# Repairing misperceptions of words early in a sentence is more effortful than repairing later words, especially for listeners with cochlear implants

Michael L. Smith[1*] & Matthew B. Winn[1]

Department of Speech-Language-Hearing Sciences
University of Minnesota, Minneapolis, MN, USA
*Corresponding authors contact information:
smit8854@umn.edu
Shevlin Hall
164 Pillsbury Dr SE
Minneapolis, MN 55455

## Abstract

The process of repairing misperceptions has been identified as a contributor to effortful listening in people who use cochlear implants (CIs). The current study was designed to examine the relative cost of repairing misperceptions at earlier or later parts of a sentence that contained contextual information that could be used to infer words both predictively or retroactively. Misperceptions were enforced at specific times by replacing single words with noise. Changes in pupil dilation were analyzed to track differences in the timing and duration of effort, comparing listeners with typical hearing or with CIs. Increases in pupil dilation were time-locked to the moment of the missing word, with longer-lasting increases when the missing word was earlier in the sentence. CI listeners showed elevated pupil dilation for longer periods of time after listening, suggesting lingering effects of effort compared to listeners with TH. When needing to mentally repair missing words, CI listeners also made more mistakes on words elsewhere in the sentence, even though these words were not masked. Stimulus-related effects were not evident in basic measures like peak pupil dilation, and only emerged when the full-time course was analyzed, suggesting the timing analysis adds new information to our understanding of listening effort. Taken together, these results demonstrate that some mistakes are more costly than others and incur different levels of mental effort to resolve the mistake, underscoring the information lost when characterizing speech perception with simple measures like percent-correct scores.

Keywords: cochlear implants, listening effort, pupillometry, speech perception, perceptual restoration

# Introduction

There is a growing appreciation for listening effort in clinical hearing science (Pichora-Fuller et al., 2016; Zekveld et al., 2018), complemented by studies aimed at understanding the mechanisms of what aspects of speech communication are effortful for listeners who are hard-of-hearing. Repetition accuracy is the most common outcome measure of speech perception abilities, but a percent correct score fails to capture how the listener arrived at their answer – particularly the cost of recovering from a perceptual mistake in the process of inferring the correct answer.

In virtually any study of speech recognition, misperceptions are uncontrolled and emerge unpredictably at any moment during listening. Two listeners who both show 75% correct repetition accuracy could be making different mistakes, and it would not be possible to explain their different listening experiences without understanding the difference between those mistake patterns. There are various types of misperceptions a listener can make during perception (phonetic mistakes, segmentation errors, syntactic errors, semantic substitutions, etc.), and these different types of misperceptions incur different amounts of listening effort (Winn & Teece, 2021). However, when using commonly used testing materials, it can be difficult to prospectively control when misperceptions occur during an experiment, relegating comparison of these mistakes to retrospective analyses. Winn and Teece (2021) used such a retrospective design to reveal that mistakes on earlier words in a sentence are more costly than making mistakes on later words, and the observed effort was linked specifically to semantic processing, rather than the degree of acoustic-phonetic matching of the stimulus and response. These observations underscore the notion that listening effort cannot be captured by a mere tally of repeated correct and incorrect words or degree of phonetic match to the original stimulus. Incidentally, that study also introduced a testing method to prospectively induce misperceptions at a specific time

during a sentence to estimate the effort of mentally repairing misperceived words. However, that design only included mistakes early in a sentence, toward the goal of specifically examining retroactive use of context. The present study extends that study design by controlling the time of misperceptions and the contextual information listeners have available to resolve the ensuing ambiguity, toward the goal of discerning the impact of misperceptions earlier or later in a sentence.

## Pupillometry as a Measure of Listening Effort

While there are multiple methodological tools that can be used to quantify listening effort, such as changes in reaction time or subjective report, the current research question demands the ability to measure moment-by-moment changes in effort as language processing unfolds in real time, because the core question is about earlier and later moments of processing within the same sentence. Pupillometry is a tool that can measure changes in pupil dilation before, during, and after language processing (Engelhardt et al., 2010; Winn, 2023), and can reflect degrees of ambiguity and post-sensory processing of the input (Satterthwaite et al., 2007). Changes in pupil dilation have a long history of being correlated with changes in cognitive demand across a variety of tasks (Kahneman & Beatty, 1966; Sirois & Brisson, 2014), with pupil dilation generally increasing when more effort is exerted (van der Wel & van Steenbergen, 2018), so long as there is sufficient motivation to complete the task. As opposed to the slowly-varying changes in *tonic* pupil size that are thought to reflect alertness (McGinley et al., 2015), short *phasic* changes in pupil size are the key physiological signature of momentary listening effort used in previous studies (Beatty, 1982; Gabay et al., 2011; Zekveld et al., 2018) and are what will be examined in the present study.

