Comprehenders Model the Nature of Noise in the Environment

Rachel Ryskin[1,2], Richard Futrell[3], Swathi Kiran[2] & Edward Gibson[1]

1. Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar St. Cambridge, MA, 02139

2. Dept. of Speech, Language, and Hearing Science, Boston University, 635 Commonwealth Avenue, Boston, MA 02215

3. Dept. of Language Science, University of California, Irvine, 3151 Social Sciences Plaza, Irvine, CA 92697

Corresponding author: Rachel Ryskin, ryskin@mit.edu

Abstract

In everyday communication, speakers make errors and produce language in a noisy environment. Recent work suggests that comprehenders possess cognitive mechanisms for dealing with noise in the linguistic signal: *a noisy-channel model*. A key parameter of these models is the *noise model*: the comprehender's implicit model of how noise affects utterances before they are perceived. Here we examine this noise model in detail, asking whether comprehension behavior reflects a noise model that is adapted to context. We asked readers to correct sentences if they noticed errors, and manipulated context by including exposure sentences containing obvious deletions (*A bystander was rescued by the fireman in the nick time.*), insertions, exchanges, mixed errors, or no errors. On test sentences (*The bat swung the player.*), participants' corrections differed depending on the exposure condition. The results demonstrate that participants model specific *types* of errors and make inferences about the intentions of the speaker accordingly.

Keywords: sentence comprehension; noisy-channel; rational inference; adaptation; error correction

Everyday language use occurs amid myriad sources of noise. In a conversation, the speaker may say one word when she intended to say another, there may be other conversations going on in the same room, and the listener may mishear what was said. Each of these types of noise serves to corrupt the signal that is transmitted from speaker to listener (Shannon, 1948). One might think that such noise would pose major impediments to efficient communication. Yet language comprehension typically unfolds without noticeable effort.

Because of this noise, comprehenders maintain uncertainty about the nature of preceding words. When reading sentences such as, "The coach smiled at the player tossed the ball" readers' eye movements indicate that they leave open the possibility that "at" was actually "and." Replacing "at" with "and" allows the interpretation of "tossed" as a finite verb rather than a past participle; the former interpretation has a much higher conditional probability (Levy et al., 2009). Thus, readers have probabilistic representations of language input—in particular, syntactic constructions—and use prior knowledge to infer the intended meaning.

Recent theories have proposed that the language processing system maintains uncertainty about the input because it is designed to optimally decode the intended meaning from a signal transmitted over a noisy channel (Bergen, Levy, & Gibson, 2012; Gibson et al., 2013; Jaeger, 2010; Levy, 2008; Levy et al., 2009). In particular, Gibson et al. (2013) lay out a framework for sentence comprehension that entails the rational (Bayesian) integration of noisy evidence and semantic priors. On their account, the producer chooses an intended sentence s_i in order to communicate her intended meaning, m_i. s_i is conveyed across a noisy channel and is corrupted by noise originating from the

producer, comprehender, or environment. The comprehender perceives sentence s_p and tries to infer s_i. Communication succeeds when the intended sentence s_i can be recovered from s_p. This process can be formalized by considering an ideal observer (Geisler & Diehl, 2003) model of language comprehension, where the comprehender engages in optimal Bayesian decoding of the intended meaning:

$$P(s_i \mid s_p) \propto P(s_i)P(s_i \rightarrow s_p) \quad (1)$$

In Equation (1), $P(s_i \mid s_p)$ represents the probability assigned by the comprehender to any hypothesized s_i, given the observed linguistic input s_p. By Bayes rule, this probability can be rewritten as the prior probability $P(s_i)$ that a producer would wish to communicate s_i, multiplied by the probability of s_i being corrupted to s_p during communication, $P(s_i \rightarrow s_p)$. The prior, $P(s_i)$, represents the comprehender's relevant linguistic and world knowledge, and biases comprehenders towards a priori plausible utterances. The noise model $P(s_i \rightarrow s_p)$ encodes the comprehender's knowledge of how sentences can be corrupted—for instance, smaller changes to a sentence are more likely than larger ones. By integrating $P(s_i)$ and $P(s_i \rightarrow s_p)$, comprehenders may arrive at interpretations which differ from the literal meanings of the acoustic or visual stream. That is, if comprehenders perceive an implausible sentence s_p (e.g., The oven cleaned the grandmother) which is "close" to a more plausible sentence (e.g., The grandmother cleaned the oven), they should infer that the producer actually uttered (or intended to utter) the plausible version.

Gibson et al. (2013) provide evidence for several predictions of the noisy-channel framework in a series of experiments where participants read implausible sentences (e.g., The oven cleaned the grandmother) followed by comprehension questions (e.g., Was the grandmother cleaned by someone/something?), which probed whether participants

interpreted the sentence literally (answer: Yes) or inferred that the intended sentence had been corrupted (answer: No). They found that comprehenders were a) more willing to forego the literal interpretation when the semantically plausible interpretation involved positing fewer changes, b) more likely to infer nonliteral meanings when the change involved a deletion compared to an insertion, consistent with the Bayesian size principle (Xu & Tenenbaum, 2007), c) more likely to endorse literal interpretations when the fillers contained errors, indicating that they had inferred a higher noise rate; and d) less likely to endorse literal interpretations when the base rate of implausible sentences was increased, suggesting that they had adjusted their semantic prior. Further, Poppels and Levy (2016) replicated these results and demonstrated that, in addition to deletions and insertions, word exchanges represent a likely form of corruption (e.g., The package fell from the floor to the table.).

Noise Variation

Gibson et al. (2013) demonstrated that participants adapt their noise model when provided with evidence of a high base-rate of syntactic errors. Further, listeners infer a higher noise rate when listening to foreign-accented speech (Gibson et al., 2017). Yet, how the noise likelihood term ($s_i \rightarrow s_p$) responds to input characteristics beyond error rate has yet to be explored. Critically, we can ask: is the noise model sensitive to the nature of errors or simply to the rate of errors?

