Structure-sensitive Noise Inference: Comprehenders Expect Exchange Errors

Till Poppels (tpoppels@ucsd.edu)[1]
[1]Department of Linguistics
University of California, San Diego
San Diego, CA 92093 USA

Roger P. Levy (rplevy@mit.edu)[1,2]
[2]Department of Brain & Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

Previous research has found that comprehenders are willing to adopt non-literal interpretations of sentences whose literal reading is unlikely. Several studies found evidence that comprehenders decide whether or not a given utterance should be taken at face value in accordance with principles of Bayesian rationality, by weighing the prior probability of potential interpretations against the degree to which they are (in)consistent with the literal form of the utterance. While all of these results are consistent with string-edit noise models, many error processes are known to be sensitive to the underlying linguistic structure of the intended utterance. Here, we explore the case of exchange errors and provide experimental evidence that comprehenders' noise model is structure-sensitive. Our results add further support to the noisy-channel theory of language comprehension, extend the set of known noise operations to include positional exchanges, and show that comprehenders' noise models are well-adapted to structure-sensitive sources of signal corruption during communication.

Keywords: rational analysis; noisy-channel comprehension; non-literal interpretation;

Introduction

Evidence that comprehenders adopt non-literal interpretations when the literal meaning of a sentence is implausible or otherwise unlikely has been around for at least 15 years. For example, Christianson, Hollingworth, Halliwell, and Ferreira (2001) found that garden-pathed readers partially retain the thematic role assignment associated with their initial misinterpretation even when it is incompatible with the literal sentence once the garden path has been resolved. Ferreira (2003) subsequently showed that such tendencies arise not only from garden-path constructions, but also in ordinary sentences with implausible literal interpretations, particularly when implausible events are expressed using non-canonical linguistic forms (e.g., in passive voice). These observations raise the questions *when* and *how* comprehenders decide whether or not a given utterance should be taken at face value and receive a literal interpretation, and if not, what alternative interpretation should be adopted instead.

According to the noisy-channel (Shannon, 1949) theory of language comprehension, interpretations arise rationally and gradiently through probabilistic inference (Levy, 2008). As per usual in the tradition of Rational Analysis (Anderson, 1990; Chater & Oaksford, 1999), the noisy-channel theory takes as a starting point the hypothesis that comprehension is the statistically optimal solution to the problem of communicating under noise. The structure of this problem is schematically represented in Fig. 1: The speaker intends to convey meaning M by encoding it linguistically in a structured representation S with surface form w. For example, if S is the intended syntactic tree, w may represent the word string that corresponds to the intended yield of the tree. More generally, we take w to be an ordered, but otherwise unstructured, string of atomic elements (e.g., words)[1], and S to contain any additional information, such as the lexical category of words, their functional position in a phrase structure, etc. /Thus, S and w characterize the underlying structure and surface form of the speaker's *intended* utterance, and jointly determine the *actual* utterance u. Since this is where the speaker's intention takes physical shape, u can now be processed by the comprehender to produce I, the input to the interpretation process, which finally results in the comprehender's inferred meaning M'.

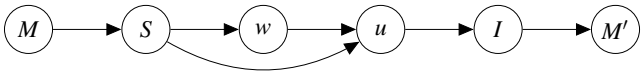

Figure 1: Schematic representation of the noisy channel.

Communication is successful when M' recovers M, which is threatened by the possibility of signal corruption at various points along this noisy channel. For example, phonetic reduction may cause the deletion of words between w and u, and interference during lexical retrieval may result in the accidental substitution of intended words with other words. These are two examples of (production-based) noise processes that may cause u to differ from w, but, and this is crucial for the present paper, these processes may be modulated by the structural information in S. For example, phonetic reduction is more common in function words than in content words (Bell, Brenier, Gregory, Girand, & Jurafsky, 2009), and accidental substitutions typically involve words of the same grammatical category (Harley & MacAndrew, 2001). Likewise, phonological exchange errors (e.g., "**c**oat-**thr**utting") frequently involve phonemes that occupy the same syllable position or share certain phonetic features (Ellis, 1980). More generally, then, the effect of S on u, captured by the curved edge in Fig. 1, is manifest in what we call "structure-sensitive" errors: it modulates the operation of noise processes between w and u, based on structural information about the atomic elements in w. The structure-sensitivity of $w \rightarrow u$ noise processes makes them qualitatively different from the noise processes that transform the actual utterance u into the input I to the comprehension system. The latter are physical and neural processes including environmental noise, precision limits of perceptual systems,