The key advantage of analyzing phasic pupil dilations is the ability to quantify precisely *when* effort occurs and *how long* effort lasts, rather than just *how much*. Although it is customary

to report summarized response of peak pupil dilation and peak latency to quantify the amount of effort in a given task (Ayasse et al., 2021; Wendt et al., 2018; Zekveld et al., 2010), there is additional information that can be gained by observing the full-time course of changes in pupil dilation by designing experiments with this specific goal in mind (Johns et al., 2024; Steinhauer et al., 2022; Winn, 2023) . Sustained increases in pupil dilation following the peak could reflect the listener having to reconcile remaining linguistic ambiguity after the initiation of that effort. Quantifying the precise timing of when effort occurs during listening, and the duration for how long this increase in effort lasts, can offer new insight into the relative cost of mistakes at different times during a sentence, as well as the ways that contextual cues and hearing status might interact with that cost.

## Sentence Context

Sentence context will play an important role in the current study for two reasons. First, people who are deaf or hard-of-hearing have been shown to rely more heavily on contextual cues, which can be shown in both accuracy scores (Hunter, 2021; O'Neill et al., 2021; Patro & Mendel, 2016; Pichora-Fuller et al., 1995; Vickery et al., 2022) and listening effort (Hunter & Humes, 2022; Winn, 2016). This is especially true for CI listeners (Başkent et al., 2016; Dingemanse & Goedegebure, 2022; Winn, 2016), likely because they hear an auditory signal that is highly degraded, requiring more compensation from non-auditory cognitive processes. The second reason is that the ability to repair misperceptions likely depends on the availability of remaining contextual cues within the utterance (or across utterances) to resolve linguistic ambiguity.

In ideal situations, listeners can use context to rapidly predict upcoming words as the speech signal unfolds in real time (Federmeier, 2007). However, some evidence suggests that CI listeners are less likely to use context quickly (Winn, 2016; Winn & Moore, 2018), perhaps as

a result of refraining from full commitment to lexical decisions until they can be more certain in what they heard (Farris-Trimble et al., 2014; McMurray et al., 2017). CI listeners may rely more heavily on using context retroactively, which has been shown to be an effortful process (Winn & Teece, 2022). It is tempting to speculate on the comparisons between CI listeners using context in a way that is typically framed as predictive (e.g. Winn 2016; (Hunter & Humes, 2022) versus using context framed as retroactive (Winn & Teece 2022). However, studies focusing on these uses of context have used different methods in terms of stimulus design and outcome measures that prevent fair comparison. For example, while the study by Winn and Teece (2022) verified that listeners used context to mentally repair misperceptions early in a sentence, previous work on predictive context mainly inferred the use of context through accuracy scores for sentence-final words, without being able to confirm if any earlier misperception took place. Therefore, it remains unclear whether the use of context to repair words at different time points during a sentence elicits a different amount or duration of effort.

## The Present Study

Combining the previous described impact of sentence context on listening effort, the present study aims to evaluate how listeners use context in different parts of the sentence to resolve linguistic ambiguity and the effect this has on listening effort. By designing stimuli where the same sentence could have a word missing either earlier or later in the same sentence, we can directly compare how different types of sentence context impacts the timing of changes in pupil dilation, the impact on the duration of the pupil response, and the potential differences in effort between CI and older and younger TH listeners. The approach of distorting or removing a portion of the signal was introduced by Warren (1970) as perceptual restoration, and later extended by Winn and Teece (2021) to address situations where the listener does not feel a sense of actually having heard the word, but instead needs to actively infer it based on later information.

The present study has three main hypotheses: 1) the timing of pupil dilation resulting from missing words will be related to the position of those words within the sentence, as opposed to being a general increase in dilation; 2) Sentences with earlier-masked words will have a larger increase and duration of pupil size compared to sentences with late-masked words, because for words early in a sentence, listeners cannot take advantage of preceding contextual information to help resolve any linguistic ambiguity; they must hold that ambiguous word in memory while they wait to accumulate more information; 3) Consistent with previous studies, CI listeners will show prolonged pupil dilation because they could be less likely to take advantage of sentence context as it unfolds in real time; 4) Responses that demand repair but which are not repaired successfully will produce prolonged pupil responses as a result of the listener's persistent effort to resolve ambiguity in the sentence.

# Methods

## Participants

All participants in this study were native speakers of North American English and reported no history of language or learning disabilities. Two groups of listeners were recruited. For the CI group a total of 20 listeners participated in this study (14 female, 6 male) with an average age of 65.3 years old (sd = 11.3 years old, range = 34-77 years old). All CI listeners were able to converse freely during face-to-face communication, and none reported cognitive difficulties. To account for the effect of age, 20 older and 21 younger TH listeners were also recruited (32 female, 9 male, mean = 47.4 years old, sd = 23.7 years, range = 20-84 years old). Listener ages are shown in Figure 1. Normal hearing status was confirmed by pure-tone audiometry screening via air-conduction at 25 dB HL from 250-4000 Hz. Participants were not evaluated for visual acuity. All gave informed written consent of procedures that were approved

by the Institutional Review Board at the University of Minnesota, which stands on Miní Sóta Makhóčhe the homelands of the Dakhóta Oyáte.

CI listeners all had at least 1 year of experience with their device (median experience 7 years). There was a median of 30 years duration of deafness until first implantation among the CI group, which included 5 unilaterally and 11 bilaterally implanted individuals, along with 4 bimodal listeners. Those listeners who regularly used a hearing aid in the contralateral ear to manage moderate to profound hearing loss continued using the hearing aid during the experiment to best simulate their everyday listening experience.