In real-world language use, many properties of the noise, beyond the rate, vary with context. For example, second language (L2) learners may make certain errors in English that a native speaker is unlikely to make and that are influenced by their native language (see MacWhinney, 1992). Native speakers of Russian tend to omit articles when

speaking L2 English (e.g., Ionin, Ko, & Wexler, 2004), while native speakers of French may exchange the orders of adjectives and nouns in L2 English (Nicoladis, 2006). If the comprehender's noise model is sensitive to the nature of errors, it will have different properties when listening to an L2 English speaker from Russia than to an L2 English speaker from France. However, if the noise model is sensitive to an overall rate of errors, it will be similar for the two speakers.

Recent findings suggest that comprehenders rapidly learn and adapt to the linguistic patterns (e.g., frequencies of syntactic constructions, phonetic category boundaries) present in their environment in order to achieve more efficient language processing (Fine, Jaeger, Farmer, & Qian, 2013; Kleinschmidt & Jaeger, 2015; Ryskin, Qi, Duff, & Brown-Schmidt, 2017; though see Harrington Stack, James, & Watson, 2018 for an example of limits on this ability). Similarly, comprehenders may track the types of errors they perceive in a given environment and rapidly adapt the likelihoods of components of the noise model. For example, after hearing a speaker repeatedly drop articles (e.g., "We had nice time at beach."), the listener's noise model may put high probability on certain words being deleted, but the probability of insertions may not change. Thus, the noise model for the article-dropping speaker would have a larger ratio of deletions to other errors, as compared to the noise model for a generic, native English speaker. Forming these fine-grained, context-specific representations of the noise would likely allow comprehenders to make more accurate inferences about the intended meaning s_i. We call such a noise model a context-specific noise model.

On the other hand, hearing a speaker repeatedly drop articles may lead the listener's noise model to put higher probability on all possible errors, perhaps on the

reasonable assumption that a speaker who makes one type of error is likely to also make other errors in the future. Under such a context-invariant noise account, the comprehender's noise model always possesses the same general properties (e.g., more edits are less likely than fewer edits, insertions are less likely than deletions) and varies only in the base-rate of corruptions. In an environment with a high base-rate of errors, comprehenders simply increase the likelihoods of all errors by a constant. In every other respect, the probabilities of different occurrences (e.g., deletions vs. insertions) maintain the same ratio across contexts. Inferring the noise model would then simply reflect the process of adjusting all the likelihoods in the noise model up or down, depending on recent evidence.

Whether comprehenders have context-invariant or context-specific noise models gets at the more general question of how people trade off complexity of models and accuracy in prediction. If the context-invariant model is correct, then this suggests that comprehenders weight model simplicity as more important than accuracy in prediction: the context-invariant noise model only has one parameter, the noise rate, and thus it should be easier to learn and deploy than a more complex model. If the context-specific model is true, then comprehenders weight accuracy as more important than model complexity in this case: the context-specific model achieves higher accuracy at the cost of greater complexity. The optimal tradeoff of accuracy and complexity will depend on the true rate of context-specificity in the world and on the exact nature of the complexity cost for noise models. These complexity-accuracy tradeoffs are at the heart of all theories of statistical learning (Solomonoff, 1964; MacKay, 2003). Investigating these two particular hypotheses in the context of noisy-channel language understanding allows us to

develop models of how complexity and accuracy trade off in language processing.

In the present experiments, we test these hypotheses by probing readers' inferences about intended meanings of sentences and manipulating the experimental context to include sentences with specific types of errors (e.g., deletions, insertions, or exchanges). If comprehenders track the base-rate of errors but don't model the nature of the errors (context-invariant), they should make more inferences when they're exposed to errors than when the context contains only error-free sentences (Gibson et al., 2013), but the pattern of inferences should not differ by type of error exposure. However, if readers track more fine-grained error information beyond the base-rate (context-specific), their inferences should be sensitive to the type of error they experienced.

The goals of Experiment 1 were to a) replicate the effect of increasing the noise rate observed in Gibson et al. (2013) using a more direct measure (retyping and editing rather than comprehension questions), and b) test whether readers are sensitive to the nature of noise in the exposure. The goal of Experiment 2 was to run a pre-registered replication of Experiment 1 using a large sample size determined by a simulation-based power analysis of Experiment 1.

Methods

Participants

In Experiment 1, participants were 293 Amazon Mechanical Turk workers with IP addresses in the US who self-reported being native English speakers. They were paid $3.00 for their participation. In Experiment 2, participants were 880 workers with IP addresses in the US who self-reported being native English speakers. They were paid

$2.50 for their participation.

Materials & Design

Participants were told that they would be reading 95 sentences that were transcriptions of someone's speech and that these transcriptions might contain errors. Participants were asked to retype each sentence in a text box and edit it if they thought the speaker had intended something different. They saw all the sentences at once and were able to return to any sentence and change their answer at any time before submitting their work. The instructions for Experiment 2 additionally stated that participants were allowed to copy and paste sentences if they did not think they contained any errors. The latter two measures were intended to minimize any errors that participants could introduce themselves while typing (though such errors would not be expected to affect the results of the experimental manipulation because they should occur randomly and uniformly across conditions).

The 95 sentences consisted of 15 test sentences, 20 exposure sentences, and 60 filler sentences. All experimental materials are available at osf.io/rkrha. The pre-registration for Experiment 2 is available at osf.io/83nsn. The test sentences were taken from Gibson et al. (2013). A norming study[1] was used to select 10 sentences that were likely to be interpreted as resulting from a deletion (e.g., The uncle sold the truck the father.), 10 as an insertion (The earthquake shattered from the house.), and 10 as an exchange (e.g., The oven cleaned the grandmother.). These 30 test sentences were then

[1] The norming study tested 144 potential test sentences, 160 potential exposure sentences (40 groups of 4 sentences where one contains no error and the other 3 are minimal variants that contain either a deletion, an insertion, or an exchange error), and 80 filler sentences. These sentences, which were split into 2 lists, were retyped and edited by 97 Mechanical Turk workers. The responses were coded in terms of what kind of error was perceived. Sentences with low inter-responder agreement about the type of error were not included in the experiment.

split into 2 lists of 15, five of each type, to make the task shorter for participants and increase the ratio of exposure sentences to test sentences. Data from the 2 lists were analyzed together. The 20 exposure sentences differed by condition in terms of what type of error they contained (see Table 1). The 60 filler sentences were mostly taken from Gibson et al. (2013) and did not contain any errors (e.g., The journalist was ignored by the politician at the press conference.). Four test sentences were changed between Experiment 1 and Experiment 2; because of the high rate of Inferred Exchanges, sentences that could only plausibly be interpreted as the result of exchanges (e.g., The essay wrote the student.) were switched for sentences from Gibson et al. (2013) that were rated most likely to be the result of exchanges but could also be interpreted as resulting from deletion (e.g., The bat swung the player).

