Don't Underestimate the Benefits of Being Misunderstood (A Replication)

Andrew Yang
Ling 245B Methods in Psycholinguistics
Judith Degen

Introduction

Often, those who speak English with an accent may feel embarrassed or less self-confident. However, in certain cases, speaking with an accent may actually provide an "advantage" - that is, native speakers of English may sometimes interpret implausible statements uttered by accented speakers as plausible. In this writeup, I present my replication of "Don't Underestimate the Benefits of Being Misunderstood," a 2017 paper by Edward Gibson, Caitlin Tan, Richard Futrell, Kyle Mahowald, Lars Konieczny, Barbara Hemforth, and Evelina Fedorenko.

In the remainder of this section, I will discuss the overall theoretical background and relevant hypotheses behind this research. In the next section, I will outline the parameters of the original experiment, and discuss the ways in which my replication deviated from the original experiment. Then, I will discuss possible reasons for why some of my findings did not replicate the original experiment.

This research follows a Bayesian view of language processing. Say we have an intended sentence *i* and a perceived sentence *p*. Given that the comprehender perceives *p*, what is the probability that they assign to *i*? In Bayesian terms, this is modeled by the following relation:

$$\Pr[i \mid p] \propto \Pr[i] \times \Pr[i \to p]\,.$$

Here, the notation $\Pr[i \to p]$ denotes the probability that i gets corrupted to p. Under the assumption that $\Pr[i \to p]$ is generally higher for non-native speakers of English (and assuming that i and p are not overly "far" one one another), it makes sense that implausible utterances are more likely to be interpreted as plausible if the speaker has an accent.

Let's take a look at a more concrete example that involves the *dative alternation*. This example shows the sixth target trial from the dative alternation sentences in the original paper:

a. DO, plausible: The girl tossed the boy the apple.
b. DO, implausible: # The girl tossed the apple the boy.
c. PO, plausible: The girl tossed the apple to the boy.
d. PO, implausible: # The girl tossed the boy to the apple.[1]

[1] Note that DO means "double object" and PO means "prepositional-phrase object."

Now, if the listener hears the implausible DO sentence (say, (b)), the probability that they assign a plausible interpretation (either (a) or (c)) is given by

$$\Pr[a \vee c \,|\, b] \propto \Pr[a]\,\Pr[b \to a] + \Pr[c]\,\Pr[b \to c]\,.$$

First, note that the prior probabilities $\Pr[a]$ and $\Pr[c]$ (referring to plausible scenarios) are most likely higher than $\Pr[b]$ and $\Pr[d]$. This helps push the comprehender in the direction of accepting a plausible interpretation (i.e. $a \vee c$) over accepting an implausible interpretation (i.e. $b \vee d$).

Next, the quantities $\Pr[b \to a]$ and $\Pr[b \to c]$ depend on a number of factors, which we won't attempt to compute here. In any case, we note that the corruption $b \to a$ is a switch in the ordering of two NPs, will may or may not be very likely. The corruption $b \to c$ is the addition of the function word *to*, which is certainly possible but perhaps less likely than the deletion of a *to*, as in the corruption $c \to b$.

Experiment

In this section, we discuss the set-up of the original experiment. To be clear, the original experiment was divided into six sub-experiments. I only replicated the fourth sub-experiment. But, for the sake of clarity, when I say "the original experiment," I shall refer to the fourth sub-experiment in the original paper.

Original experiment

The original experiment proceeded as follows. Each respondent was given an 80-trial survey. Each trial consists of an audio snippet of a sentence, and a corresponding yes/no question. For example, a trial might involve a sentence like *The toddler hugged a teddybear because he was scared*, and a corresponding question like *Did the toddler hug someone/something?*. To proceed through the experiment, respondents had to listen to the audio and answer the corresponding questions.

The auditory materials were produced by two speakers. Speaker 1 is able to speak in near-native American English, as well as with a heavy Israeli accent. Speaker 2 is able to speak in native American English, as well as with a heavy Hindi accent. In each survey, 60 of the sentences were filler sentences, and 20 were target sentences. One speaker always read *all* of the filler sentences, and the other speaker always read *all* of the target sentences.