[1]In principle, these atomic units could also represent phonemes or orthographic characters, but the present paper focuses on word-based operations.

and imperfection in memory traces of previous input representations. These operate on the actual utterance u and thus must be conditionally independent of the upstream variables w and S, which have representational status only internal to production.

Given this representation of the communicative process, the comprehender's decision whether or not a given utterance should be interpreted literally, comes down to the conditional distribution $P(M'|I)$. We hypothesize that comprehenders infer this distribution by rationally integrating their prior expectations about meanings with the likelihood of the observed input under their model of the noisy channel:

$$P(M|I) \propto P(I|M)P(M) \tag{i}$$

This general characterization of comprehension as Bayesian inference underscores the importance of understanding the noise model, $P(I|M)$, that comprehenders use to reverse-engineer the process that may have generated the input they observed. In keeping with the tradition of Rational Analysis, we take as a starting point the noise model of an optimal comprehender (Fig. 2), which mirrors the production process and thus explicitly represents the possibility of various types of noise corruption, as discussed above.

The noisy-channel proposal has found general support in online reading-time studies (Levy, Bicknell, Slattery, & Rayner, 2009; Levy, 2011; Bergen, Levy, & Gibson, 2012), as well as in offline measures of comprehension accuracy (Gibson, Bergen, & Piantadosi, 2013). However, there are two important limitations in these previous studies, which are addressed in the present paper: First, the noise operations that comprehenders have been shown to detect and repair during interpretation so far only contain deletions, insertions, and substitutions, which represent only a small subset of the noise processes known to affect production. One prominent source of noise that is known to operate at various levels of linguistic representation involves the accidental exchange of elements (e.g., phonemes, affixes, or words), yet no previous study has found evidence that comprehenders expect and repair such errors (in fact, as discussed below, Gibson et al. (2013) report evidence that appears to be inconsistent with this possibility). A second, related limitation of previous research is that all of the extant evidence is compatible with structure-*in*sensitive noise models, since all of the noise processes that comprehenders have been shown to repair in these studies are string-based operations that do not require reference to, or knowledge of, the underlying structure of the input.

The central difference between such structure-*in*sensitive noise models and the structure-sensitive model we are testing in the present paper is captured by the dashed edge in Fig. 2: in a structure-*in*sensitive noise model, this connection does not exist and w' d-separates u' and S'. Consequently, the underlying structure of the intended word string has no bearing on the actual utterance except through the word string itself. If, however, as we hypothesize, comprehenders do consider the possibility of structure-*sensitive* noise operations,

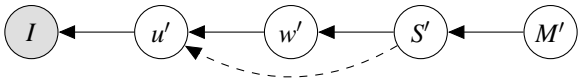

Figure 2: The comprehender's noise model ideally mirrors the production process. u', w', and S' are the comprehenders model of u, w, and S (cf. Fig. 1); the dashed edge corresponds to the structure sensitivity hypothesis pursued in the present paper. The shading emphasizes that only I is known to the comprehender and forms the basis for inferring M' (cf. Eq i).

their noise model completely mirrors the production process as represented in Fig. 1, and inferring the intended meaning of an utterance requires marginalizing over w', S', and u':

$$P(M'|I) \propto P(M') \int_{w',S',u'} P(I|u')P(u'|w',S')P(w'|S')P(S'|M')\,\mathrm{d}w'\,\mathrm{d}S'\,\mathrm{d}u' \tag{ii}$$