Participants were randomly assigned to one of 10 lists (2 lists of test items x 5 exposure conditions). The order of the 95 items was pseudo-randomized for each participant with the constraints that the first 3 items were fillers and 2 test items did not directly follow each other.

Condition	Example
Deletion	A bystander was rescued by the fireman in the nick time
Insertion	A bystander was rescued by the fireman to in the nick of time.
Exchange	A bystander was rescued by the fireman in the time of nick.
Mixed	1/3 Deletion, 1/3 Insertion, 1/3 Exchange
No Error	A bystander was rescued by the fireman in the nick of time.

Table 1. Example exposure items by condition

Coding

Participants' responses (typed sentences) were compared to the sentences they

read (e.g., The oven cleaned the grandmother.) and coded[2] as an Inferred Deletion if their response was mostly the same as the prompt but contained one or more extra words (e.g., The oven was cleaned by the grandmother), as an Inferred Insertion if the response contained fewer words than the prompt (e.g., The cleaned grandmother.), an Inferred Exchange if the response contained the same words but their order was changed (e.g., The grandmother cleaned the oven.), as Inferred No Error if the sentence was identical to the prompt and as Inferred Other when the sentence was edited but not in a way that fit into any of the preceding categories (e.g., The soap cleaned the grandmother.). If the participant's response formed an ungrammatical sentence (e.g., "The actor handed the director script") it was coded as Inferred Other. If there was an obvious typographical error (e.g., "the the actor…") this was ignored. If responses contained two changes, for example both a change in the order of words and the insertion of a word (e.g., "The oven cleaned the grandmother" becoming "The grandmother was cleaned by the oven") the response was coded as Inferred Other.

Predictions

The context-specific noise hypothesis predicts that inferred errors should increase

[2] Coding was performed in two steps, both of which were blind to condition: 1) Using the "stringi" package in R, each response sentence was compared (character by character) to the corresponding stimulus sentence. Those responses that were identical to the stimulus were automatically coded as "Inferred No Error." For the responses that returned a non-zero difference value, the lengths of the two sentences were compared. If the lengths of stimulus and response sentences were the same, the response was coded as "Inferred Exchange." If the length of the response was longer than the stimulus sentence, this was coded as "Inferred Deletion." If the length of the response sentences was shorter than the stimulus sentence, the response was coded as "Inferred Insertion." 2) All codings were then checked manually by the first author (with condition information hidden during this process) and the fifth category "Inferred Other" was added for cases that were not adequately covered by step 1. This checking procedure changed 6.0% and 6.8% of the coding in Experiment 1 and 2 respectively. The majority of these changes were due to capitalization and punctuation discrepancies. For example, a stimulus and response sentence that differed by the absence of a terminal period in the response would be marked as "Inferred Deletion" in step 1. This would have been corrected to "Inferred No Error" in step 2. The results of both coding steps can be found in the datasets made available on OSF (correction.type and correction.type2 for the coding resulting from step 1 and 2 respectively; see Supplementary Materials).

in the presence of any type of additional noise relative to a baseline of no (additional) noise. Thus, Inferred No Error responses should be more likely in the No Error Exposure condition than the Mixed condition (and other error Exposure conditions). Further, the context-specific hypothesis predicts that the increase in inferred errors should be sensitive to the nature of the additional noise, so the rates of Inferred Deletions, Inferred Insertions, and Inferred Exchanges should differ by Exposure condition. More specifically, Inferred Deletions should be significantly more likely in the Deletion Exposure condition than the Mixed condition, Inferred Insertions should be significantly more likely in the Insertion Exposure condition than the Mixed condition, and Inferred Exchanges should be significantly more likely in the Exchange Exposure condition than the Mixed condition.

The context-invariant noise hypothesis also predicts that inferred errors should increase in the presence of noise. Thus, Inferred No Error responses should be more likely in the No Error Exposure condition than any of the four Exposure conditions that introduced additional noise (Mixed, Deletion, Insertion and Exchange). However, the increase in inferred errors should not be sensitive to the nature of the additional noise, so the rates of Inferred Deletions, Inferred Insertions, and Inferred Exchanges should not differ by Exposure condition.

Results

Figure 1 shows the average proportion of each type of response (Inferred Deletions, Inferred Insertions, Inferred Exchanges, and Inferred No Errors) by Exposure Condition, in both Experiment 1 and 2 (the replication). Across Exposure conditions, participants were not equally likely to infer Exchanges, Insertions, Deletions, Other Errors, or No Error ($\chi^2(4)=49894.4$, $p< .001$). Participants inferred Exchanges (E1: 53%,

E2: 54% of responses) more than Deletions (E1: 27%, $\chi^2(1)=627.5$, $p_{Holm\text{-}adjusted} < .001$; E2: 27%, $\chi^2(1)=2024.9$, $p_{Holm\text{-}adjusted} < .001$) and Insertions (E1: 13%; E2: 12%) less than Deletions (E1: $\chi^2(1)=288.2$, $p_{Holm\text{-}adjusted} < .001$; E2: $\chi^2(1)=1001.6$, $p_{Holm\text{-}adjusted} < .001$).