*[[ Figure 1: Age Distributions ]]*

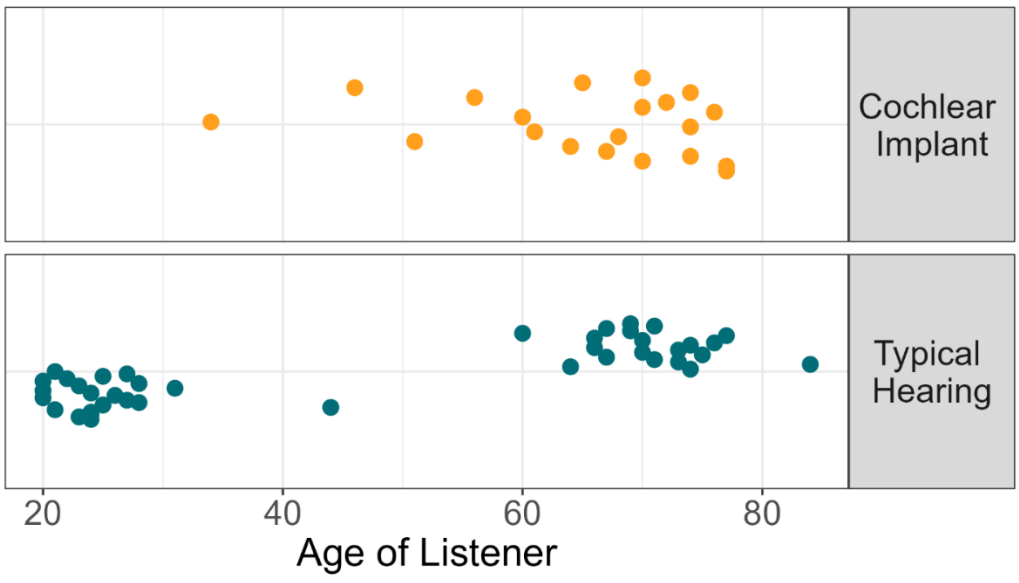


Figure 1: Distributions of listener ages for the two different hearing groups (color online).

## Stimulus variations

Stimuli included 108 sentences written and recorded by our lab. Each sentence was designed so that there were at least three semantically related key words, such that when the earliest or latest of these words were masked by noise, it could be inferred from the remaining related words that were intact. Importantly, this enabled the same sentence to be used as a stimulus in either the early- or late-missing word variations, which is crucial in order to address the research question. The “Fully Intact” version was the fully spoken sentence with no alterations. The other two versions were designed to force the listener to engage in the mental repair process to disambiguate the missing word. In the “Early Repair” condition an early target word was replaced with noise, and in the “Late Repair” condition the final word was replaced with noise. The stimulus types are illustrated in Figure 2. The noise used to replace the words was matched in duration and intensity to the target word, and the frequency spectrum matched the long-term spectrum of the entire stimulus corpus.

The contextual constraint on the words was verified using an online cloze probability test (Kutas & Hillyard, 1984), where a separate group of 30 online participants were shown text versions of the sentences with missing words and had to type what they thought the missing word was. These responses were then analyzed to determine if a particular sentence had either high or low cloze probability, with high probability considered to be situations in which at least 67% agreement in responses to any individual item (Block & Baldwin, 2010). Both missing-word variations of a given sentence had to have a high cloze probability in order to be included in the final stimulus list. Using this criterion, 108 of the original set of 119 candidate sentences had high cloze probability and were included as stimuli for the experiment. The sentences were divided into three lists of 28 and one list of 24 sentences, with sentences having an average of 9 words and an average duration of 2.98 seconds. It was important the sentences be highly

intelligible to minimize mistakes on other words in the sentence. Sentences were recorded by a person who was sex assigned female at birth from Wisconsin, with an emphasis on a clear speaking style and effort to minimize regional dialects of particular vowels (e.g. /æ/-/eɪ/ variation in “bag”).

