Suhas Arehalli and Eva Wittenberg*

Experimental filler design influences error correction rates in a word restoration paradigm

<https://doi.org/10.1515/lingvan-2020-0052>

Received June 11, 2020; accepted March 30, 2021; published online August 30, 2021

Abstract: Including fillers or distractors in psycholinguistic experiments has been standard for decades; yet, relatively little is known how the design of these items interacts with critical manipulations. In this paper, we ask about the role that contextual statistical information in filler items plays in determining if and how to correct a given error, and how grammatical expectations interact with context. We first replicate a speech restoration experiment conducted by Mack, J. E., C. Clifton, L. Frazier & P. V. Taylor. 2012. (Not) hearing optional subjects: The effects of pragmatic usage preferences. *Journal of Memory and Language* 67. 211–223, measuring usage preferences of null-subject constructions. Then we report two additional experiments in which we manipulated only the filler items, either having noise appear uniformly at random, or with a particular bias. Our results (1) demonstrate that listeners are sensitive to statistical patterns in the distribution of noise within the experiment, and (2) suggest that this paradigm can be used to investigate interaction between the mechanisms that govern grammatical preferences, and those that govern error correction processes.

Keywords: error correction; experimental design; expletives; filler structure; psycholinguistics; replication; subject deletion

1 Introduction

Sentence processing is context-dependent. Speakers constantly adapt to their linguistic environments, sometimes resulting in different acceptability judgements for the same sentence (e.g., Keller 2000; Sorace and Keller 2005). Thus, understanding the relationship between the grammar, contextual information, and processing is crucial for understanding the language processing system as a whole.

The relationship between context and grammar is also methodologically important: many psycholinguistic arguments hinge on the assumption that a sentence is processed in isolation. To lend a basis to this assumption, filler items are often included in an experiment, to water down any contextual effects that interfere with the experimental manipulation: being submerged in a sea of random sentences, the critical manipulation is thought to be less salient, and participants do not adapt to it (Cowart 1997). However, there has been little systematic research on how filler design influences processing, although there is evidence that not enough fillers lead to repetition effects or unnatural comprehension strategies (e.g., Havik et al. 2009). In this paper, we will attempt to help characterize one particular instance of this kind of relationship: the relationship between statistical information about the distribution of acoustic noise in the context and error correction processes.

*Corresponding author: Eva Wittenberg, Department of Linguistics, University of California, San Diego, La Jolla, CA, USA, E-mail: ewittenberg@ucsd.edu. <https://orcid.org/0000-0002-3188-6145>

Suhas Arehalli, Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA, E-mail: suhas@jhu.edu

1.1 Designing experimental fillers

The design details of linguistic experiments matter. Taking the widespread example of acceptability judgments, a recent paper has found that several task features, from the mode of stimuli presentation, the number of response options, and the use of response labels, significantly influence the sensitivity of an experimental paradigm (Marty et al. 2020; also see Schütze 2016). More generally, it has been shown that the number of both participants and items (Mahowald et al. 2016), and the kind of participants — native speakers versus second-language learners, or linguistics experts versus laypeople — (Dąbrowska 2010; Gross and Culbertson 2011; Schütze and Sprouse 2014; Wasow and Arnold 2005) will likely make a difference in experimental results.

While there is a dense literature on the design of judgment data, recent methods used by (psycho)linguists are not limited to probing linguistic intuitions, but include the measurement of reaction times, eye-movements, or brain activity. However, apart from increasingly popular one-shot tasks (e.g. several experiments in Morgan et al. 2020; von der Malsburg et al. 2020), most of these experimental setups have one design feature in common: the customary inclusion of fillers.

It is all the more surprising, then, that the role of filler items has not been addressed in more detail by the psycholinguistic literature. Here, we start addressing this issue by manipulating the two design decisions which usually influence the creation of filler items: the *ratio* of critical items to filler items, and the *structure* of the filler items themselves.

The first question of the correct *ratio* is most often discussed based on recommendations to include fillers in order to make critical items less visible across the experiment, by reducing the latters' density, and, in turn, reduce the likelihood that participants detect the critical manipulations and consciously adapt their responses (Cowart 1997; Keller 2000; Schütze and Sprouse 2014). For instance, Marinis and Cummings (2018: 14), advise a ratio of “at least 1:1”. However, as Table 1 shows, there is large variation in the critical-to-filler ratio, ranging from three critical items for every two fillers to four critical items for every twenty-one fillers.

Table 1: An overview of critical-to-filler ratios in a sample of psycholinguistic papers published in the journal *Glossa* in 2020.

Experiment	Critical items	Filler items	Ratio
Green et al. 2020, Exp. 1	30	20	3:2
Green et al. 2020, Exp. 2	30	60	1:2
Grant et al. 2020, Exp. 1	30	52	15:26
Grant et al. 2020, Exp. 2	24	126	4:21
Feiman et al. 2020, Exp. 1	32	64	1:2
Feiman et al. 2020, Exp. 2	16	64	1:4
Franck et al. (2020)	32	64	1:2
Pañeda et al. (2020)	32	72	4:9
Baumann and Schumacher (2020)	40	120	1:3

Some of this variation can be explained by the practice of experimental ‘piggy-backing’: independent experiments are combined, so critical items in one study serve as the next study’s fillers; this usually results in higher ratios. Examples of this strategy are Grant et al.’s (2020) Experiment 2, or Pañeda et al. (2020). The assumption underlying this strategy is that unrelated fillers do not affect a given study’s outcome; by combining multiple experiments into one, more participants can be run, improving each study’s statistical power.