The effects of Exposure conditions on responses were analyzed using four logistic mixed-effects models[3] with random intercepts for participants and items nested within type of item (test sentences belonged to one of three types that are differentially likely to elicit deletions, insertions, or exchanges based on the norming). Random by-items slopes were included for Exposure condition. The Mixed Errors Exposure condition was used as the reference level in all models. Estimated parameters for all fixed effects are reported in Table 2 and the full set of model parameters is reported in Appendix A. Inferred Deletions were more likely in the Deletion Exposure condition than the Mixed condition (only numerically[4] in E1: b= 0.68, p= 0.40; E2: b = 1.34; p < .001) and significantly less likely in the Exchange Exposure condition in Experiment 1 (b=-1.34; p= 0.03). In Experiment 2, Inferred Deletions were also more likely in the Insertion Exposure condition (b = 1.29; p = 0.04) than the Mixed condition. Inferred Insertions were only numerically more likely in the Insertion Exposure condition than the Mixed condition in both experiments[5]. Inferred Exchanges were significantly more likely in the Exchange Exposure condition than the Mixed condition (E1: b = 1.25, p = .003; E2: b = 0.85; p = .001) and significantly less likely in the Deletion Exposure (E1: b = -0.88, p = 0.02; E2: b

[3] Given the multinomial nature of the responses we also fit a Bayesian multinomial mixed-effects model. However, this analysis was post-hoc and thus we report the pre-registered logistic regressions in the main text and report the multinomial analysis in Appendix B. The results from both analyses are consistent.
[4] Likely due to insufficient power in the first experiment.
[5] The fact that the Exposure effect does not reach significance in the case of Inferred Insertions may be a consequence of how unlikely comprehenders perceive insertion errors to be from the outset or, relatedly, an artifact of attempting to fit a logistic model to a dataset with a large number of zeros (where the log-odds approach negative infinity).

= -0.65; p = 0.04). Inferred Exchanges were also significantly less likely in the Insertion Exposure condition than the Mixed condition in Experiment 2 (b = -0.76; p = 0.003). Finally, Inferred No Errors responses were significantly more likely in the No Errors Exposure condition than the Mixed condition (E1: b = 1.16; p = 0.04; E2: b = 0.67; p = 0.01).

Experiment 1					Experiment 2			
	b	*SE*	*z*	*p*	*b*	*SE*	*z*	*P*
Inferred Deletions								
Mixed (Intercept)	-3.397	2.331	-1.457	0.145	-4.544	2.859	-1.589	0.112
Deletion	0.683	0.808	0.845	0.398	1.34	0.381	3.516	0.000*
Insertion	0.647	0.507	1.275	0.202	1.285	0.621	2.07	0.038*
No Error	-0.824	0.766	-1.076	0.282	0.162	0.51	0.318	0.751
Exchange	-1.336	0.61	-2.191	0.028*	-0.662	0.429	-1.543	0.123
Inferred Insertions								
Mixed (Intercept)	-7.984	4.803	-1.662	0.096	-6.976	2.475	-2.819	0.005
Deletion	-1.22	7.234	-0.169	0.866	-3.397	3.129	-1.086	0.278
Insertion	3.095	3.949	0.784	0.433	-2.889	2.823	-1.023	0.306

No Error	2.475	4.058	0.61	0.542	-3.425	3.063	-1.118	0.263
Exchange	2.495	4.102	0.608	0.543	-4.422	3.623	-1.22	0.222
Inferred No Errors								
Mixed (Intercept)	-4.855	0.739	-6.569	0.000	-4.419	0.735	-6.015	0.000
Deletion	-0.408	0.823	-0.496	0.620	-0.088	0.334	-0.262	0.793
Insertion	-0.397	0.683	-0.581	0.562	0.109	0.287	0.378	0.705
No Error	1.161	0.576	2.014	0.044*	0.672	0.271	2.482	0.013*
Exchange	-1.342	2.085	-0.644	0.520	0.03	0.347	0.088	0.930
Inferred Exchanges								
Mixed (Intercept)	0.523	1.807	0.29	0.772	0.472	1.538	0.307	0.759
Deletion	-0.882	0.383	-2.303	0.021*	-0.647	0.312	-2.074	0.038*
Insertion	-0.632	0.391	-1.618	0.106	-0.757	0.25	-3.024	0.002*
No Error	0.002	0.408	0.005	0.996	-0.039	0.212	-0.185	0.853

Exchange	1.251	0.428	2.926	0.003*	0.852	0.261	3.261	0.001*

Table 2: Fixed Effects from Logistic Mixed-Effects Regression Analyses (Experiment 1 and 2; see Appendix A for full models with random effects structure). Asterisk denotes significant differences at $\alpha = 0.05$.