*[[ Figure 2: Stimuli schematic here ]]*

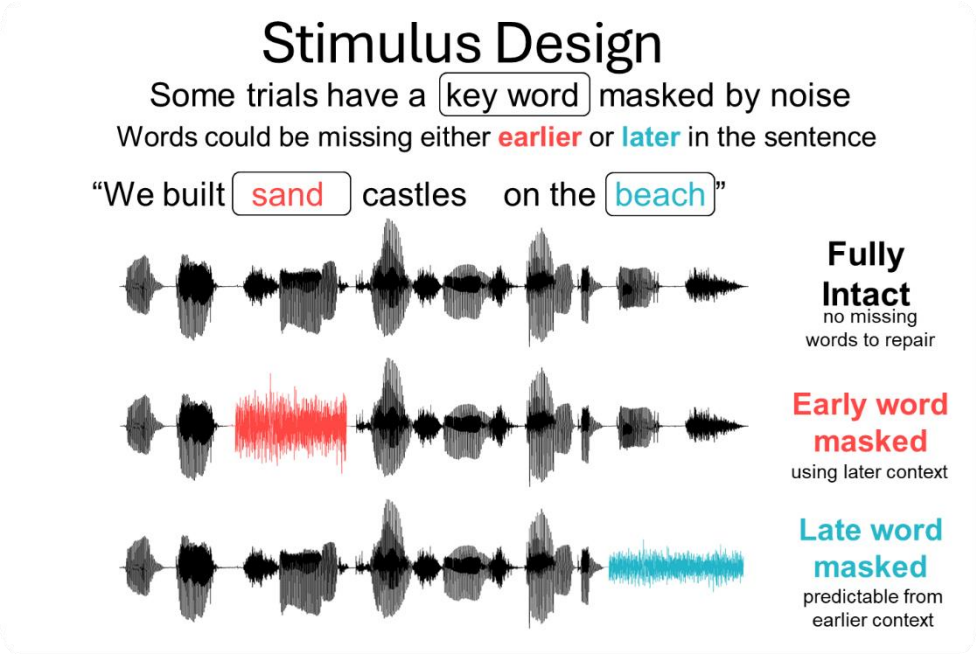


Figure 2: Three different stimulus types, all variations of an example sentence “We built sandcastles on the beach”. Replacing either “sand” or “beach” with speech-shaped noise would create the early-masked or late-masked conditions, respectively. (color online).

## Procedure

Each participant completed a sentence-repetition task with a total of 108 stimuli, which resulted in 36 trials for sentences that were in the intact, backward, and forward condition, respectively. Each stimulus list began with an intact sentence, followed by a pseudo-random ordering of sentence type, with no more than three consecutive trials of the same sentence type. Presentation lists were rotated and counterbalanced across listeners, and the sentence type (fully intact, early word masked, late word masked) for each item was rotated for each listener, except the first trial in each list, which was always intact.

During the experiment, listeners sat in a chair with their forehead position stabilized by the upper bar of a chinrest whose base was sufficiently lowered to allow comfortable jaw movement for speaking. They visually fixated on a red cross in the middle of a medium-dark gray background on a computer screen that was 50 cm away. Each trial was initiated by the experimenter, and the participant heard a beep marking the onset of the trial. There was 2 seconds of silence and then the sentence was played at 65 dBA through a single loudspeaker in front of the listener. Two seconds after the offset of the sentence, the red cross turned green, which was the cue for the listener to give their verbal response. They were instructed to repeat what they thought was spoken, filling in missing words when necessary. The participants' verbal responses were scored on paper, with incorrect responses documented for further analysis of error patterns. The participant's eye position and pupil size were recorded by an SR Research Eyelink 1000 Plus eye tracker recording at 1000 Hz sampling rate, tracking pupil diameter in the remote-tracking mode, using the desktop-mounted 25mm camera lens. Lighting in the testing room was kept constant. A schematic of an example trial is shown in Figure 3.

*[[ Figure 3: Task schematic here ]]*

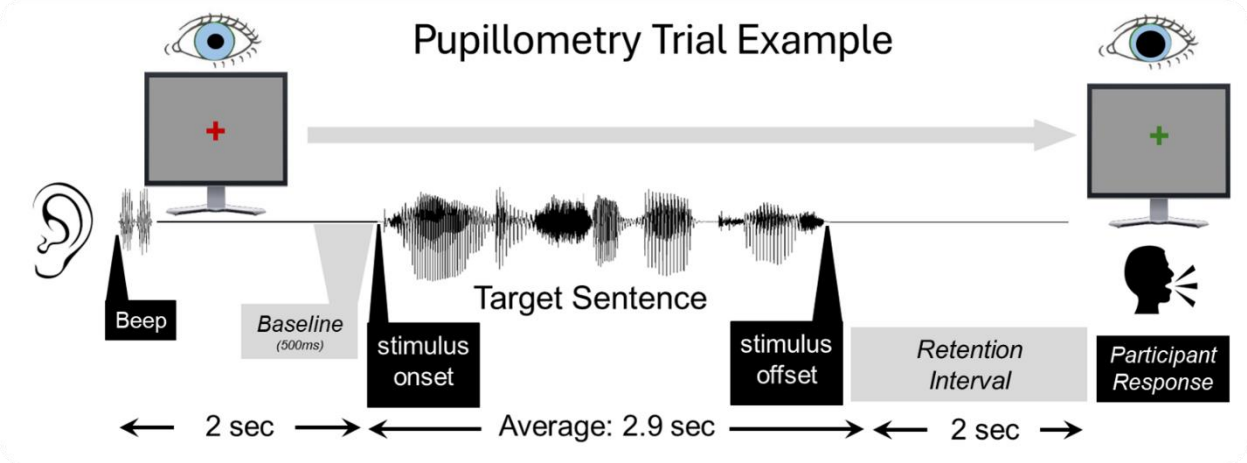


Figure 3: Schematic of overall task design. Listeners were instructed to fixate on the red crosshair on the screen. A beep was played to signal that the sentence would start 2 seconds later. One of the three sentence types was presented at random. After the sentence was over, they waited another 2 seconds before the crosshair turned green indicating the listener should repeat the sentence they heard (including the missing word). Changes in pupil diameter were measured throughout the trial as an index of listening effort.