However, what makes this strategy potentially problematic is that it has been observed that the ratio of fillers to critical items affects effect size (Bodner et al. 2006; Green et al. 2020; Harrington Stack et al. 2018), and indeed, some experiments explicitly manipulate the number of their fillers in order to understand the obtained effects better. For instance, Green et al. (2020) increased the number of fillers between Exps. 1 and 2 by a factor of three, and found that their result of interest became weaker.

The other question of interest is that of the filler items' *structure*. In the acceptability and grammaticality judgment literature, it has been argued that filler design can, for instance, ensure that the full scale of rating options is used, by including fillers that have certain grammatical or semantic properties (Schütze and Sprouse 2014). Indeed, many experiments are designed with this reasoning in mind, and changing the structure of the filler items has been shown to alter the effects observed (e.g., Garnham et al. 1992).

The picture that emerges is that filler properties can have important effects on the experimental outcome, although their purpose in the first place is to ensure that the critical experimental manipulations can unfold their effect undisturbed by confounds. This presents researchers designing experiments with a problem: how can they ensure that the ratio and structure of their fillers do not introduce a bias of theoretical relevance?

One approach to understanding filler effects is from the larger perspective of context effects: as a participant is building up a representation of the experimental situation, statistical information about the stimuli she encounters plays a major role. One corner of the psycholinguistic literature that has leveraged this insight particularly well is that of error correction.

1.2 Error correction paradigms and statistical information within an experiment

When there is noise in the signal, whether it be background noise, competing auditory signals, or the product of speech errors, listeners often effortlessly correct for it (Levy 2008). This process is necessarily controlled by at least two factors: low-level auditory information, and high-level, top-down expectations derived from contextual information and grammatical knowledge broadly construed (e.g., Poppels and Levy 2016; Warren 1970). Prior work has shown that error correction processes are also sensitive to a third factor, long-term statistical information in the signal (Gibson et al. 2013), and that in some cases listeners adapt their processing strategy based on that information (Schotter et al. 2014). This suggests that the statistical properties of noise throughout an experiment, including filler items, also contributes to the correction process.

This principle is most straightforwardly examined by considering acoustic noise that prevents a listener from hearing the speaker's full production. For example, if a listener hears "Noah and I can't decide when to #### the tomatoes," (with #### representing a segment of the signal where some acoustic noise obscures the identity of the underlying word), they would need to repair the utterance before it can be properly processed – that is, infer what was intended to be in place of the noise to form a Gestalt of a sentence. When doing this, the listener follows top-down expectations: in the utterance above, grammatical expectations would predict that the noise covers a verb, like "plant" or "eat." Of course, listeners would also be sensitive to bottom-up information, like auditory information gleaned from the noise – does the acoustic information under the noise sound more like /p/ or /i:/?

However, another source of information lies in (subconsciously) taking into account patterns in what we will refer to as the *distribution* of the noise. Formally, this refers to the probability that a particular word in the original, uncorrupted signal, given its context, becomes noisy before being heard by the listener. More practically, this probability distribution characterizes the listener's model of the process that generates the noise, and any biases it might have that would help the listener discern what lies under the noisy segments they hear.¹ In the context of the example above, if most of the noise in a context appears over what is likely to be a word that begins with a vowel, listeners might be more likely to believe that the noisy segment was meant to be "eat" rather than "plant." This project investigates whether and how this third source of information – the listener's beliefs about the distribution of the noise in the context – interacts with grammatical expectations when correcting for noise.

¹ Note that what we call the distribution of noise is a special case of what is referred to as the *noise model* or *noise likelihood function* in the noisy-channel framework (Gibson et al. 2013, Ryskin et al. 2018): while a proper noise model specifies the likelihood of a word being corrupted in a particular manner (i.e., whether it is deleted or exchanged with another word), we assume only one kind of corruption, the superimposition of acoustic noise over the word, and have the noise distribution refer simply to the probability that particular kind of corruption is placed over a particular word in its context.

A similar question has been investigated by Gibson et al. (2013). In their experiments, participants were asked to read a series of sentences and answer a comprehension question for each. Critical items consisted of sentences that were semantically implausible (i.e., “The girl was kicked by the ball”) paired with a question that probed whether the subject accepted the sentence as written (and adopted an implausible reading) or corrected the sentence to obtain a more plausible reading (i.e., “Did the girl kick something?”). Crucially, the rate of ungrammatical and implausible sentences in the filler items was manipulated between subjects (none vs. a 3:8 ungrammatical to grammatical ratio, and a 1:8 vs. a 5:16 implausible to plausible ratio). Gibson et al. (2013) found both that an increase in syntactic violations in the fillers lowers the rate of literal interpretations of critical items, and that an increase in semantically implausible sentences increased the rate of literal interpretation. These results were taken to indicate that the distribution of potential noise (in this case, in the form of syntactic errors and semantically implausible sentences) affects the error correction process.

