth charge effect also in sentences with *so*, we put the focus on *too* here, as it appears to contribute strongly to the effect, but it should be possible to adapt the account to other degree particles as well.

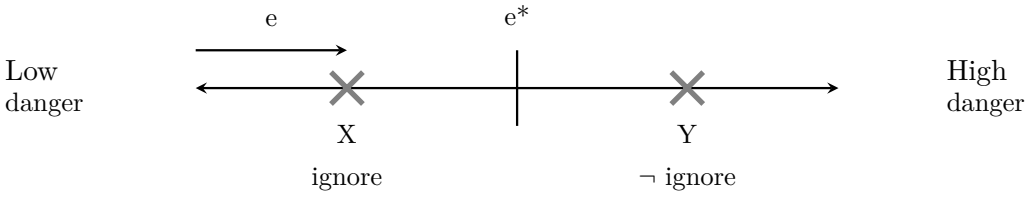
Our proposed semantic account is based on the analysis of *too* given by Meier (2003), which is in turn based on the extent semantics of von Stechow (1984). Extents are intervals on scales associated with gradable adjectives such as *dangerous* which are bounded by zero and positive infinity. Meier (p. 92) states that "[i]nformally, the *too*-construction is true in a world if the extent that satisfies the extent predicate expressed by the main clause is greater than the maximal extent that satisfies the conditional corresponding to the infinitival clause." According to Meier's analysis, the internally coherent sentence *This head injury is too dangerous to be ignored* would have the semantics shown in (1), where the head injury in question would be Y on the scale.

(1) **No global negation, no adjectival negation: Coherent**

This head injury is too dangerous to be ignored.

The maximal e such that the head injury is e-dangerous > the maximal e such that, if the head injury is e*-dangerous, it should be ignored.*

Conclusion: *treat head injury*



The nonsensical sentence *This head injury is too un-dangerous (trivial) to be ignored* with adjectival negation would instead have the semantics shown in (2), where the head injury in question is again Y:

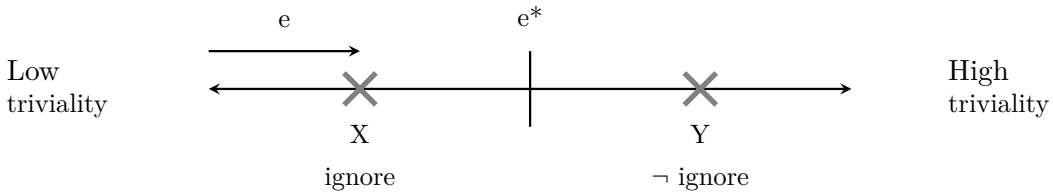
(2) **No global negation, adjectival negation: Transparently incoherent**

This head injury is too un-dangerous to be ignored.

The maximal e such that the head injury is e-un-dangerous > the maximal e such*

that, if the head injury is e-un-dangerous, it should be ignored.*

Conclusion from incoherent premise: *treat head injury*



The sentence thus states that the extent of "un-dangerousness" (read: triviality) of the head injury is greater than the extent of triviality for which it would still be permissible to ignore it (note that this is nonsensical), so the conclusion is that it should be treated. In the graphical representation, head injury Y could be the head injury referred to.

Global negation adds a negative existential quantifier to the above semantics. If the conclusion from (2) is that the head injury should be treated, it stands to reason that the negated version should mean that all head injuries should be ignored; this is indeed the resulting meaning in (3). The claim made by the sentence would be that head injuries like Y do not exist; therefore, all head injuries fall below the triviality threshold introduced by *too*.

(3) **Depth charge sentence with global negation: Add negated existential quantifier**

No head injury is too un-dangerous to be ignored.

There is no head injury such that the maximal e such that the head injury is e-un-dangerous > the maximal e such that, if the head injury is e*-un-dangerous, it should be ignored.*

Conclusion from incoherent premise: *treat none*

We can think of at least three ways in which this semantics could be changed to yield the inverted reading. Two of them are based on our proposed heuristics *negation cancellation* and *negate the verb*.

1 The *negation cancellation* heuristic

Negation cancellation assumes that both the global negation and the negation on the adjective are deleted because the reader infelicitously assumes that they cancel each other out, as shown in (4), whose scale is the same as the one for the sensible (1). The claim made by the sentence is that there is at least one head injury that crosses the threshold, such as Y.