## Analysis

### Intelligibility

Repetition accuracy was scored in real time by the experimenter and participant responses were manually entered into a data-tracking spreadsheet after the experiment visit for further analysis. For trials where a target word was replaced with noise, any response that was not semantically coherent with the stimulus was counted as an error, as well as any errors elsewhere in the sentence. If the participant's guess at the word replaced by noise was not the

"intact" version of the word but still made sense (e.g. "Please *clean* the floor with this broom", instead of "please *sweep* the floor with this broom"), it was counted as correct. In the case of intact sentences, the target word was defined as the word that would have been masked by noise in the alternate version of the stimulus, to facilitate fair comparison across stimulus types. We also tracked whether participant responses were linguistically coherent, and the presence of multiple errors within trials. An example of an incoherent response would be "The plant hit the soccer ball with the door" (see Winn & Teece 2021 for further discussion of incoherent responses). The goal of evaluating repetition accuracy in this way was to track whether participant responses had any errors, rather than only focusing on the *number* of errors within the response. This approach was taken specifically because the words in high-context sentences are not independent; multiple errors within a sentence would not be a conclusive sign that multiple words were misperceived. For example, misperception of a word might result from the listener trying to create coherence with an earlier word that was misperceived, and participants tend to produce these secondary errors when trying to resolve linguistic ambiguity. Winn and Teece (2021) and Gianakas et al. (2022) provided evidence of this effect in both forward- and backward direction within the sentence, and suggestion that a secondary error tends to reduce effort because it promotes coherence.

To evaluate the differences in overall intelligibility between listener groups, errors were estimated on a per-trial level using a binomial (i.e., logistic) mixed-effects model that included fixed effects and interactions between stimulus type and condition, as well as random intercepts and correlated random effects of condition per listener. Estimated marginal means were calculated from this model to statistically compare the difference in intelligibility scores across hearing groups and express them in plain terms. The following model formula was used in the prevailing model:

*glmer( AnyError ~ Condition + Hearing + Condition*Hearing + (1 + Condition | Listener) )*

Although tracking any error in the participants' verbal responses already reveals the impact of stimulus type and listener group, there is additional information to be gained from analyzing the different types of errors that were made by CI listeners and how those errors may be related to mental repair. Sentence repetition scores were analyzed in more depth for the CI group using a series of GLMMs that estimated various outcome measures, including, 1) the presence of any error within the response, and 2) errors on words other than the target word. These models were restricted only to CI listeners because TH listeners did not make many errors, resulting in implausibly high or low beta estimates due to model estimates including values at or close to zero. The model for estimating the presence of an error on words other than the target had the same structure as the model for any errors. Other types of errors, such as target word errors and incoherent responses, were also counted, however these errors were not frequent enough to be statistically evaluated. The model formula was declared as follows:

*glmer(AnyError ~ Condition + (1 + Condition | Listener) )*

*glmer(ErrorElsewhere ~ Condition + (1 + Condition | Listener) )*

For the two models described above, when a specific comparison was not available in the original model because both sides of the comparison were deviations from the default (and therefore not directly compared to each other), comparisons were obtained by rotating the same model with the default reassigned, rather than running a post-hoc model limited to the specific comparison of interest.

## Pupillometry Data Preprocessing

Pupil data were processed as described by Winn et al. (2018) and Winn and Teece (2022). Blinks were detected as a decrease in pupil size to 0 pixels, with the stretch of time corresponding to the blink expanding backward by 80ms and forward by 120ms to account for the partial occlusion of the pupil by the eyelids during blinks. The signal was low-pass filtered at 5 Hz using a $4^{th}$-order Butterworth filter and then down-sampled to 25 Hz. The baseline pupil size was calculated as the mean pupil size in the time spanning 500ms before stimulus onset to 500ms after sentence onset. Each pupil size data point in the trial was expressed as the proportional difference from the trial-level baseline.

Trials were discarded if 30% or more data points were missing between the start of the baseline to three seconds past the onset of the stimulus. CI listeners on average had fewer trials discarded due to missing data (14.2%) and less variation among individuals (s.d. of 10.4%) compared to TH listeners (average of 22.1% trials discarded with s.d. of 14.8%). Other outliers and contaminations were automatically detected through an algorithm that accumulated multiple “flags”, such as high-intensity low-frequency fluctuations (hippus) activity during baseline, baselines that had extraordinary deviation from both the previous and the next baseline, significant slope of change in pupil size during the baseline, or a significant negative swing in proportional dilation immediately after the stimulus onset. Three or more flags resulted in a trial being dropped. If a participant had fewer than 12 trials remaining in any condition following outlier detection, that participant’s entire dataset was dropped. One listener was excluded from analysis for this reason, leaving 61 total listeners to be included for analysis.