However, this paradigm has limitations: first, as the authors mention, participants were allowed to read and re-read each stimulus and corresponding comprehension question as often as desired. Thus, these results cannot distinguish between automatic correction processes and other, conscious corrections. Second, this measure is coarse-grained, as the syntactic errors and semantic implausibilities had to be strong enough to cause listeners to reject the literal reading of a grammatical sentence for a correction to be detected. Finally, Gibson et al. (2013) manipulate the frequency of implausible and ungrammatical sentences in the filler items, which may be unnatural enough to draw the conscious attention of participants to the manipulation.

We aim to reduce some of these shortcomings by adopting a different methodology for investigating the interaction between grammatical knowledge and the distribution of noise in error correction. To do this, we adapt the Speech Restoration paradigm (Warren 1970) to investigate the interaction between grammatical knowledge and automatic error correction processes. In this paradigm, participants listen to stimuli with white noise superimposed over segments of the speech, including one critical region designed such that the addition of noise causes the identity of the underlying segment to be ambiguous. Participants are then instructed to repeat exactly what they had heard. Given this instruction, participants will not consciously correct any errors in the stimuli, and thus any “restoration” of the linguistic material underneath the ambiguous noise in the critical region can be taken as the result of automatic, unconscious error correction processes.

Thus, in order to study the manner in which grammatical knowledge and statistical information interact in error correction, we can simply manipulate the distribution of the superimposed noise across filler items, and measure the resulting effect on error correction rates. This paradigm, and the particular implementation of this paradigm in Mack et al. (2012), offers several advantages over the method used in Gibson et al. (2013): first, the task does not rely on participants’ rejections of literal readings of stimuli, which should lead to a greater sensitivity to factors that influence error correction. In addition, the use of a single-shot auditory stimulus with an “exact” repetition task minimizes the effect of post-comprehension processes on our measure: if participants are asked to report exactly what they believe they heard (which ostensibly includes speech errors) the only repairs from the heard material to the reported material should be those caused by automatic, unconscious processing, though some conscious reconstruction due to memory limitations will inevitably occur. Finally, since the design only requires that statistical information (i.e., the distribution of noise) in the filler items influences the interpretation of the ambiguous noise, the paradigm can be extended to domains in which the literal syntax versus plausible semantics choice provided to participants in Gibson et al. (2013) cannot be used.

We demonstrate the effectiveness of this paradigm by testing a hypothesis about the interaction of the effect of grammatical preference and that of the distribution of noise. Specifically, we suggest that in situations where noise is distributed uniformly at random (hereafter the *random noise* condition), listeners correct noise based on top-down expectations like grammaticality, since a listener can discover no informative pattern in the noise distribution through which they can recover the segment underneath the noise and thus their best remaining option is to correct the segment using their prior linguistic knowledge. In situations where there is highly systematic distribution of noise (in the above example, a high likelihood that the noise always covers a vowel), a listener relies less on factors like grammaticality, since they now have more direct information about the uncorrupted signal: while our grammatical knowledge can help us determine what we are likely to hear in general, knowledge about the current conversation can more directly inform us as to what a particular speaker

was likely to have said in this context. We additionally predict that different grammatical expectations vary in their robustness to this manipulation: stronger constraints will withstand the pressure to rely on distributional information more than weaker rules.

This hypothesis is motivated by the observation that, while Bayesian integration of both factors would be the ideal mechanism to integrate multiple signals into the decision making process (see the Noisy Channel Model (Gibson et al. 2013; Levy 2008; Ryskin et al. 2018)), this neglects the costs associated with tracking and manipulating all available information (Wittenberg and Jackendoff 2018: among many others). An alternative strategy might therefore be to adjust the weight of top-down and bottom-up information based on statistical patterns in the materials (for a similar question, see Schotter et al. 2014). Thus, the listener would rely heavily on top-down, grammatical information more when the inferred noise distribution is uninformative, but would prioritize using the more contextually relevant information from the inferred noise distribution when it is informative.

1.3 Current studies

In three experiments, we test the influence of different distributions of noise in the filler structure on the restoration of a noise covered function word: optional, sentence-initial “it” (Yoon 2001). We gradually increase the ratio of systematically distributed noise to random noise between experiments (see Figure 1) by including both fillers with random and systematic noise (Exp. 1, grey box in Figure 1), and then selectively only fillers with random (Exp. 2, red box in Figure 1) and systematic noise (Exp. 3, green box in Figure 1), while keeping the critical items constant across experiments.

Figure 1: A diagram of the materials used in each of our experiments. Each dot represents an item used in Experiment 2 of Mack et al. (2012), with yellow dots representing the critical items, dark green dots representing fillers with an optional, sentence initial “that,” and other colored dots representing fillers with a different structure. Exp. 1 (grey box) replicated Mack et al. (2012) and used all of the items. Exps. 2 and 3 used only a subset of the fillers, with Exp. 2 using only the 24 “that”-fillers and Exp. 3 using 24 of the other fillers. Crucially, noise was systematically placed over the sentence-initial “that” in “that”-fillers, while each noise segment was distributed uniformly at random throughout the other filler items.

2 Experiment 1: replication

Given that we will be working with a new population, both in space (speakers in Southern California vs. Massachusetts) and time, we will first attempt to replicate Mack et al. (2012).