To test the structure sensitivity hypothesis and address limitations of previous research, we test whether comprehenders consider the positional exchange of words a potential source of signal corruption. Initial support for this hypothesis comes from the abundance of exchange errors in the phonological domain (MacKay, 1987), which comprehenders appear to repair routinely during comprehension. For example, upon encountering the phrase "coat-thrutting", the non-literal alternative "throat-cutting" appears to surface naturally (MacKay, 1987). It is important to note at this point that detecting exchange errors does not require structure sensitivity *per se*, since the exchange of two randomly sampled elements in a string is possible without reference to (or knowledge of) the underlying structure. However, if comprehenders expect the exchange of some elements but not others, that would support the structure sensitivity hypothesis because that distinction does require reference to the underlying structure of the surface string.

Gibson et al. (2013) found evidence that suggests that comprehenders do not consider the positional exchange of at least some words. They presented participants with sentences that varied in terms of semantic plausbility and syntactic structure and probed their interpretation through comprehension questions. Overall, comprehenders readily adopted non-literal interpretations when (a) the literal utterance was semantically implausible and (b) a more plausible alternative interpretation could be obtained by postulating the deletion or insertion of individual words. However, implausible transitive sentences, such as (1-a), were almost always interpreted literally, although a more plausible reading was available on the assumption that the two nouns had been exchanged by mistake (1-b). The finding that comprehenders retained the faithful but implausible reading in such cases suggests that the word exchange that would recover the more plausible reading is not a likely operation under the their noise model.

(1) a. The ball kicked the girl.
b. The girl kicked the ball.

However, the observation that comprehenders do not contemplate exchanges in such cases does not imply that their noise model precludes any possibility of exchange errors. For example, the noise model may place higher probability on the exchange of function words than content words or favor exchanges out of syntactic adjuncts compared to complements, which would make the noun-noun exchanges that are required for repairing (1-a) a low-probability noise operation.

The present paper provides evidence from applying the experimental paradigm used in the Gibson et al. to further explore the status of exchange errors under comprehenders' noise model. We presented participants with sentences whose thematic role assignment was either plausible or implausible on a literal interpretation and, crucially, could be reversed through the positional exchange of prepositions. If comprehenders are tempted by these non-literal interpretations, that would suggest not only that exchange errors *are* among the noise operations they consider, but also that their noise model is structure-sensitive in that it assigns greater probability to the exchange of some elements (e.g., prepositions) than others (e.g., nouns).

Methods

Participants were presented with sentences that were plausible or implausible on a literal interpretation, paired with comprehension questions that were designed to distinguish between literal and non-literal interpretations in a 2-alternative forced-choice task. The materials were constructed to afford alternative interpretations through word exchanges. To foreshadow the results, participants' response patterns strongly suggest that these exchanges were indeed among the noise operations they considered during comprehension.

Participants. 60 self-reported native speakers of English were recruited via Amazon.com's Mechanical Turk. 1 participant with less than 75% accuracy on filler items was excluded from the analysis.

Materials. Following a 2x2 within-subject design, the semantic plausibility of the literally described event was crossed with the order of prepositional phrase (PP) adjuncts, as illustrated in Table 1. More precise estimates of the semantic plausibility of the utterances were obtained in a separate norming experiment and the canonicality of PP orders was estimated in a corpus analysis (both described below).

Plausibility	PP order	The package fell...
plausible	canonical	...from the table to the floor.
plausible	non-canonical	...to the floor from the table.
implausible	canonical	...from the floor to the table.
implausible	non-canonical	...to the table from the floor.

Table 1: Example item in 2x2 within-subject design.