## Pupillometry Data Analysis: Generalized Additive Mixed-effects Modeling (GAMMs)

Our goal is to quantify differences in the timing and duration of listening effort when listeners have to mentally repair a missing word. To achieve this goal, filtered data that were

summarized for each individual in each stimulus condition were estimated using generalized additive mixed-effects models (GAMMs; van Rij et al., 2019). One of the distinct advantages of using GAMMs is the ability to identify stretches of time where there is a meaningful difference between curves, and this can be done during the entire time-course of the pupil response during listening and linguistic processing. GAMMs model the data using a combination of Gaussian basis functions that are summed in weighted combination to match the shape of non-linear data (e.g the pupil response) allowing for statistical analysis of the entire pupil response function without the need for different analysis windows. The number of basis functions to calculate each smooth function can be specified for each predictor variable and each interaction term, as well as the specified random effects. Another advantage of GAMMs is accounting for the autocorrelation of time-series data (Baayen et al., 2016), or the tendency for the data point at time t to be similar to its preceding data point at time t-1, which is problematic because it increases the probability of Type I errors. GAMMs have previously been used to model pupillometry data in studies of listening effort (Boswijk et al., 2020; Porretta & Tucker, 2019; Winn, 2024). The details of the GAMMs model presented here are below, and we refer to van Rij et al (2019) for a more thorough overview of using GAMMs to analyze pupillometry data.

All the models and statistical analyses were executed in R (R Core Team, 2021) and R Studio (RStudio Team, 2020). GAMMs were implemented using the R package “mgcv” version 1.8-42 (S. Wood, 2023; S. N. Wood, 2017) and the R package “itsadug” version 2.4.1 (van Rij et al., 2022) was used for interpretation, validation, and visualization of the statistical analyses. An initial model was used to calculate the autocorrelation lag value (rho) that would be used in the final model. The final model included hearing group (CI and TH) and stimulus type (early masked word, late masked word, fully intact) as fixed effects with different smooth functions fitted over time for each interaction of hearing status and stimulus type. There were random

effects of time as a smooth factor for each listener for each stimulus type. The final model terms are shown below.

```
bam(pupil ~
        # parametrics
        is_CI + is_early + is_late + is_early_CI + is_late_CI +

        # basic smooth for time
        s(time, k = 20, bs = "cr") +

        # difference curve for hearing group
        s(time, by = is_CI, k = 20, bs = "cr") +

        # interactions of condition x hearing
        s(time, by = is_early, k = 20, bs = "cr") +
        s(time, by = is_late, k = 20, bs = "cr") +
        s(time, by = is_early_CI, k = 20, bs = "cr") +
        s(time, by = is_late_CI, k = 20, bs = "cr") +

        # random time smooth per listener
        s(time, Listener,  bs = 'fs',  m = 1, k = 5) +

        # random time smooth per listener interacting with condition
        s(time, Listener, by = is_early, bs = 'fs',  m = 1, k =5),
        s(time, Listener, by = is_late, bs = 'fs',  m = 1, k = 5),

        # inputs for computational efficiency
        method = "fREML", discrete = TRUE, family = "scat",

        # account for autocorrelation of each timepoint in the data
        AR.start = start_event, rho = 0.986,

        data = df)
```

# Results

## Intelligibility

Intelligibility scores (percentage of sentences that were repeated with all words correct) were high for all sentence types for both listener groups, with performance at 85.5% for CI listeners and 96.5% for listeners with TH. These high intelligibility scores indicate that

performance did not decrease into the range where motivation and effort to complete the task would render the pupil data difficult to interpret (Wendt et al., 2018).

CI listeners made a statistically greater number of errors on sentences that demanded mental repair, as shown by the estimated marginal means and confidence intervals for this analysis in Figure 4. There was no statistical difference in the error rates for sentences that involved early versus late repair for CI listeners. When separated by sentence type, CI listeners had intelligibility scores of 79.2% when an early word was masked, 84% when a late word was masked, and 93.3% when the sentence was fully intact. Listeners with TH showed near ceiling levels of performance, with 95.9%, 95.6%, and 97.9% for each sentence type, respectively, with no statistical difference between performance for the three stimulus types.