English usually requires an explicit grammatical subject. Mack et al. (2012) tested whether ungrammaticality due to the elision of subjects is ameliorated in utterances that express “immediate” judgments — judgements that are made immediately before a corresponding utterance is produced (for example, “ $\emptyset$ seems to me like it’s raining”). They manipulated two factors of immediacy: temporal immediacy, and personal immediacy, capturing the intuition that immediate judgments are likely to be about events that happened

recently, and that are likely to express the speaker's opinion rather than those of others. Since tense (past vs. present) and person (first vs. third person) are the grammatical realization of these factors, they predicted (and subsequently found) that in a speech restoration paradigm, participants would produce more null subject constructions in first person, present tense sentences ("... seems to me ...") as opposed to third person, past tense sentences ("... seemed to her ..."; see Figure 2 for sample sentences).

- (1)
 - a. Noah and I can't decide when to plant the tomatoes.
 - b. Why?
 - c. (FIRST/PRES) ### seems to me like the weather isn't cooperating. ###
 (THIRD/PRES) ### seems to her like the weather isn't cooperating. ###
 (FIRST/PAST) ### seemed to me like the weather wasn't cooperating. ###
 (THIRD/PAST) ### seemed to her like the weather wasn't cooperating. ###
- (2)
 - a. ### What did the reporter reply?
 - b. ### the local news is boring.
- (3)
 - a. I think the latest candidate for mayor is pretty gutsy.
 - b. Oh yeah?
 - c. ###ew politicians seldom directly criticize their opponents.

Figure 2: (1) An example critical stimulus, (2) An example filler stimulus with sentence-initial "that"; (3) A filler stimulus with random noise. ### denotes noise.

While the original study used this paradigm to measure the strength of grammatical preferences, we view this methodology as a tool to measure listeners' error correction processes: since participants hear sentences with noise in place of the subject and then are asked to repeat what they heard, they are, in effect, producing what segment they believe was masked by auditory noise. English grammar constrains the possibilities for what lies underneath the noise to either an "it" or nothing, providing us with a simple binary measure of the result of the participant's error correction process based on participants' top-down expectations.

2.1 Methods

2.1.1 Participants

We recruited 48 self-identified English native speakers from the subject pool of the University of California, San Diego (37 female; mean age: 21.2; age range: 18–35) for course credit.

2.1.2 Stimuli and design

Mack et al. (2012) generously provided their original stimuli (as shown in Figure 3), all of which were used without modification in our replication. The 24 critical stimuli (Figure 2) were three-utterance dialogues ending in a potential null-subject construction with the subject position masked by noise. Each critical stimulus also contained a randomly placed additional segment of noise. Each item was manipulated across two factors: TENSE (past or present) and PERSON (first or third).

The 60 fillers were either two- or three-utterance dialogues. These stimuli each contained two or three segments of noise distributed throughout the dialogue (see Figure 2).

2.1.3 Procedure and analysis

After each trial, participants were recorded repeating the final sentence of the dialogue using a simple user interface.² They were then asked to type their response into a text box before moving on to the next trial. Before the experiment began, participants were instructed to repeat exactly what they had heard, in an effort to ensure

² The experiment's code, along with our data and analysis code, can be accessed at <https://osf.io/jcu75/>.

Exp. 1:
Mack et. al.

Figure 3: Each dot represents an item used in (Mack et al. 2012)'s Experiment 2, with yellow dots representing the critical items, dark green dots representing fillers with noise over an optional, sentence initial “that,” and other colored dots representing fillers with a different structure and all noise distributed randomly. Exp. 1, like (Mack et al. 2012), used all of the items.

that participants did not consciously correct errors in the stimuli. Six practice trials, including informal and nonstandard language, preceded the 84 filler and critical trials. Participants saw six trials in each condition (manipulations of TENSE and PERSON in a Latin square design). Trials were presented in random order. This design, including the instructions and the typed component of the response, was chosen in order to match the design of Mack et al. (2012) as closely as possible.

Participants' spoken responses were transcribed by research assistants blind to the purpose of the study. The transcriptions were then automatically coded for the presence of a sentence-initial “it” and fit with a logistic mixed effects model using R and the lme4 library (Bates et al. 2015). Models were selected by first fitting to the model predicting the “it” restoration rate from predictors TENSE and PERSON with maximal random effects structure, and removing interactions between random effects until the model converged.³

2.2 Results and discussion

The mean restoration rate (% of expletive subjects produced) was 55.2%, lower than in the original experiment (67%). We replicated the effects of immediacy: participants were more likely to produce “it” in past tense than in present tense sentences, and in third person more than first person sentences (see Figure 4). Just as in Mack et al. (2012), TENSE was significant ($\beta = -0.73$, $p < 0.01$, $\chi^2(1) = 11.29$), and PERSON was marginally significant ($\beta = -0.44$, $p = 0.051$, $\chi^2(1) = 3.780$) with no significant interaction ($\beta = 0.11$, $p > 0.80$, $\chi^2(1) = 0.062$).

We found a lower “it”-restoration rate than the original experiment, and consider this an interesting likely effect of language change, where our participants found null subject constructions more acceptable.