Target sentences followed the pattern in the DO/PO sentences shown in the previous section. Target sentences were either exclusively DO, or exclusively PO. Of the 20 target sentences per trial, 10 of them were plausible and 10 of them were implausible. The filler sentences did not follow any of the the DO/PO patterns, and all described plausible scenarios.

Note that there are *three* sets of binary parameters that the experiment alternates between:

DO S1 target S1 has accent	DO S1 target S1 no accent	DO S2 target S2 has accent	DO S2 target S2 no accent
PO S1 target S1 has accent	PO S1 target S1 has accent	PO S2 target S2 has accent	PO S2 target S2 has accent

It can be seen that this arrangement gives rise to 8 sub-experiment conditions.

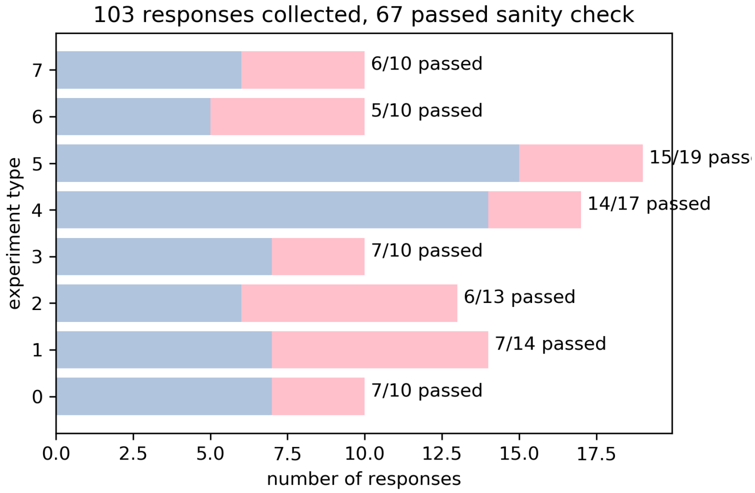
Gibson et al. (2017) 320 gathered responses on Amazon Mechanical Turk. Respondents were paid, though the paper does not state how much. Participants were excluded from analysis if they did not have a United States IP address, if their native language was not English, if they filled out more than one survey, or if they incorrectly answered more than 25% of the filler questions. According the the paper, respondents took around "10 - 15 minutes" to complete the survey.

Replication

For my replication, I tried to keep as many experimental parameters faithful to the original experiment as I could. However, this was not always possible given budgetary constraints and my own technical limitations.

First, I reduced the number of filler trials from 60 per experiment to 40 per experiment. Each time, I randomly sampled 40 filler sentences from the original 60[2] provided by the authors in their OSF repository. I also surveyed only 103 workers on Mechanical Turk, paying $2.75 per survey ($3.30 per survey including the Mechanical Turk fee).

Also, the original experiment sampled equal numbers of participants for each sub-experiment condition. However, for each respondent I picked randomly from the eight sets of conditions. **Figure 1** shows the number of participant I gathered for each set of conditions, as well as the fraction of participants that passed the filler trial questions (namely, by answering at least 75% of the filler questions correctly). DO trials were encoded as 0 through 3, and PO trials were encoded as 4 through 7. Even indices encode target speakers with an accent, and odd indices encode target speakers without an accent. Types 0, 1, 4, 5 encode set-ups where the target speaker was the one who could speak in a heavy Israeli accent, and the rest encode set-ups where the target speaker could speak in a heavy Hindi accent.