*[[ Figure 4: Any Error Marginal Means ]]*

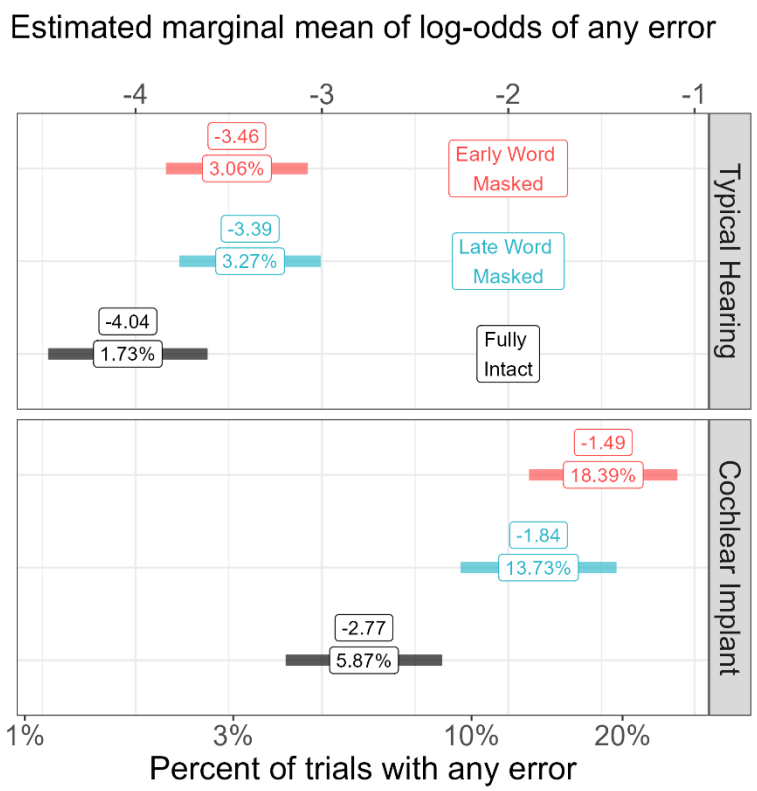

Figure 4: Model estimates of the prevalence of making an error on any word in the sentence, represented as model-inherent log-odds (top x-axis) and converted to percentage (bottom x-axis) for ease of reading. The marginal means estimates for each of the different sentence types are shown with the shaded ribbons indicating the estimated 95% confidence interval (color online).

There was a clear ordering effect of sentence type, with CI listeners making the most errors when an earlier word was missing compared to a late missing word ($\beta = 0.36$, $z = 2.32$, $p = 0.02$), and fewer errors overall when the sentence was fully intact ($\beta = -1.00$, $z = -4.71$, $p < 0.001$). The estimated marginal means shows that there was no overlap in the 95% confidence bands across the listener groups for any stimulus type, suggesting a significant increase in errors for the CI group.

## Different kinds of Errors for CI Listeners

Figure 5 shows the percentage of errors by CI listeners for each specific error types, along with a panel showing data for the previous tally of any error. The number of true target errors (not correctly repairing the masked word, or in the case of intact sentences, making an error on the word that would have been masked) was small enough that no statistics were conducted. A raw count of target-word errors for CI listeners revealed more errors for sentences with early (n = 45) or late (n = 26) missing words compared to those same words when the sentence was fully intact (n = 10).

The third panel of Figure 5 illustrates how mentally repairing a missing word affected perception elsewhere in the sentence. Compared to when the sentences were fully intact, there were more errors on non-target words when CI listeners had to repair earlier ($\beta = 1.02$, $z = 4.27$, $p < 0.001$) or later ($\beta = 0.69$, $z = 2.86$; $p = 0.004$) missing words. Although CI listeners made

more errors elsewhere in the sentence when forced to repair an early missing word (n = 112) versus a late missing word (n = 86), this difference did not reach the conventional criterion for statistical significance ($z$ = 1.715; $p$ = 0.086).

No statistical comparisons were made for incoherent response, although they occurred more often when repairing early (n = 30) or late (n = 20) missing words compared to when the sentence was intact (n = 7).

*[[ Figure 5: CI Intelligibility here ]]*

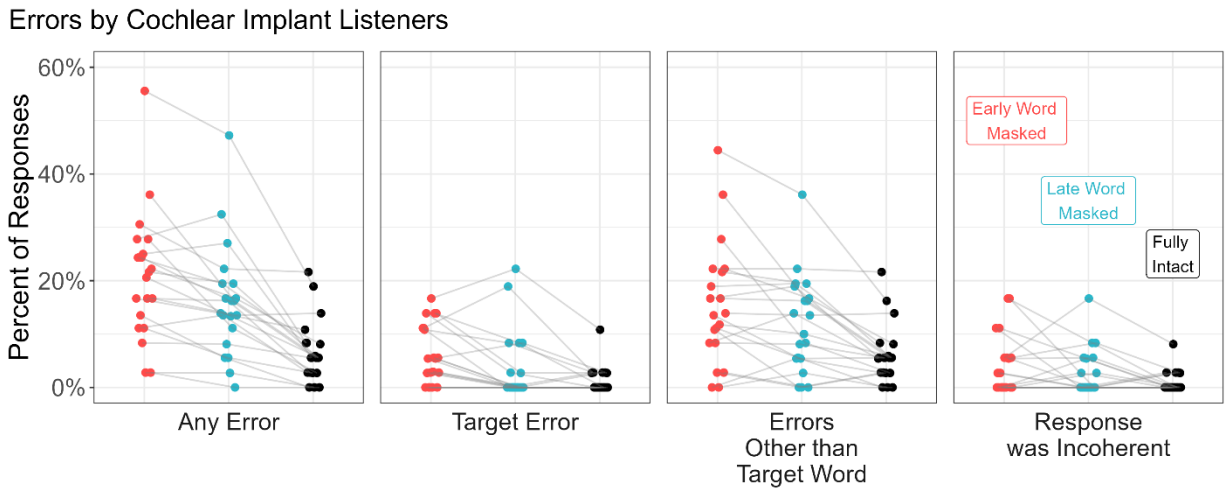


Figure 5: Percentage of responses that contained any errors, for the CI listener group only. Each panel is a different type of error, colored by the different stimulus conditions. Individual listeners are represented by points with connecting lines across the stimulus conditions within a panel.