(4) **Possible inverted reading 1** (*"Negation cancellation"*)

There is $\boxed{\textit{at least one}}$ *head injury such that the maximal e such that the head injury is e-*$\boxed{\textit{dangerous}}$ *> the maximal e* such that, if the head injury is e*-*$\boxed{\textit{dangerous}}$*, it should be ignored.*

Conclusion: *treat all*

As can be seen, the result is an inverted semantics where the conclusion has changed into the illusory *treat all* reading.

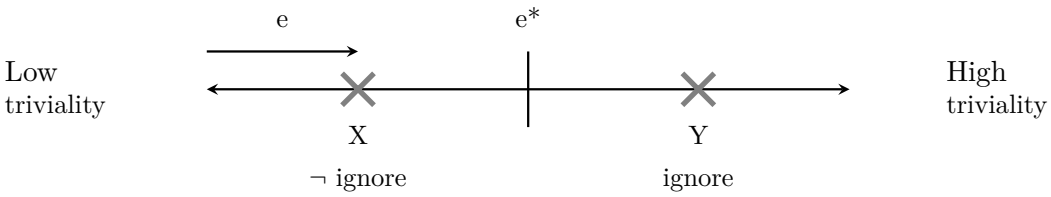
2 The *negate the verb* heuristic

The second heuristic, *negate the verb*, instead introduces a "phantom" negation that scopes locally over *ignored*, as shown in 5. The result is, again, an inverted conclusion: The globally negated sentence now claims that head injury Y does not exist, that is, all head injuries fall below the threshold and should be treated.

(5) **Possible inverted reading 2** (*"Negate the verb"*)

There is no head injury such that the maximal e such that the head injury is e-un-dangerous > the maximal e such that, if the head injury is e*-un-dangerous, it should* $\boxed{\textit{not}}$ *be ignored.*

Conclusion: *treat all*



Interestingly, in the terms of Horn (2009), *negate the verb* points to hyponegation – that is, the sentence containing fewer negations than needed for the resulting interpretation – rather than hypernegation – that is, overt negative elements remaining uninterpreted – as the reason for the illusion, contrary to what is assumed by O'Connor (2015).

3 Turning *too* into *enough*

Finally, there is a third way in which the inverse reading can be generated, namely by flipping the sign of the critical comparison of extents. This way, the semantics of *too* is turned into that of *enough*: Instead of none of the head injuries satisfying the condition for treatment, all head injuries will now satisfy the condition, just like in the sensible sentence *No head injury is un-dangerous (trivial) enough to be ignored.* The meaning shown in (6) is the result of dropping the potentially problematic implicit negation from the semantics of *too.*

(6) **Possible inverted reading 3** (*"too-as-enough"*)

There is no head injury such that the maximal e such that the head injury is e-un-dangerous < the $\boxed{\textit{minimal}}$ *e* such that, if the head injury is e*-un-dangerous, it should be ignored.*

Conclusion: *treat all*

This third possibility agrees with the suggestion by O'Connor (2015) that *too* may be (at least in depth charge configurations) ambiguous between its "normal" meaning and one in which the implicit negation is rendered inert. Whether *too* really turns into a copy of *enough* in depth charge contexts is somewhat doubtful, however: In his comparison of the two lexical items, Fortuin (2014) argues that a depth charge sentence such as *No head injury is too trivial to be ignored* cannot be paraphrased using its counterpart with *enough*, and the sentence *No head injury is trivial enough to be ignored* is labeled as "pragmatically odd or at least marked" (p. 280). It is thus an open question whether *too*, after having lost its implicit negation, has the exact same semantics as *enough*, or whether the latter needs to be analyzed differently.

References

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), 249–292.

Horn, L. R. (2009). Hypernegation, hyponegation, and parole violations. In *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* (Vol. 35, pp. 403–423).

Kizach, J., Christensen, K. R., & Weed, E. (2015). A verbal illusion: Now in three languages. *Journal of Psycholinguistic Research*, 1–16.

Meier, C. (2003). The meaning of too, enough, and so... that. *Natural Language Semantics*, *11*(1), 69–107.

O'Connor, E. (2015). Comparative illusions at the syntax-semantics interface. *Los Angeles, CA: University of Southern California dissertation*.

von Stechow, A. (1984). My reaction to Cresswell's, Hellan's, Hoeksema's and Seuren's comments. *Journal of Semantics*, *3*(1-2), 183–199.

Wason, P. C., & Reich, S. S. (1979). A verbal illusion. *The Quarterly Journal of Experimental Psychology*, *31*(4), 591–597.