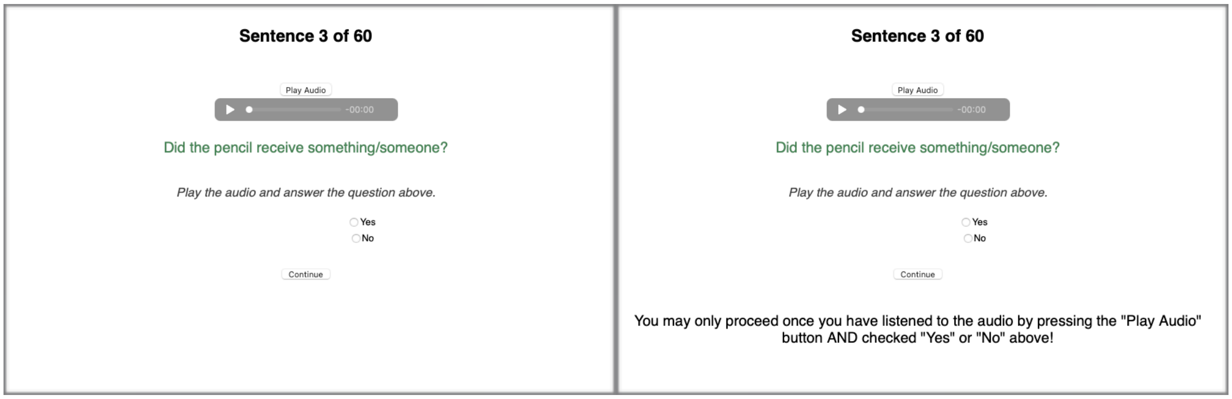


[2] To be precise, two of the 60 were excluded, as I used them as example trials at the beginning to familiarize respondents with the experiment.

Figure 1. Responses collected by sub-experiment conditions, and how many participants passed the sanity check (answering 30 out of 40 (75%) filler questions correctly).

After I collected 80 to 90 responses, I noticed that the randomization in experimental parameters resulted in a deficiency of type 4 and 5 sub-experiments, which correspond to PO, Israeli-accented speaker, with/without accent. As a result, I tried "balancing out" the distribution by collecting more responses for those two set-ups. Because it takes a nontrivial amount of time for all assignments in a HIT to finish, I posted more assignments than was necessary, and planned to terminate the HITs once I gathered enough responses. However, it turned out that after I manually expired those HITs, I still had to wait for the "Pending" assignments to complete - this unfortunately resulted in a markedly higher number of responses for those two set-ups. In hindsight I probably should not have done this, although my overall results do not vary if I exclude these additional data.

In my experiment, participants had to click on a "Play Audio" button and answer the yes/no question before proceeding (see **Figure 2** below). If the participant does not press "Play Audio" or does not answer the yes/no question, a message appear to remind the participant to do those things.



Sentence 3 of 60

Did the pencil receive something/someone?

Play the audio and answer the question above.

○ Yes
○ No

Sentence 3 of 60

Did the pencil receive something/someone?

Play the audio and answer the question above.

○ Yes
○ No

You may only proceed once you have listened to the audio by pressing the "Play Audio" button AND checked "Yes" or "No" above!

Figure 2. My experiment.

As a result of my own lack of know-how in web programming, I could not devise a way to ensure that the *entirety* of the audio had finished playing before the participant could advance to the next trial. This meant that, in theory, a participant could press "Play Audio," quickly answer the yes/no question, and then successfully advance to the next trial. Regardless, I assume that if a participant could answer 75% of the filler questions correctly, then they probably weren't exploiting the experiment in this way.

Also, before the actual trials began, participants were shown two example trials with two hardcoded filler examples. In addition to playing the audio and answering the yes/no questions, participants were asked to type in the last word they heard in the audio (the words were "bees" and "loyalty"). Participants could not move on unless they entered those words correctly (up to simple misspellings and capitalization differences). I reasoned that these two example trials would encourage respondents to listen to the complete audio snippets during the actual experiment. These two hardcoded filler trials were excluded from the 40 filler trials during the experiment.

The original study also asked an additional cohort of respondents to transcribe the target sentences to determine if the words in each target sentence were hard to understand. For reasons of budget and time, I did not do this. The authors of the original study noted that this led to the exclusion of a small number of target sentences from analysis, but did not specifically state which ones, so I excluded this consideration from my analysis entirely.