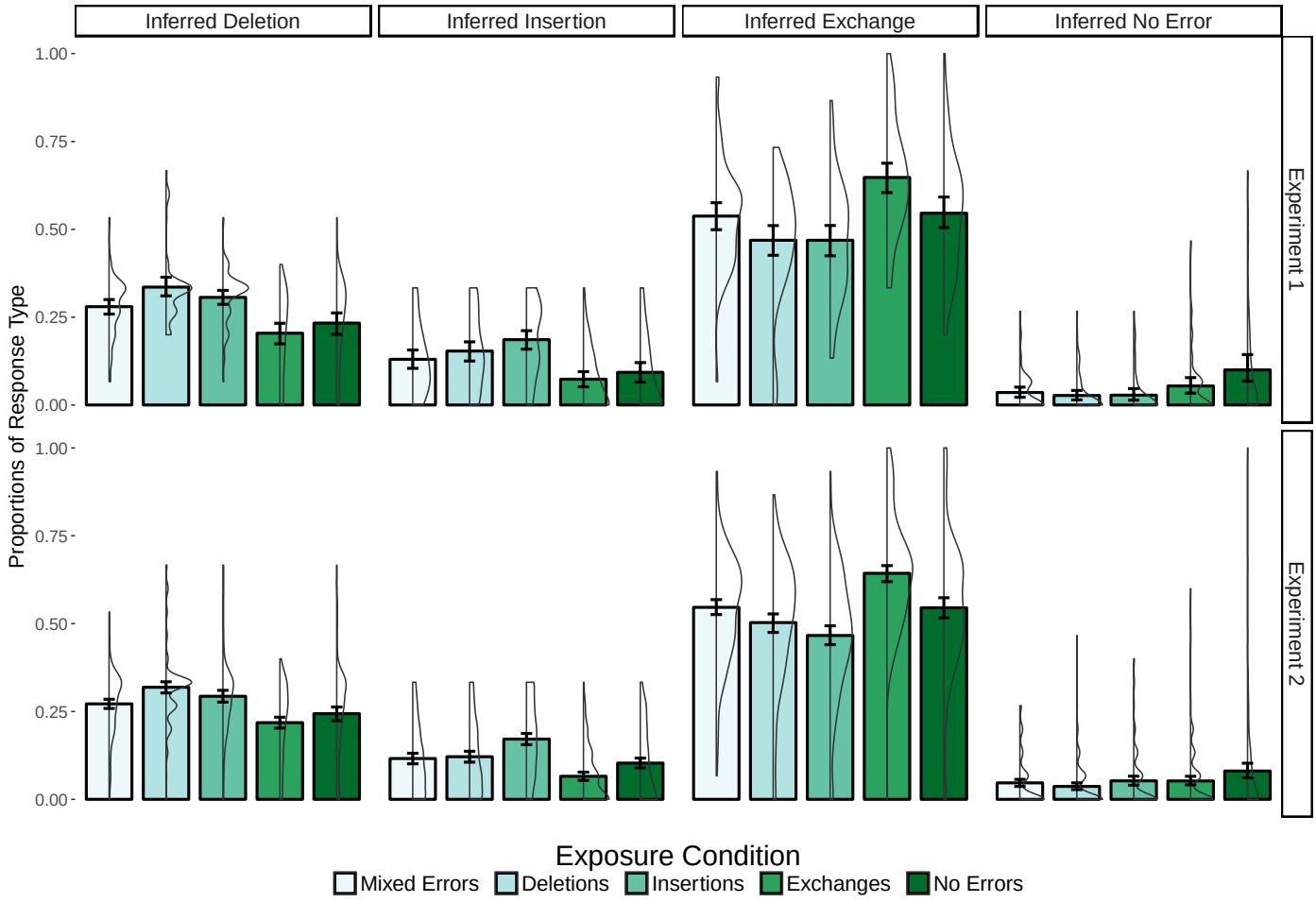

Figure 1. Proportions of edits by type of inference and Exposure Condition in Experiments 1 and 2. Vertical densities indicate distribution of individual responses included in each average. Error bars represent bootstrapped 95% confidence intervals. (See Appendix C for proportions of each response by test item.)

General Discussion

In two experiments, readers corrected sentences based on what they thought the speaker had originally intended. We observed that participants have a noise model in which Insertions were less likely than Deletions overall and Exchanges were the most likely type of error. These patterns are consistent with previous observations (Gibson et al., 2013; Poppels & Levy, 2016), though in the present paradigm idiosyncrasies of the stimulus set may contribute to these different baselines. Edits were overall more likely when participants were exposed to sentences with a mixture of errors than when there were no errors in the input, consistent with Gibson et al. (2013). Furthermore, the inferences readers made— about the nature of noise corruption that was applied to sentences—were influenced by the nature of errors most common in their input. In particular, participants inferred that an Exchange had occurred most often when the experimental context contained many sentences with Exchange errors and they inferred that a Deletion had occurred most often when the context contained many sentences with Deletion errors. Similarly, they were most likely to infer no corruption in the case where there were no errors in the experimental context.

In addition, Inferred Deletions were more likely in both the Deletion and Insertion Exposure conditions whereas Inferred Exchanges were less likely in both Deletion and Insertion Exposure conditions compared to Mixed Exposure, which suggests a categorical distinction in the noise model that groups insertions and deletions together as analogous types of errors and exchanges as a different category of error. This somewhat unexpected result suggests that comprehenders' model of the noise assumes multiple potential generators of errors in written language: one that produces deletions

and insertions and one that produces exchanges. One possibility is that deletions and insertions are the kinds of errors that one might expect to occur when the producer is typing fast and omitting words or forgetting to delete words that no longer fit in the sentence. On the other hand, exchanges may reflect a higher-order language planning error that can occur independently of the producer's typing speed or competence. Recall that, in these experiments, participants were told that they were reading transcriptions of a speaker's productions, thus both error generators could have played a role in corrupting the sentences they perceived. One very speculative interpretation could be that participants who observed many deletions in the exposure may have inferred that the speech was pristine but the transcriber who was typing was error-prone, thus making insertions seem more plausible as well (and vice versa for those in the Insertions Exposure condition). On the other hand, participants in the Exchange exposure may have inferred that the transcriptions were careful but the speaker was inattentive. Further work is needed to determine whether these groupings of errors hold across experiments and whether they reflect comprehenders' hypotheses about likely mechanisms underlying the errors they observe in the input.

One important difference between the current approach and prior investigations of noisy-channel sentence comprehension, is the use of a more explicit measure of noise inference: re-typing and correcting errors. Previous tasks have relied on implicit measures of noisy-channel processing, such as comprehension questions (Gibson et al., 2013) or regressive eye-movements (Levy et al., 2009). Those methods have the benefit of being (perhaps) more typical of everyday comprehension activities, though explicit correction is a common activity for anyone who edits papers, for example. In addition, those tasks

reveal that noisy-channel correction can be covert and readers often are not aware that they have interpreted the sentence in a way that is different than the literal string. However, these implicit measures allow for only a very indirect assessment of the noise model that readers are deploying: a given perceived sentence could have originated from a variety of intended sentences via different noise corruption processes. Understanding, at a more fine-grained level, which noise corruptions comprehenders find more or less probable is a critical step in developing computational models of noisy-channel sentence comprehension and a goal of the current work. It seems parsimonious to assume that the same noise model is at play during implicit and explicit corrections. Thus, we probe the noise model directly by asking readers to reverse whichever corruption they find most plausible given a sentence.

Conclusion

In this work, we provide robust evidence that language comprehenders model the noise that is present in their environment and make rational inferences in accordance with that model. Moreover, our results support the context-specific noise hypothesis: the model of the noise is not only sensitive to the base rate of noise, but also to more fine-grained information about the nature of errors that are present. This novel finding extends prior work showing that comprehension involves rational inference accounting for potential noise in the transmission of a linguistic message by integrating information about the prior probability of a perceived message with the likelihood of a particular corruption (Levy et al., 2009; Gibson et al., 2013). It also informs models of language processing that account for noise in the input by suggesting that there is a dynamically adaptive component to the noise model.