Crucially for our purposes, Exp. 1 successfully replicated the results found in Mack et al. (2012), showing that manipulating tense and person does affect the rate at which listeners restore the “it” at the beginning of critical items in the predicted direction: first person sentences are restored less than third person sentences, and that present tense sentences are restored less than past tense sentences.

This allows us to use this methodology to investigate how grammatical preferences interact with other influences on the correction process in Exps. 2 and 3. We manipulate the level of systematicity in the distribution of the noise in the filler items to test whether softer grammatical violations will cease to be detected with a reduced filler ratio (Exp. 1 vs. Exps. 2 and 3): based on both Mack et al. (2012) and our results, while the effect of TENSE may weaken slightly, the marginal effect of PERSON should weaken or disappear with a reduced filler-to-critical-item ratio, since, under prediction (2), weaker grammatical effects (like that of PERSON) are liable to

³ We chose an approach different from the forward-selection process used in Mack et al. (2012), in which models were made incrementally larger until the model fit did not improve beyond a threshold, due to concerns about the conservativity of such methods (Flom and Cassell 2007).

Figure 4: Results from Exp 1. Participants were more likely to produce expletive subjects in place of noise in past-tense, third person sentences than in present-tense, first person sentences. Error bars are 95% CIs.

weaken or disappear entirely in environments with stronger cues from the distribution of noise. Crucially, reducing the filler ratio necessarily increases the strength of cues from the distribution of noise, since all of the critical items have noise systematically placed over the segment containing the expletive subject

We now test whether this ratio alone, or also the distribution within filler items, influences error correction behavior. In order to do this, we selectively changed the number and structure of fillers in the experiment.

3 Experiment 2

In Exp. 2, we selected 24 filler items with randomly distributed noise, while using the same critical items as in Exp. 1, resulting in a 1:1 ratio of fillers to critical items. This allowed us to understand whether and how changing the ratio of critical items to fillers affects the rate of restoration, as well as the effects we observed in Exp. 1.

3.1 Methods

3.1.1 Participants

A power analysis using *simr* (Green and MacLeod 2016) determined that 77 subjects would be required to obtain 80% power to identify a main effect of noise structure between Exps. 2 and 3. We recruited these 77 self-identified English native speakers from the subject pool of the University of California, San Diego (52 female; mean age: 20.49; age range: 18–33) for course credit.

3.1.2 Stimuli and design

The stimuli and design were identical to that of (Mack et al. 2012) and Exp. 1, except that we only used a 24-item subset of the fillers in which the noise was distributed randomly (see Figures 2 and 5). Thus, the ratio of structured noise (critical items) to random noise (fillers) was increased to 1:1.

3.1.3 Procedure and analysis

The procedure and analysis were identical to those of Exp. 1.

Exp. 2: "Random" Noise

Figure 5: Each dot represents an item used in Exp. 2, with yellow dots representing the critical items, dark green dots representing fillers with noise over an optional, sentence initial “that,” and other colored dots representing fillers with a different structure and all noise distributed randomly. Exp. 2 used the 24 critical items and 24 fillers with randomly distributed noise.

3.2 Results and discussion

The mean “it” restoration rate was 20.95%. The effect of TENSE was significant ($\beta = -0.64$, $p < 0.05$, $\chi^2(1) = 6.21$). In contrast to Exp. 1, the effect of PERSON was not significant ($\beta = -0.23$, $p = 0.30$, $\chi^2 = 1.07$), and neither was the interaction of TENSE and PERSON ($\beta = 0.74$, $p = 0.15$, $\chi^2 = 2.06$). That is, participants were again more likely to produce expletive subject constructions in past-tense sentences than in present tense sentences, but the marginal effect of person disappeared. As in Exp. 1, the interaction was not significant (Figure 6). Thus, TENSE seems to be a stronger grammatical preference than PERSON, whose effect was only marginal in Exp. 1, and is not significant with a higher proportion of structured noise in the materials (Exp. 2).

Figure 6: Results from Exp. 2. Participants were more likely to produce expletive subjects in place of white noise in past-tense, third person sentences than in present-tense, first person sentences.

4 Experiment 3

Exp. 3 used only fillers containing noise overlaid on the sentence-initial function word “that,” which was always optional in the contexts used (see Figures 2 and 7; e.g. Ferreira 1997). Thus, participants were only

exposed to items with systematically distributed noise: in the critical items, with noise laid over “it”, and in the fillers, with noise laid over “that.”

Figure 7: Each dot represents an item used in Exp. 3, with yellow dots representing the critical items, dark green dots representing fillers with noise over an optional, sentence initial “that,” and other colored dots representing fillers with a different structure and all noise distributed randomly. Exp. 3 used only the 24 critical items and the 24 “that”-fillers.

4.1 Methods

4.1.1 Participants

A different set of 78 native speakers from the subject pool of the University of California, San Diego participated (56 female; mean age: 20.36; age range: 18–33) for course credit.

4.1.2 Stimuli and design

The stimuli and design were identical to the previous studies, except that the 24 fillers used were all two-utterance dialogues ending in sentence fragments beginning in “that” (see Figures 2 and 7). Each of these sentence-initial “that”s were masked in noise, with an additional segment of noise overlaid elsewhere in the dialogue at random.

4.1.3 Procedure and analysis

The procedure and analysis were identical to those of Exps. 1 and 2.