In addition, the original paper did not specify the dates over which the study was conducted. My experiment, however, was conducted first over the weekend of Friday May 21, 2021 through Sunday May 23, 2021, and then over the weekend of Friday May 28, 2021 through Saturday May 29, 2021. Finally, to use up the remaining $10.00 in my Mechanical Turk account, I conducted three final trials in the afternoon of Tuesday June 1, 2021. In hindsight I probably should have ran the trials during mornings from Mondays through Thursdays, but unfortunately I was not aware that data tends to become more noisy on outside of those hours.

My OSF registration is available at https://osf.io/egt8s. All of the code and data used for my replication is available on Github at https://github.com/ycm/ling245b-replication.

Results

In this section I will present the results that I obtained in my replication.

Noisier data

First, the data I collected was much noisier than what was reported in the original paper. The authors of the original paper report that the mean correctness on filler trials was "over 90%." However, this value was a mere 80% on my filler trials, leading to the exclusion of over a third of the responses (see **Figure 1**).

Effect of accent on implausible DO/PO sentences

The primary result in the original paper was that implausible sentences were more likely to admit a plausible interpretation if the sentence was read with a heavy accent. This is summarized in the following table[3], which juxtaposes the original findings with the corresponding finding from my replication:

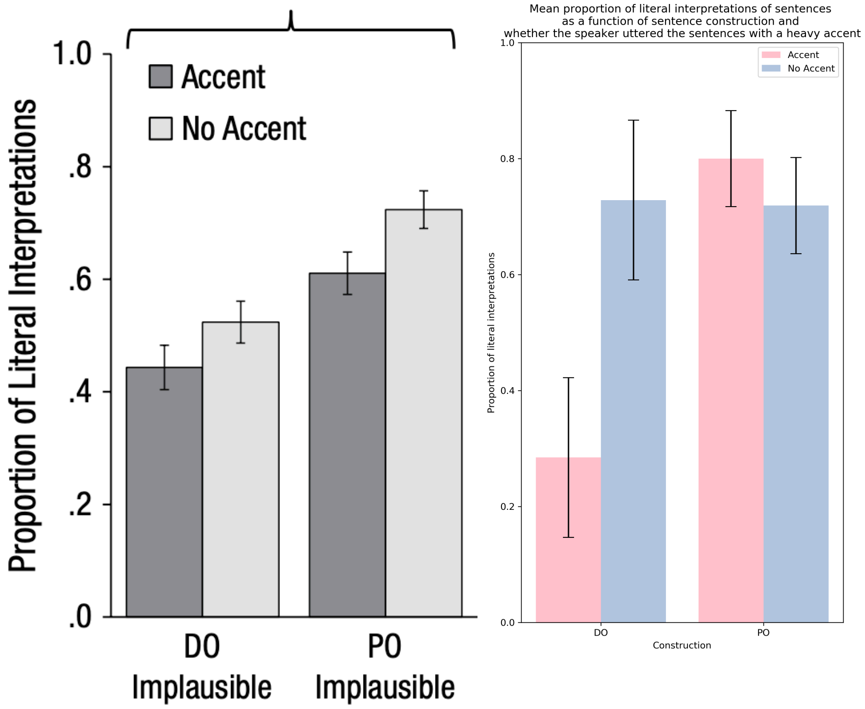


Table 1. Proportion of literal interpretations (namely, implausible interpretations assigned to implausible sentences) versus DO/PO alternation and whether the target speaker spoke with an accent. Error bars denote 95% confidence intervals.

[3] Because numerical values were not provided in the paper, I could not generate the original bar plots on my own.

Since only implausible target sentences were considered for analysis both in the original analysis and in my replication, the height of each bar in **Table 1** shows the proportion of implausible target sentences for which the respondent answered the corresponding yes/no question in such a way that implies they assumed the implausible interpretation.