The literally encoded event is either plausible (package falling down) or implausible (falling up), and in each case the other reading is available by exchanging the prepositions while leaving everything else in place. Both plausibility and canonicality affect the prior probability of the literal interpretation: since the implausible event is less likely to occure, it is less likely to reflect the speaker's communicative event, but irrespective of her intended meaning, she is more likely to express it using the canonical PP order "from...to..." than the non-canonical "to...from...".[2]

Procedure. Participants were presented with 20 experimental items in a random order, interspersed with 48 filler items. In a full latin square, one of the four versions of each item was displayed together with a yes/no comprehension question that was designed to distinguish between plausible and implausible thematic-role assignments. For example, the sentences in Table 1 were followed by the question *Did something fall to the floor?*, to which completely literal comprehenders would respond "Yes" following the plausible sentences and "No" following the implausible ones. Optimal noisy-channel comprehenders, on the other hand, would weigh the possibility of an exchange error against the prior probability of the candidate interpretations, and thus may exhibit the reverse response pattern when the literal parse is semantically implausible and/or syntactically non-canonical. Across items, the plausibility of the comprehension question was counterbalanced to ensure that the literally correct answer was equally often "Yes" and "No".

Canonicality norming. To estimate the relative canonicality of PP orders, we performed a frequency analysis of the Brown corpus (Kucera & Francis, 1967) and the Wall Street Journal (Paul & Baker, 1992) sections of the Penn Treebank (Marcus, Santorini, & Marcinkiewicz, 1994), using Levy and Andrew's (2006) `Tregex` software package for tree searches.[3] The relative frequency with which each preposition pair occurred in the more frequent order is summarized in Fig. 3. For the purpose of the 2x2 design, each utterance was categorized as canonical or non-canonical (based on the threshold indicated by the dashed line), but the analyses reported below made use of the continuous estimates.

Plausibility norming. To quantify differences in plausibility between sentences both within and across items, we conducted a separate norming experiment, in which participants ($n = 12$) rated the plausibility of the events described by the sentences in question (e.g., *The package fell from the table to*

[2]The corpus analysis described below found that in over 90% of sentences with PP adjuncts headed by "from" and "to" the former preceded the latter (cf. Fig. 3).

[3]The queries for this analysis were, for example, `__< (@PP <<# from $++ (@PP <<# to))` and `__< (@PP <<# to $++ (@PP <<# from))`, for preposition pair *from-to*.

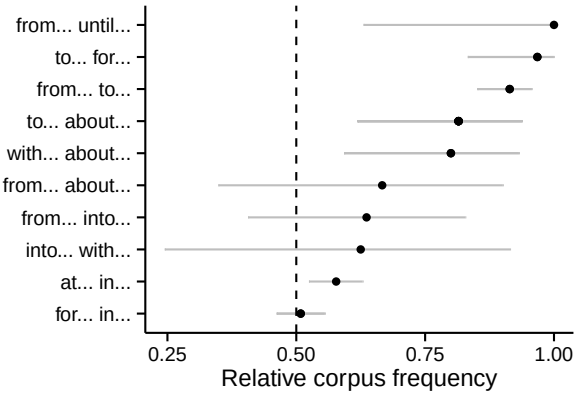

Figure 3: Relative frequency of canonical order of each PP pair. The dashed line indicates the threshold for categorizing orders as canonical vs. non-canonical. Errorbars represent Clopper-Pearson confidence intervals for proportions, calculated with the `PropCIs` R package (Scherer, 2014).

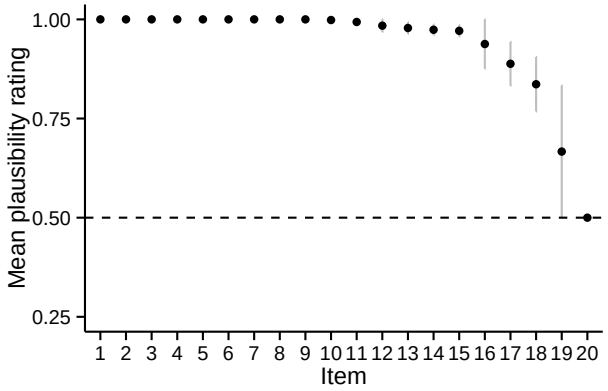

Figure 4: Mean normalized plausibility ratings of canonical PP exchange items. As in Fig. 3, the dashed line indicates the categorization threshold for establishing the 2x2 design. Errorbars represent SEs of the mean.