4.2 Results and discussion

The overall restoration rate was 31.3%. As in Exps. 1 and 2, the effect of TENSE was significant ($\beta = -0.53$, $p < 0.05$, $\chi^2(1) = 6.06$): That is, participants were again more likely to produce expletive subject constructions in past-tense sentences than in present tense sentences. In contrast, the effect of PERSON was not significant ($\beta = -0.13$, $p = 0.50$, $\chi^2(1) = 0.45$), and neither was the interaction of TENSE and PERSON ($\beta = 0.30$, $p = 0.41$, $\chi^2(1) = 0.67$). After the marginal PERSON effect of Exp. 1, and the non-significant effect in Exp. 2, we take this as supporting evidence that TENSE generates more robust grammatical top-down expectations (Figure 8).

Figure 8: Results from Exp. 3. Participants were significantly more likely to produce expletive subject constructions in past tense sentences than in present tense sentences, but we found no significant effect of person. In addition, restoration rates were significantly higher than in Exp. 2

5 Comparison between experiments

The three experiments individually establish that listeners use top-down, grammatical expectations to inform the noise correction process. On the other hand, comparisons between experiments can reveal to what extent filler design modulates behavior, as the experiments vary in both the number of fillers (60 in Exp. 1 vs. 24 in Exps. 2 and 3) and in the structure of fillers (a highly informative noise distribution in Exp. 3 vs. a relatively uniform noise distribution in Exp. 2). Note that if participants' error correction processes are sensitive to the noise distribution throughout the experiment, then both of these factors should affect restoration rates, either by simply increasing the ratio of critical items with noise over a potential expletive subject to fillers, or by using only filler items that contain noise distributions similar to that of the critical items (as we do with that-fillers). Thus, if we assume such a sensitivity, we would expect significant differences in restoration rates between Exps. 1–3. In addition, given our hypothesis that the presence of a highly informative noise distribution would reduce our reliance on top-down grammatical influences on correction, we predict an interaction between grammatical factors (here, *TENSE*) and our filler structure manipulation (Exps. 2 vs. 3), where the effect of grammatical factors is smaller when the noise distribution is more informative (as in Exp. 3).

This analysis was conducted using a logistic mixed effects model predicting the restoration rate with *EXPERIMENT NUMBER*, *TENSE* and *PERSON* as fixed effects, and with *item* as a random effect with maximal random effects structure (random slope and intercept).

Between Exp. 1 and Exps. 2 and 3, the overall correction rates dropped significantly (Exps. 1 vs. 2: $\beta = -1.57$, $p < 2e-16$, $\chi^2(1) = 178.83$; Exps. 1 vs. 3: $\beta = -1.25$, $p < 2e-16$, $\chi^2(1) = 122.58$). Since the primary difference between Exps. 2 and 1 was the *ratio* of filler to critical items, this suggests that the number of fillers, and thus the ratio between stimuli with systematically distributed noise (in Exp. 2, only the critical items) and those with no such systematicity (the fillers), has a strong influence on the error correction process. This is mirrored in Exps. 1 and 3, where in Exp. 3 all experimental items share relevant statistical properties (noise over optional, sentence initial words).

As shown in Figure 9, we also found a significant main effect of filler *structure*, comparing Exps. 2 and 3 ($\beta = 0.33$, $p < 10^{-4}$, $\chi^2(1) = 16.55$). Contrary to our predictions, we did not find an interaction with *TENSE* ($\beta = 0.072$, $p = 0.68$, $\chi^2(1) = 0.17$) or *PERSON* ($\beta = 0.015$, $p = 0.90$, $\chi^2(1) = 0.02$).

Figure 9: The effect of the TENSE manipulation across Exps. 2 and 3. Though we find a significantly higher rate of restoration in Exp. 3 than in Exp. 2, we did not find our predicted interaction between the experiment and TENSE.

6 Discussion

In this paper, we asked how contextual information can influence error correction, and how grammatical expectations interact with experimental context. Specifically, we asked how listeners arrive at the Gestalt of a sentence when they correct a noisy auditory signal, and whether different grammatical preferences are more robust to filler ratio and structure.

First, we successfully replicated Mack et al.'s (2012) results, demonstrating again that listeners restore a missing subject based on grammatical constraints: the ungrammaticality of subject elision is ameliorated in present tense, and less so, with first person. While we found a lower overall restoration rate than Mack et al. (2012), both the direction and pattern of results replicated those of the original study. One possible explanation of the difference in the overall restoration rate could be recent language change: given the notable time difference between our data collection and that of Mack et al. (2012), as well as the geographical differences in the collection sites, the population we sampled may have had a grammar that found null subject constructions more acceptable across the board. While investigating this potential change is interesting, it is outside of this paper's purview.

In Exps. 2 and 3, we chose subsets of the filler items from Exp. 1 to increase the ratio of systematically distributed noise to random noise in a step-wise fashion. This allowed us to investigate the role of filler *structure* and *ratio* in an experimental setting. We found a substantial effect of this between-experiment manipulation, confirming that listeners adapt to the distributional patterns of noise: fewer fillers (Exp. 1 vs. Exps. 2 and 3) led to a drop in restoration rates, and a higher rate of fillers with noise distributions similar to critical items (Exps. 2 vs. 3) resulted in higher restoration rates.