In the original experiment, having an accent appears strongly correlated with the respondent favoring a plausible interpretation of an implausible sentence. However, my results suggests that this only applied to DO constructions. In fact, my results suggest the opposite is true for PO constructions, though the 95% confidence intervals for the PO trials overlap considerably, likely as a result of my smaller sample size. Furthermore, my results suggest that there is a much larger difference between accent and no-accent trials for DO trials: about 30% of [DO, +accent] trials resulted in an implausible interpretation, compared to about 45% in the original paper; and about 75% of [DO, -accent] trials led to an implausible interpretation, compared to less than 55% in the original paper. For PO trials, [PO, +accent] and [PO, -accent] resulted in about 80% and 70%, respectively, whereas the original paper reported about 60% and over 70%. I will provide a more in-depth discussion of this in the **Discussion** section to follow.

The authors of the original study also supply a numeric breakdown of the results for each of the eight sub-experiment conditions. Because I had the numeric values this time, I was able to plot these values alongside my own, shown in **Table 2** below:

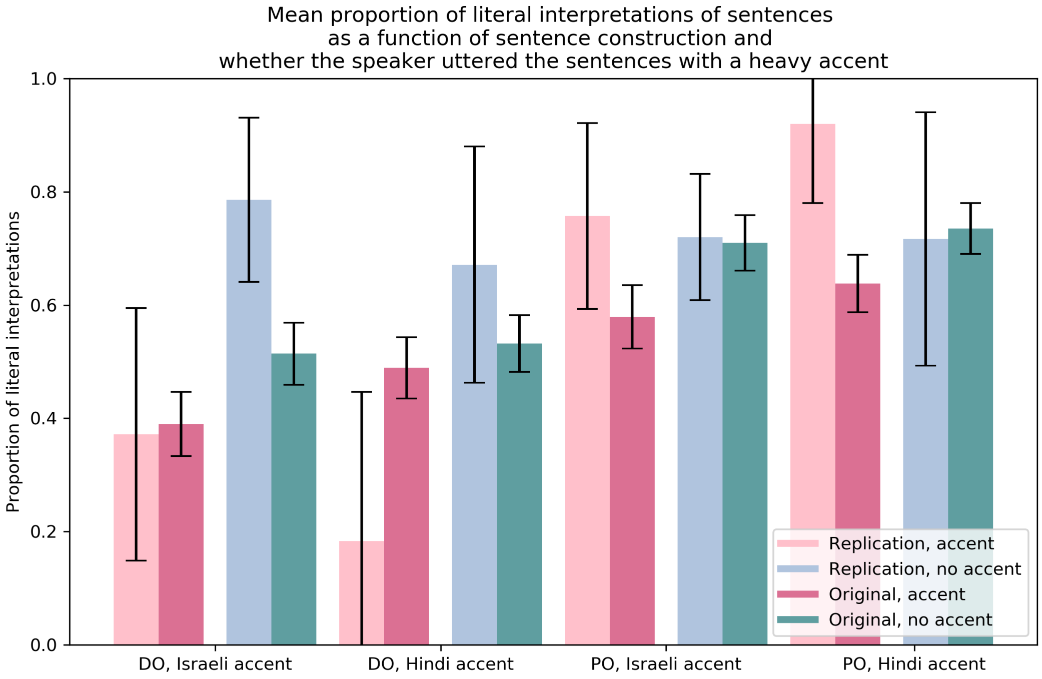


Table 2. Proportion of literal interpretations of implausible sentences versus DO/PO alternation and whether the target speaker spoke with an accent, by sub-experiment condition. Error bars denote 95% confidence intervals.

For the original study, the aggregate results in **Table 1** are consistent with the by-sub-experiment results in **Table 2** (shown by the darker-colored bars). For my replication, the aggregate results that I found were also consistent with my own results in **Table 2**, which suggests the actual identity of the target speaker was not a confounding factor for my deviation from the original results.

Discussion

Before analyzing why the DO results successfully replicated while the PO results did not, I first asked why my results were noisier than those in the original study (i.e. why only 80% of my respondents answered at least 75% of the filler questions correctly).