the floor vs. *The package fell from the floor to the table*), using two independent sliders, each ranging from "completely plausible" to "completely implausible". Only canonical item versions were used, so that the sentence pairs that participants rated differed only in their thematic role assignment, and not in the order of the prepositions. The two measures obtained for each sentence pair were normalized and scaled to range over the interval [0,1] by dividing each by their sum.[4]

Replication of Gibson et al. experiments. To situate our results within the context of the Gibson et al. (2013) findings, we replicated three of the experiments that were reported in that paper, each testing a different syntactic alternation. We chose one that according to Gibson et al. had received mostly non-literal interpretations (DO/PO benefactives), one that produced a balanced number of literal and non-literal interpretations (transitive/intransitive), and one that triggered almost no non-literal interpretations (active/passive). Example sentences from each experiment are shown below:

(2) Active/passive:
  a. The girl kicked the ball. [plausible]
  b. The girl was kicked by the ball. [implausible]

(3) Transitive/intransitive:
  a. The chemotherapy shrank the tumor. [plausible]
  b. The chemotherapy shrank from the tumor.[impl.]

(4) Direct-object/prepositional-object benefactives:
  a. The father bought his son a bicycle. [plausible]
  b. The father bought his son for a bicycle. [impl.]

[4]For example, if the plausible sentence was rated as "completely plausible" (corresponding to a score of 100), and the implausible alternative received the score 20, the normalized plausibility scores were $\frac{100}{100+20} \approx 0.83$ and $\frac{20}{100+20} \approx 0.17$, respectively.

The items used in these three experiments were included in the norming experiment described above, which presented all items in a within-subject design to ensure that plausibility norms were comparable across experiments (cf. Fig. 5).

Predictions. If accidental exchanges of prepositions are possible under comprehenders' noise model, we expect them to adopt non-literal interpretations at least on some trials. More specifically, the noisy-channel model predicts that the extent to which comprehenders perform such noise inferences should be inversely proportional to the prior probability of the literal utterance, and should therefore be driven by the semantic plausibility as well as syntactic canonicality of the displayed sentence.

If comprehenders' noise model is structure-sensitive, they may assign different probabilities to exchanges of some elements compared to others. Thus, we predict that comprehenders adopt non-literal interpretations in the case of prepositional exchanges more readily compared to the noun-noun exchanges that would permit the repair of implausible active/passive sentences. In other words, the rate of noise inference should be higher for sentences like *The package fell from the floor to the table* compared to sentences like *The ball kicked the girl*.

Results

Plausibility and canonicality norming. The mean normalized plausibility ratings of the PP exchange items are shown in Fig. 4, summarized alongsides the replication materials in Fig. 5 (since ratings for plausible sentences and their implausible counterparts summed to 1, only plausible versions are shown). Overall, sentences were rated as either highly plausible or highly implausible (all normalized mean ratings for plausible items were above 0.9), although the PP exchange materials were slightly less polarizing than the materials used

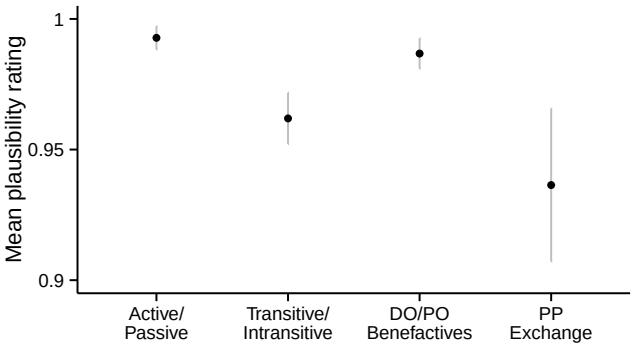

Figure 5: Mean normalized plausibility ratings for plausible items from all 4 experiments. Errorbars show SEs of by-item means.

in the Gibson et al. (2013) study, especially the active/passive sentences.