This suggests independent, competing contributions of filler *ratio* (where a larger filler to critical ratio leads to lower restoration rates) and filler *structure*, here operationalized as noise distribution within those fillers (where an informative distribution of noise leads to higher restoration rates).

However, we failed to find the predicted interaction between noise distribution and the grammaticality manipulations. While it is possible that such an interaction simply does not exist, we suspect that its absence may be due to a floor effect: While tense and person factor into grammaticality preferences differently (as in Exp. 1), the marginal effect of person disappears when the ratio of filler to critical items is reduced, while the overall restoration rates dropped significantly (from Exp. 1 to Exps. 2 and 3).

This would indicate that participants unconsciously reacted to the critical manipulation based on the *ratio* between critical and filler items, but the *structure of noise* within the filler items at the same ratio did not selectively affect subtle differences in strength between grammatical constraints: we may be observing low filler-to-critical item ratios in Exps. 2 and 3 driving down restoration rates (and as a consequence also the size of the main effects of tense and person), resulting in the lack of an interaction effect. Further experiments with higher filler-to-critical ratios are necessary to determine whether this is the case.

In general, we see these results as a validation of the use of a Speech Restoration paradigm to probe the context dependence of speaker's error correction processes. Just as in Gibson et al. (2013), we found that listeners are sensitive to statistical patterns in the linguistic context of an utterance when choosing whether, and how to, correct. In addition, the Speech Restoration paradigm has several strengths relative to the methodology of Gibson et al. (2013): first, the use of ambiguous noise segments allows for sensitivity to *small* influences on error correction, while the experiments in Gibson et al. (2013) require that such an influence was large enough to bias the participant away from a faithful reading of a grammatical (but implausible) sentence.

Second, the one-shot auditory presentation makes post-comprehension conscious correction less likely, as, unlike in the Gibson et al. (2013) paradigm, participants have only temporary access to the true stimulus. Because of this, any post-comprehension process must work solely with the output of automatic error correction processes, with no opportunity to verify whether that representation is consistent with the recording that was presented. Of course, participants still could have applied conscious corrections to this automatically corrected representation, but we think this is relatively unlikely: first because this representation would have already been corrected once, and second, because participants were explicitly instructed to repeat the stimuli *exactly* as they heard them. We take the fact that we found a lower overall restoration rate than Mack et al. (2012) as additional indication that our participants were unlikely to overcorrect.

Regarding the difference between automatic and conscious corrections, the speech restoration paradigm we used is also distinct from the approach used by Ryskin et al. (2018) in their conceptual replication of Gibson et al. (2013). As in our work here, Ryskin et al. (2018) like our work, directly measure the rate at which participants correct noise in stimuli in an effort to obtain more fine-grained information about the sensitivity of the error correction process to contextual information. However, the methodology they adopt serves a fundamentally different purpose than the speech-restoration paradigm we have chosen: as noted above, we see the fact that the speech-restoration paradigm almost entirely prevents conscious error correction as a benefit, as our object of study is specifically unconscious error correction processes. Ryskin et al. (2018), on the other hand, aim to investigate the noise model directly, which they hypothesize is shared between conscious and unconscious error correction processes. We thus see these two paradigms as complementary approaches, with their merits differing depending on the particular research question and set of assumptions one chooses.

Third, Mack et al.'s (2012) paradigm provides the ability to measure the relative amenability of other soft grammatical violations, like subject elision, within the error correction process. We have shown that when the ratio between structured and unstructured noise decreases (from Exp. 1 to Exps. 2 and 3), people rely more on bottom-up information, and the effect of weak grammatical preferences (the effect of person) disappears. Thus, this paradigm can function to evaluate the relative strength of grammatical preferences.

We see this as an exciting correlate to vision research, which has shown how sparse, noisy information suffices for humans to correctly infer underlying visual input based on top-down expectations, but also, that statistical information within the noise is crucially important for performance (Gosselin and Schyns 2001). This sensitivity seems to be an important factor *whenever* people are (re)constructing the Gestalt of a percept – regardless of the cognitive domain involved.

Finally, these results provide additional, systematic evidence that participants are sensitive to subtle details in fillers, influencing their behavior on critical items. Modifying either the ratio of filler to critical items (Exp. 1 vs. Exps. 2 and 3) or the statistical properties of the filler items (Exp. 1 vs. 2) results in significantly different patterns of results despite the critical items being identical. We therefore recommend following the most conservative suggested ratios in experimental design in order to avoid accidentally biasing participants (Cowart 1997) – both in favor and against one's own prediction.

Research funding: This research was supported by University of California, San Diego (Chancellor's Research Excellence Scholarship).