More broadly, these results are consistent with recent findings that language users are sensitive to the statistics of the input and form expectations about upcoming linguistic material at many levels of representation—acoustic features, syntactic constructions, referential form, pragmatic cues inter alia—that are tailored to the distributional information in the input (Brown-Schmidt, 2009; Brown-Schmidt, Yoon & Ryskin, 2015; Fine et al., 2013; Kleinschmidt & Jaeger, 2015; Kurumada, Brown, & Tanenhaus, 2017; Ryskin et al., 2017). In terms of everyday language use, the results reported here suggest that, when meeting an L2 speaker of English with an idiosyncratic pattern of errors (e.g., omitting articles or inverting word order), listeners do not simply assume that the speaker is more likely to make errors across the board. Rather, listeners learn what kinds of errors the speaker is prone to and fine-tune their inferences about the intended meaning of the speaker's utterances to account for those specific errors.

References

Bergen, L., Levy, R., & Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. In *Proceedings of the thirty-fourth annual conference of the cognitive science society* (p. 1320-1325).

Brown-Schmidt, S. (2009). Partner-specific interpretation of maintained referential precedents during interactive dialog. *Journal of Memory and Language*, *61*(2), 171–190. http://doi.org/10.1016/j.jml.2009.04.003

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as Contexts in Conversation. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 62, pp. 59–99). Academic Press. http://doi.org/10.1016/bs.plm.2014.09.003

Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28.<doi:10.18637/jss.v080.i01>

Fine, A., Jaeger, F., Farmer, T., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE , 8* , e77661.

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

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*, 8051-6.

Gibson, E., Tan, C., Futrell, R., Mahowald, K., Konieczny, L., Hemforth, B., & Fedorenko, E. (2017). Don't Underestimate the Benefits of Being Misunderstood. *Psychological Science, 1-10*. http://doi.org/10.1177/0956797617690277

Harrington Stack, C., James, A. & Watson, D. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*. In press.

Ionin, T., Ko, H., & Wexler, K. (2004). Article Semantics in L2 Acquisition: The Role of Specificity. *Source: Language Acquisition*, *12*(1), 3–69.

Jaeger, F. (2010). Redundancy and reduction: speakers manage syntactic information density. *Cognitive Psychology, 61*, 23-62.

Kleinschmidt, D., & Jaeger, F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148-203.

Kurumada, C., Brown, M., & Tanenhaus, M. (in press). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin & Review*. http://doi.org/10.3758/s13423-017-1332-6

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th conference on empirical methods in natural language processing* (p. 234-243).

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 21086-90.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*.

Cambridge, UK: Cambridge University Press.

MacWhinney, B. (1992). Transfer and Competition in Second Language Learning. *Advances in Psychology*, *83*, 371–390. https://doi.org/10.1016/S0166-4115(08)61506-X

Nicoladis, E. (2018). Cross-linguistic transfer in adjective–noun strings by preschool bilingual children. *Bilingualism: Language and Cognition*, *9*(1), 15–32. https://doi.org/10.1017/S136672890500235X

Poppels, T., & Levy, R. (2016). Structure-sensitive noise inference: Comprehenders expect exchange errors. In *Proceedings of the 38th annual meeting of the cognitive science society* (p. 378-383). Poster presentation.

Ryskin, R. A., Qi, Z., Duff, M. C., & Brown-schmidt, S. (2017). Verb Biases Are Shaped Through Lifelong Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 781–794.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal , 27*, 379-423.

Solomonoff, R. J. (1964). A formal theory of inductive inference. Part I. *Information and Control , 7*(1), 1–22.

Xu, F., & Tenenbaum, J. (2007). Word learning as bayesian inference. *Psychological Review , 114*, 245-72.

Author Contributions

All authors contributed to study concept and design. Testing and data collection were performed by RR. RR and RF performed the data analysis and interpretation under the supervision of EG. RR drafted the manuscript, and RF, EG, and SK provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

This work was funded by NIH F32DC015163 to RR and grant 1534318 from the NSF linguistics program to EG.

Appendix A

Experiment 1					Experiment 2			
Inferred Deletions								
Fixed Effects	**β**	**SE**	**z-value**	**p-value**	**β**	**SE**	**z-value**	**p-value**
Mixed (Intercept)	-3.397	2.331	-1.457	0.145	-4.544	2.859	-1.589	0.112
Deletion	0.683	0.808	0.845	0.398	1.34	0.381	3.516	0
Insertion	0.647	0.507	1.275	0.202	1.285	0.621	2.07	0.038
No Error	-0.824	0.766	-1.076	0.282	0.162	0.51	0.318	0.751
Exchange	-1.336	0.61	-2.191	0.028	-0.662	0.429	-1.543	0.123
Random Effects	**SD**				**SD**			
(Intercept) Sentence : Type	0.999				0.679			
Deletion	0.131				0.159			
Insertion	0.047				0.106			
No Error	0.131				0.227			
Exchange	0.183				0.036			
(Intercept) Type	3.996				5.171			
Deletion	1.095				0.397			
Insertion	0.283				0.977			
No Error	0.825				0.695			
Exchange	0.257				0.367			
(Intercept) Participant	1.735				1.735			
Inferred Insertions								
Fixed Effects	**β**	**SE**	**z-value**	**p-value**	**β**	**SE**	**z-value**	**p-value**
Mixed (Intercept)	-7.984	4.803	-1.662	0.096	-6.976	2.475	-2.819	0.005
Deletion	-1.22	7.234	-0.169	0.866	-3.397	3.129	-1.086	0.278
Insertion	3.095	3.949	0.784	0.433	-2.889	2.823	-1.023	0.306
No Error	2.475	4.058	0.61	0.542	-3.425	3.063	-1.118	0.263
Exchange	2.495	4.102	0.608	0.543	-4.422	3.623	-1.22	0.222
Random Effects	**SD**				**SD**			
(Intercept) Sentence : Type	0.711				0.529			
Deletion	0.504				0.308			
Insertion	0.487				0.203			
No Error	0.618				0.169			
Exchange	0.615				0.238			
(Intercept) Type	5.229				4.52			
Deletion	1.144				2.604			
Insertion	1.539				2.898			
No Error	2.286				2.349			
Exchange	2.559				2.471			
(Intercept) Participant	1.352				1.459			
Inferred No Error								
Fixed Effects	**β**	**SE**	**z-value**	**p-value**	**β**	**SE**	**z-value**	**p-value**
Mixed (Intercept)	-4.855	0.739	-6.569	0	-4.419	0.735	-6.015	0
Deletion	-0.408	0.823	-0.496	0.62	-0.088	0.334	-0.262	0.793
Insertion	-0.397	0.683	-0.581	0.562	0.109	0.287	0.378	0.705