I see three possible reasons for this. First, it could be possible that I did not pay respondents enough. The original study does not state how much respondents were paid, so it is possible that they paid significantly more. For what it's worth, however, 97 out of 103 participants reported that the pay was fair. Secondly, my experiment may have been lacking in aesthetic value. The only aesthetic difference between my experiment and the supplied experiment template was a change of font from serif to sans-serif, so it is possible that a more visually appealing experiment webpage could have led to less noisy results. Third, my experiment was conducted over weekend days, which could result in noisier data, as participants might be more fatigued or less interested on those days. Due to time constraints, I did not conduct an analysis of the number of filler questions completed correctly versus the time of day that a respondent completed a survey. Still, this seems like a plausible reason for the large difference between my results and those reported in the original paper.

Now I turn to the main difference between my results and the results from the original paper - that is, why my DO trials show a significantly larger difference as result of accent, and why my PO trials suggest the opposite effect than what was reported by the original paper.

First, I provide an explanation for why participants were much more likely to adopt a plausible interpretation for implausible [DO, +accent] sentences. Due to my smaller number of filler sentences, it is possible that participants were more likely to distinguish between filler and target sentences, in the sense that they "expected" that they were supposed to interpret accented sentences in a plausible way.

In particularly, Jurafsky et al. (1998) show that the function word *to* has a strong tendency for "reduction" in conversational speech[4], as recorded in the Switchboard corpus (Godfrey et al., 1992). Further, all of the target trials in the original Gibson et al. (2017) study shared the following form:

DO: The *N V* the *N* the *N*.
PO: The *N V* the *N* to the *N*.

Since each noun was preceded by *the* (which always began with a coronal consonant regardless of accent), it is possible that participants were more likely to assume that implausible [DO, +accent] sentences like (1) contained a (very) reduced *to*, as in (2).

[4] That is, a basic form like [tu], [tʉ], or [ɾu] is very often reduced to [tə], [tɨ], or [ə].

(1) # The daughter passed the bowl the mother.
(2) The daughter passed the bowl ($\text{to}_{\text{reduced}}$) the mother.

On the other hand, for [DO, -accent] trials, the smaller number of filler sentences in my replication could have led participants to believe that sentences like (1) were truly meant to "trick" them, thus leading them to adopt the implausible interpretation.

For PO trials, the analysis is much more obscure. A straightforward explanation would be to fall back on my limited sample size - indeed, only 40 PO responses were included for analysis, and the 95% confidence intervals overlap considerably, as shown previously in **Table 1**.

Still, I will attempt to offer another explanation for my observed results. To do so, I examined the verbs used in the target sentences, as well as the overall number of times they occurred in the DO and PO forms in the Switchboard corpus, as reported in Bresnan et al. (2007) and supplied in the `dative` table in the `languageR` package in R. There were 10 verbs that were each used twice in the target sentences, shown in Table 3.

Verb	DO occurrences	PO occurrences
mail	7	7
give	1410	256
sell	34	172
lend	10	10
lease	0	4
toss	N/A	N/A
pass	N/A	N/A
rent	N/A	N/A
hand	9	6
throw	N/A	N/A

Table 3. Verbs appearing in target sentences, and their DO/PO occurrences (if any) in the Switchboard corpus

Due to the missing counts, it is unfortunately not possible to provide a full analysis of the PO results. However, *if* it is the case that the verbs in target trials are more favored in the DO construction, it is possible that respondents would actually assume that a non-accented (native) speaker of English is *more likely to have intended a plausible (DO) interpretation*.

To illustrate this point, I provide the following pair of target sentences, which contains the DO-favored verb *give*:

(3) # The mother gave the daughter to the candle.
(4) The mother gave the daughter the candle.

If uttered by a non-accented (in a sense, native) speaker of English, (3) may be more likely to be interpreted as (4), the plausible version. Again, the smaller number of filler trials may have played a role here; since every sentence in a [PO, -accent] survey was uttered without an accent, respondents who could distinguish between filler and target trials may have interpreted the experiment as "expecting" that all trials lead to plausible interpretation. Similarly, implausible target sentences in [PO, +accent] trials may have resulted in a higher proportion of implausible interpretations because the respondents' subconscious representations of English grammar led them to disfavor the possibility of (4) corrupting to (3) when uttered by a non-native speaker.