References

- Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.
- Baumann, Stefan & Petra B. Schumacher. 2020. The incremental processing of focus, givenness and prosodic prominence. *Glossa: A Journal of General Linguistics* 5(1). 6.
- Bodner, Glen E., Michael E. J. Masson & Norann T. Richard. 2006. Repetition proportion biases masked priming of lexical decisions. *Memory & Cognition* 34(6). 1298–1311.
- Cowart, Wayne. 1997. *Experimental syntax*. Thousand Oaks, CA: Sage.
- Dąbrowska, Ewa. 2010 Naive v. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review* 27(1). 1–23.
- Feiman, Roman, Mora Maldonado & Jesse Snedeker. 2020. Priming quantifier scope: Reexamining the evidence against scope inversion. *Glossa: A Journal of General Linguistics* 5(1). 35.
- Ferreira, Victor S. 1997. *Syntactic and lexical choices in language production: What we can learn from "That"*. Champaign, IL: University of Illinois at Urbana-Champaign Doctoral Dissertation.
- Flom, Lynda & David L. Cassell. 2007. Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. In *Proceedings of the NorthEast SAS Users Group Inc 20th annual conference, Baltimore, 11–14 November*.
- Franck, Julie, Farhad Mirdamadi & Arsalan Kahnemuyipour. 2020. Object attraction and the role of structural hierarchy: Evidence from Persian. *Glossa: A Journal of General Linguistics* 5(1). 27.
- Garnham, Alan, Jane Oakhill & Hannah Cruttenden. 1992. The role of implicit causality and gender cue in the interpretation of pronouns. *Language & Cognitive Processes* 7(3–4). 231–255.
- Gibson, Edward, Leon Bergen & Steven T. Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110(20). 8051–8056.
- Gosselin, Frédéric & Philippe G. Schyns. 2001. Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research* 41(17). 2261–2271.
- Grant, Margaret, Shayne Sloggett & Brian Dillon. 2020. Processing ambiguities in attachment and pronominal reference. *Glossa: A Journal of General Linguistics* 5(1). 77.
- Green, Jeffrey, Michael McCourt, Ellen Lau & Alexander Williams. 2020. Processing adjunct control: Evidence on the use of structural information and prediction in reference resolution. *Glossa: A Journal of General Linguistics* 5(1). 112.
- Green, Peter & Catriona J. MacLeod. 2016. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* 7(4). 493–498.
- Gross, Steven & Jennifer Culbertson. 2011. Revisited linguistic intuitions. *The British Journal for the Philosophy of Science* 62(3). 639–656.
- Harrington Stack, Caoimhe M., Ariel N. James & Duane G. Watson. 2018. A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition* 46(6). 864–877.
- Havik, Else, Leah Roberts, Roeland Van Hout, Robert Schreuder & Marco Haverkort. 2009. Processing subject-object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning* 59(1). 73–112.
- Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Edinburgh: University of Edinburgh Doctoral Dissertation.
- Levy, Roger. 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, Honolulu, 25–27 October*.
- Mack, Jennifer E., Charles Clifton, Lyn Frazier & Patrick V. Taylor. 2012. (Not) hearing optional subjects: The effects of pragmatic usage preferences. *Journal of Memory and Language* 67. 211–223.
- Mahowald, Kyle, Peter Graff, Jeremy Hartman & Edward Gibson. 2016. SNAP judgments: A small N acceptability paradigm (SNAP) for linguistic acceptability judgments. *Language* 92(3). 619–635.
- von der Malsburg, Titus, Till Poppels & Roger P. Levy. 2020. Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 United States and 2017 United Kingdom elections. *Psychological Science* 31(2). 115–128.
- Marinis, Theo & Ian Cummings. 2018. *Using psycholinguistic techniques in a second language teaching setting. Mind matters in SLA*. Bristol: Multilingual Matters.
- Marty, Paul, Emmanuel Chemla & Jon Sprouse. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: A Journal of General Linguistics* 5(1). 72.
- Morgan, Adam M., Titus von der Malsburg, Victor S. Ferreira & Eva Wittenberg. 2020. Shared syntax between comprehension and production: Multi-paradigm evidence that resumptive pronouns hinder comprehension. *Cognition* 205. 104417.
- Pañeda, Claudia, Sol Lago, Elena Vares, João Verissimo & Claudia Felser. 2020. Island effects in Spanish comprehension. *Glossa: A Journal of General Linguistics* 5(1). 21.

- Poppels, Till & Roger Levy. 2016. Structure-sensitive noise inference: Comprehenders expect exchange errors. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, 10–13 August*.
- Ryskin, Rachel, Richard Futrell, Swathi Kiran & Edward Gibson. 2018. Comprehenders model the nature of noise in the environment. *Cognition* 181. 141–150.
- Schotter, Elizabeth R., Klinton Bicknell, Ian Howard, Roger Levy & Keith Rayner. 2014. Task effects reveal cognitive flexibility responding to frequency and predictability: Evidence from eye movements in reading and proofreading. *Cognition* 131. 1–27.
- Schütze, Carson. 2016. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Berlin: Language Science Press.
- Schütze, Carson & Jon Sprouse. 2014. Judgment data. In Robert J. Podesva & Devyani Sharmi (eds.), *Research methods in linguistics*, 27–50. Cambridge: Cambridge University Press.
- Sorace, Antonella & Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11). 1497–1524.
- Warren, Richard M. 1970. Restoration of missing speech sounds. *Science* 167. 392–393.
- Wasow, Thomas & Jennifer Arnold. 2005. Intuitions in linguistic argumentation. *Lingua* 115(11). 1481–1496.
- Wittenberg, Eva & Ray Jackendoff. 2018. Formalist modeling and psychological reality. *Linguistic Approaches to Bilingualism* 8(6). 787–791.
- Yoon, Hang-Jin. 2001 Expletive it in English. *Studies in Generative Grammar* 11. 543–562.