No Error	1.161	0.576	2.014	0.044	0.672	0.271	2.482	0.013
Exchange	-1.342	2.085	-0.644	0.52	0.03	0.347	0.088	0.93
Random Effects	**SD**				**SD**			
(Intercept) Sentence : Type	0.74				0.576			
Deletion	1.229				0.215			
Insertion	0.389				0.017			
No Error	0.562				0.147			
Exchange	0.911				0.249			
(Intercept) Type	1.015				1.212			
Deletion	0.875				0.333			
Insertion	0.509				0.155			
No Error	0.419				0.155			
Exchange	2.679				0.36			
(Intercept) Participant	1.6				1.389			
Inferred Exchange								
Fixed Effects	**β**	**SE**	**z-value**	**p-value**	**β**	**SE**	**z-value**	**p-value**
Mixed (Intercept)	0.523	1.807	0.29	0.772	0.472	1.538	0.307	0.759
Deletion	-0.882	0.383	-2.303	0.021	-0.647	0.312	-2.074	0.038
Insertion	-0.632	0.391	-1.618	0.106	-0.757	0.25	-3.024	0.002
No Error	0.002	0.408	0.005	0.996	-0.039	0.212	-0.185	0.853
Exchange	1.251	0.428	2.926	0.003	0.852	0.261	3.261	0.001
Random Effects	**SD**				**SD**			
(Intercept) Sentence : Type	0.791				0.602			
Deletion	0.245				0.242			
Insertion	0.281				0.072			
No Error	0.49				0.053			
Exchange	0.405				0.187			
(Intercept) Type	3.099				2.712			
Deletion	0.343				0.424			
Insertion	0.354				0.295			
No Error	0.397				0.18			
Exchange	0.374				0.31			
(Intercept) Participant	1.466				1.504			
Observations: 4395, Participants: 293, Sentences: 30, Types: 3					Observations: 13200, Participants: 880, Sentences: 30, Types: 3			

Table A. Full model estimates for logistic mixed–effects regressions predicting types of inferences from Exposure conditions in Experiment 1 and 2.

Appendix B

Population-Level Effects		Estimate (mu)	Error	95% CI - lower	95% CI - upper	% of posterior samples > 0
Inferred Deletion	Intercept	-0.37	1.08	-2.67	1.59	37.59
	Deletion Exposure	**0.69**	**0.31**	**0.09**	**1.29**	**98.69**
	Insertion Exposure	0.34	0.31	-0.26	0.96	86.18
	Exchange Exposure	-0.86	0.31	-1.47	-0.25	0.31
	No Error Exposure	-0.66	0.3	-1.23	-0.06	1.41
Inferred Insertion	Intercept	-5.09	2.09	-9.95	-1.73	0.08
	Deletion Exposure	-0.19	0.29	-0.75	0.39	25.88
	Insertion Exposure	**0.82**	**0.3**	**0.23**	**1.43**	**99.61**
	Exchange Exposure	-0.86	0.3	-1.45	-0.29	0.11
	No Error Exposure	-0.58	0.27	-1.12	-0.05	1.55
Inferred Exchange	Intercept	3.58	0.34	2.92	4.25	1
	Deletion Exposure	-0.34	0.27	-0.86	0.19	10.24
	Insertion Exposure	-0.33	0.28	-0.88	0.22	11.7
	Exchange Exposure	**0.4**	**0.27**	**-0.13**	**0.94**	**92.89**
	No Error Exposure	-0.37	0.26	-0.88	0.14	7.63
Inferred No Error	Intercept	-0.06	0.4	-0.87	0.71	44.46
	Deletion Exposure	-0.18	0.34	-0.83	0.48	29.75
	Insertion Exposure	0.26	0.35	-0.42	0.92	77.79
	Exchange Exposure	-0.18	0.33	-0.86	0.47	29.7
	No Error Exposure	0.19	0.32	-0.44	0.82	72.53
Group-Level Effects		Estimate (sd)	Est.Error	95% CI - lower	95% CI - upper	
Items (30 levels)						
Inferred Deletion	Intercept	5.36	1.02	3.75	7.77	
Inferred Deletion	Deletion Exposure	0.22	0.15	0.01	0.58	
Inferred Deletion	Insertion Exposure	0.21	0.16	0.01	0.6	
Inferred Deletion	Exchange Exposure	0.21	0.17	0.01	0.62	
Inferred Deletion	No Error Exposure	0.25	0.17	0.01	0.63	

Inferred Insertion	Intercept	7.33	2.02	4.35	12.38	
Inferred Insertion	Deletion Exposure	0.25	0.18	0.01	0.65	
Inferred Insertion	Insertion Exposure	0.29	0.19	0.02	0.73	
Inferred Insertion	Exchange Exposure	0.18	0.15	0.01	0.55	
Inferred Insertion	No Error Exposure	0.14	0.12	0.01	0.44	
Inferred Exchange	Intercept	1.54	0.25	1.14	2.1	
Inferred Exchange	Deletion Exposure	0.15	0.11	0.01	0.41	
Inferred Exchange	Insertion Exposure	0.13	0.1	0	0.38	
Inferred Exchange	Exchange Exposure	0.12	0.09	0	0.34	
Inferred Exchange	No Error Exposure	0.12	0.09	0.01	0.34	
Inferred No Error	Intercept	1.77	0.28	1.3	2.4	
Inferred No Error	Deletion Exposure	0.2	0.16	0.01	0.59	
Inferred No Error	Insertion Exposure	0.27	0.2	0.01	0.73	
Inferred No Error	Exchange Exposure	0.2	0.15	0.01	0.57	
Inferred No Error	No Error Exposure	0.15	0.11	0.01	0.43	
Participants (880 levels)						
Inferred Deletion	Intercept	1.49	0.08	1.33	1.66	
Inferred Insertion	Intercept	1.25	0.09	1.07	1.43	
Inferred Exchange	Intercept	1.45	0.06	1.32	1.57	
Inferred No Error	Intercept	1.54	0.11	1.34	1.76	