It should be apparent that this explanation is quite weak - for this reason, I conducted a mixed-effects logistic regression on the implausible PO target trials I collected. Each target sentence in an each survey is assigned one data point. The response variable was whether the respondent had a plausible interpretation of the sentence, and the only effect was a random effect due to the choice of verb in the sentence. **Table 4** below shows the modes of the random effects due to the choice of verb:

give	-0.04316053
hand	0.18544615
lease	0.03734856
lend	0.04591625
mail	-0.11738836
pass	-0.06985102
rent	-0.05047437
sell	0.21626270
throw	-0.03152855
toss	-0.18639975

Table 4. Random effect modes by verb, after fitting a mixed-effect logistic regression model on implausible PO trials. The formula was *plausible interpretation* ~ 1 + (1 | *choice of verb*).

Verbs like *hand* were more likely to elicit a plausible interpretation, while verbs like *toss* were more likely to elicit an implausible interpretation. However, if we account for whether or not the speaker has an accent as well, the result mixed-effected logistic regression gives that the *p*-value for the effect of accent on whether the respondent takes a plausible interpretation is around 0.18, regardless of whether the choice of verb is accounted for as a random effect. As such, we can cannot directly conclude with much confidence whether having an accent has an effect on the plausible or implausible interpretations.

As a result, the data I collected may not be sufficient for concluding that PO sentences are more or less likely to elicit plausible interpretations based on the whether the target speaker has an accent.

Problems with Original Experimental Design

Though I was not able to fully replicate the results of the original experiment, I will still comment on two primary concerns I have with the overall experimental design.

Incongruences between target and filler trials

On average, sentences contained 9.8 words, while target trials only contained 7.65 words. Filler sentences varied greatly in length - the standard deviation of sentence lengths across filler trials was 2.7, whereas the standard deviation of sentence lengths across target trials was only 0.6. These incongruences suggest that participants may have been able to distinguish between target and filler trials, especially in my replication where there was a 2:1 filler-to-target ratio.

Moreover, the filler sentences were always uttered by one speaker and the target trials were always uttered by the other. Even in surveys where the target speaker spoke without an accent, it can be relatively easy for participants to distinguish between the two voices. A possible improvement to this, then, is to introduce another sub-experiment condition where both filler and target trials were uttered by the same speaker. Alternatively, we can run experiments with more filler speakers and more target speakers.

A problem with using edit distance as a proxy for the likelihood for corruption

In the original study, the authors state that the *edit distance* between two sentences is a proxy for the likelihood of corruption; intuitively, it is reasonable to say that a sentence *A* is more likely to be corrupted to *B* if *A* and *B* only differ by (say) a single insertion or deletion. However, the authors do not account for potential differences in the *costs* of these edits. For instance, it is possible to conjecture that *to*-deletion has a lower cost than *to*-addition - this is a strong argument for why, both in the original study and in my replication, implausible DO sentences were significantly more likely to admit a plausible interpretation compared to implausible PO sentences.

Further, the cost of adding or deleting *to* is compounded by the relative preferences of the specific verb in question - indeed, we can infer from **Table 3** that verbs like *lease* may have a stronger preference for PO constructions over DO constructions.

References

Bresnan, J., Cueni, A., & Nikitina, T. (2007). Predicting the dative alternation. *Cognitive Foundations of Interpretation*, 69-94. https://web.stanford.edu/~bresnan/CFI04.pdf

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, 28*(6), 703-712. https://doi.org/10.1177/0956797617690277

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). SWITCHBOARD: telephone speech corpus for research and development. *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. 1*, 517-520. https://doi.org/10.5555/1895550.1895693

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. (1998). Reduction of English function words in Switchboard. *International Conference on Spoken Language Processing*. http://www.icsi.berkeley.edu/pubs/speech/reductionofenglish98.pdf