Noise inference. Fig. 6 shows the results from all four experiments, broken up by construction and binned plausibility. Consider first the replications of the Gibson et al. experiments in panels 1, 3, and 4 (from the left). As expected, accuracy on plausible items (light grey) was largely at ceiling for all of the experiments, and comparable to the mean accuracy on filler items (white), which is consistent with Gibson et al.'s results. Implausible items (dark grey) received non-literal interpretations to different extents across experiments. Most notably, implausible passives triggered noise inferences significantly more often than plausible passives ($p = 0.02$) and marginally more often than implausible actives ($p = 0.1$). This result is consistent with Ferreira's (2003) findings and slightly different from what Gibson et al. (2013) reported, who found passives to be interpreted literally on more than 95% of trials. In all other respects our results closely replicate those from the Gibson et al. study.

The central prediction of the present study was that implausible (and non-canonical) sentences would provoke non-literal interpretations that can be reached through the positional exchange of elements within the sentence. We replicated Gibson et al.'s finding that such exchanges appear to be impossible—or at least highly unlikely—in the case of active/passive constructions, which were overwhelmingly interpreted literally, even when they described highly implausible events (cf. panel 1 in Fig. 6). However, as illustrated in panel 2, implausible sentences with exchangeable PP adjuncts did receive non-literal interpretations on 31-37% of trials, while their plausible counterparts were interpreted literally more than 90% of the time, which is comparable to the mean comprehension accuracy on filler items. Consistent with this, a binomial mixed-effects regression analysis[5] revealed a main effect of plausibility ($\beta = 1.303$, $p < 0.0001$) as well as a small, but significant, main effect of canonicality ($\beta = 0.276$, $p = 0.038$). Both main effects were in the predicted direction: non-literal interpretations were more likely for utterances with implausible semantics and non-canonical form than for those that described plausible events and made use of the more canonical PP order.

Finally, the proportion of literal interpretations of implausible PP exchange items was significantly lower compared to implausible active/passive sentences ($p < 0.001$).

[5]The analysis was carried out using the `lme4` R package. The full formula was `response ~ plausibility * canonicality (1 + plausibility * canonicality || item) + (1 + plausibility * canonicality || subject)`.

Discussion

The results presented here demonstrate that word exchanges are among the noise operations comprehenders consider when interpreting implausible or otherwise unlikely utterances. Moreover, they suggest that comprehenders' noise model is structure-sensitive, since the possibility of exchanging two words for gains in plausibility was exploited in the case of PP adjuncts, but not active/passive constructions.

It is not clear from our results, however, exactly what elements comprehenders expect to be subject to exchanges. Notice that the plausible interpretation of

(5) [$_{PP}$ to [$_{NP}$ the table]] [$_{PP}$ floor [$_{NP}$ the floor]]

can be achieved by exchanging either the prepositions or the nouns (or noun phrases). The significant effect of canonicality does suggest that prepositions were considered the object of the exchanges at least some of the time because exchanging nouns does not achieve a change in the order of prepositions and can therefore not explain the effect of canonicality. However, this does not imply that noun-noun exchanges are impossible. Although we have exemplified possible structure-sensitive constraints on exchanges in terms of lexical categories or content/function word status throughout the paper, it is also possible that the crucial difference between PP exchanges and noun-noun exchanges in active/passives is that the former involve adjuncts whereas the latter affect elements in complement position. It is possible that either or both of these constraints exist, but since our data do not allow us to test either of these admittedly post-hoc explanations directly, they remain in the realm of speculation.