Table B. Summary of parameter estimates from a multi-level multinomial model of how participants edited test sentences (whether they Inferred a Deletion, an Insertion, an Exchange, No Error, or Other, which was set to 0), based on what Exposure condition they were in (Mixed—the reference level, Deletions, Insertions, Exchanges, or No Errors). Participants were included as random intercepts. Items were included as random intercepts with random slopes for Exposure condition (correlation between slopes and intercepts was not estimated). The model was fit using the "brms" package for Bayesian Multilevel Modeling in R (Bürkner, 2017) with weakly regularizing priors for all population-level effects (normal distribution with a mean of 0 and standard deviation of

10). Bold rows indicate parameters for which there is evidence for an effect of Exposure condition on a particular response category probability, relative to the Mixed Exposure condition (either because the 95% Credible Interval, CI, doesn't contain 0 or over 90% of the posterior distribution exceeds 0). Broadly, these results are in line with the logistic regression analyses and the data patterns observed in Figure 1: Inferred Deletions were most likely when participants were in the Deletion Exposure condition, Inferred Insertions were most likely in the Insertions Exposure condition and Inferred Exchanges were most likely in the Exchange Exposure condition (though the 95% CI does overlap with 0). Unlike in the previous analysis, there is not strong evidence for a higher probability of Inferred No Error responses in the No Error Exposure condition compared to the Mixed Exposure condition. This may be partially due to the overall very low number of Inferred No Error responses across items (see Appendix C).

Appendix C

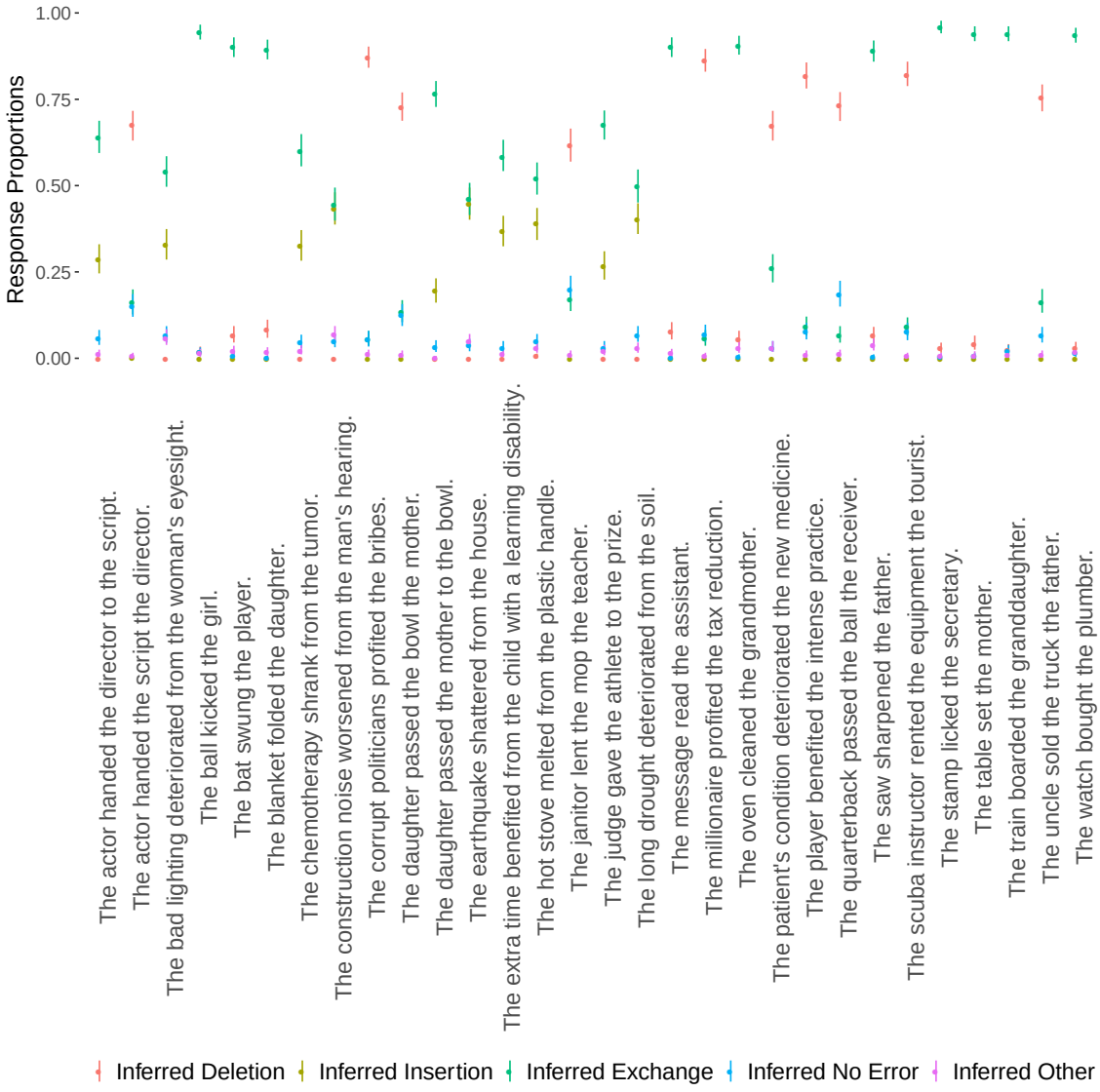

Figure C. Proportion of responses by type for each test item. (Each participant saw 15 items).

Supplementary Material

Raw data, materials, analysis code, and pre-registration available at https://osf.io/rkrha/