Conclusion

We have reported experimental evidence that comprehenders expect the communicative signal to be corrupted by accidental word exchanges, and that this expectation is structure-sensitive. These results build on previous findings suggesting that the language comprehension machinery chooses rationally from a range of possible interpretations. Going beyond previous formulations of comprehenders' noise model, we showed that it more completely mirrors the production process than previously assumed. Thus, our results add to a growing pile of evidence that language comprehension is well-adapted to the problem of communication across a noisy channel, and makes use of all available information in the

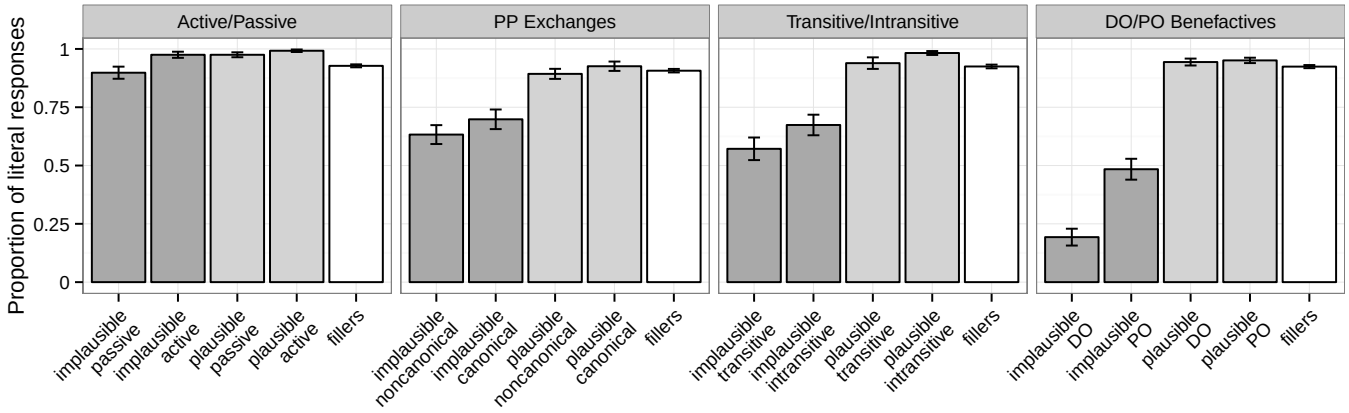

Figure 6: Proportion of literal responses across experiments. Errorbars represent SEs of by-subject means.

search for the interpretation that is most likely to reflect the speaker's communicative intent.

Acknowledgments

This work has benefitted from discussion with audiences at the Linguistics BrownBag at UCSD, AMLaP 2015, and the 2016 Annual Meeting of the LSA. We gratefully acknowledge support from NSF grant IIS-0953870, NIH grant HD065829, and an Alfred P. Sloan Research Fellowship to RPL.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92–111.

Bergen, L., Levy, R., & Gibson, E. (2012). Verb omission errors: Evidence of rational processing of noisy language inputs. In *Proceedings of the 34th annual meeting of the Cognitive Science Society* (pp. 1320–1325).

Chater, N., & Oaksford, M. (1999). Ten Years of the Rational Analysis of Cognition. *Trends in Cognitive Sciences*, *3*(2), 57–65.

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

Ellis, A. W. (1980). Errors in speech and short-term memory: The effects of phonemic similarity and syllable position. *Journal of Verbal Learning and Verbal Behavior*, *19*(5), 624–634.

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

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

Harley, T. A., & MacAndrew, S. B. (2001). Constraints upon word substitution speech errors. *Journal of Psycholinguistic Research*, *30*(4), 395–418.

Kucera, H., & Francis, W. (1967). The Brown University Standard Corpus of Present-Day American English (Brown Corpus). *Providence: Brown University*.

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

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In (pp. 1055–1065). Association for Computational Linguistics.

Levy, R., & Andrew, G. (2006). Tregex and tsurgeon: tools for querying and manipulating tree data structures. In *In 5th international conference on language resources and evaluation.*

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

MacKay, D. G. (1987). Constraints on theories of sequencing and timing in language perception and production. *Language perception and production: Relationships between listening, speaking, reading and writing*, *3*, 407.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Paul, D. B., & Baker, J. M. (1992). The design for the Wall Street Journal-based CSR corpus. In *Proceedings of the workshop on speech and natural language* (pp. 357–362).

Scherer, R. (2014). *PropCIs: Various confidence interval methods for proportions.* R package version 0.2-5.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, *37*(1), 10–21.