## Pupillometry

### Main effects of stimulus type (mental repair of missing words)

Changes in pupil dilation for both hearing groups when listening to the different sentence types are shown in Figure 6. The results of the GAMMs analysis are shown as colored bars at the bottom of each panel, which denote regions of statistical differences between the curves. The stimuli with missing early words elicited greater pupil dilation in multiple time windows. First, TH listeners showed larger increases compared to intact sentences between -1.53 to 4 s relative to sentence offset, and a similar time window was observed for CI listeners (between -1.22 to 4 s relative to sentence offset). Second, sentences with early missing words also elicited greater pupil dilation than sentences with later missing words. Both hearing groups showed a similar duration of increased pupil dilation during listening (TH: -1.68 – 0.58 s relative to sentence offset; CI: -1.45 – 0.65 s relative to sentence offset); however, CI listeners showed a longer duration of increased pupil dilation *after* listening (1.33 – 4 s relative to sentence offset) compared to TH listeners where no difference was observed. The stimuli with late missing words also elicited larger increases in pupil dilation compared to intact sentences, but the effect emerged later in time, both for TH listeners (between 0.43 to 4 s relative to sentence offset), and for CI listeners (between 0.58 to 4 s relative to sentence offset). The pupil response to the fully intact sentences was larger than sentences with a late-masked word for a brief window from -0.85 to 0.05 s for TH listeners, and -1.15 to 0.13 s for CI listeners.

*[[ Figure 6: TY Pupil + GAMMs || CI Pupil + GAMMs ]]*

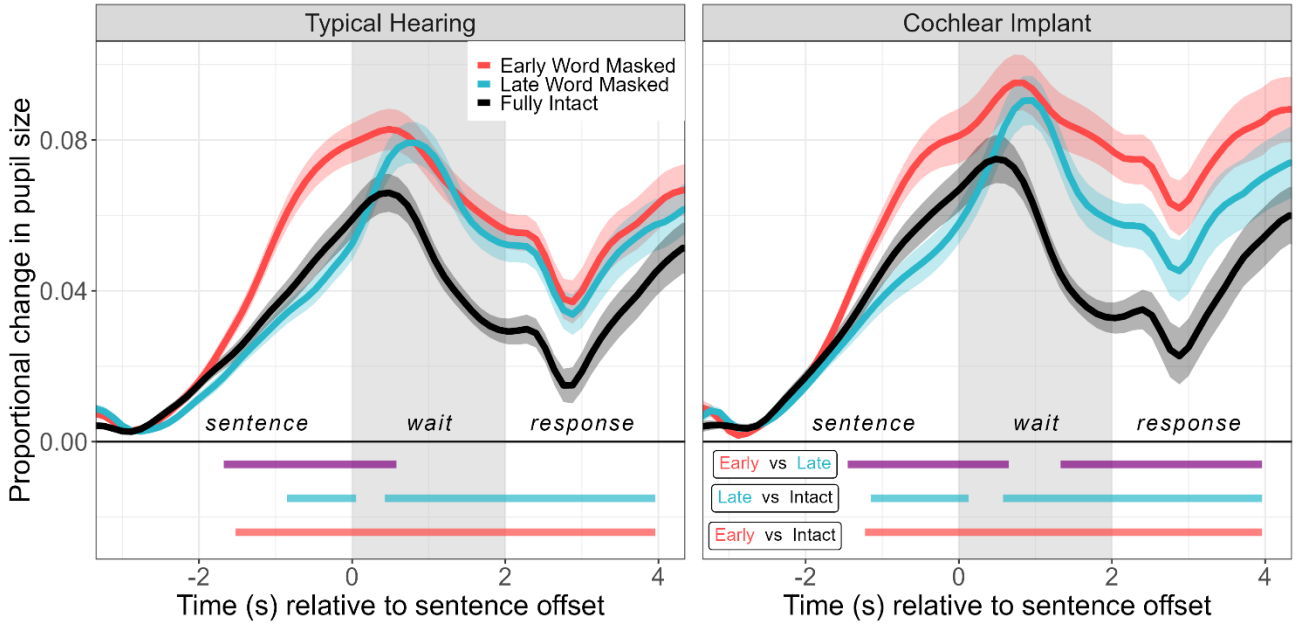


Figure 6: Average proportional change in pupil size for each listener group in response to the three stimulus types. X-axis is time in seconds, with 0 representing sentence offset. Ribbons around the line indicate one standard error. Stretches of time that were found in the GAMM to be statistically different are shown at the bottom of the plot as lines labeled with the two comparison lines. (Color available online).

## Differences between TH and CI groups

Whereas both groups demonstrated increased pupil size in response to stimuli that demanded mental repair, the *degree* of this increase was different across groups. The *difference* of pupil dilation between repair conditions and intact conditions was compared across groups using a GAMM that included interaction terms between stimulus type and hearing group. Figure 7 shows the modeled differences of pupil responses within groups for each stimulus

comparison (left panels; significant stretches already described above), and the comparison of these difference curves between hearing groups (right panels). The increase in pupil dilation resulting from repair of an early masked word in the sentence was greater and longer lasting for the CI listener group, specifically during the stretch of time spanning 0.73 to 4 s relative to sentence offset (top panel Figure 7B). CI listeners also showed a relatively larger effect of late-masked words compared to the TH group during two small time windows from 0.80 to 1.25 s and 2.61 to 3.66 s relative to sentence offset (bottom panel Figure 7B). No differences were observed for responses to early- versus late-masked words between hearing groups (middle panel Figure 7B).

*[[ Figure 7: Diff curves]]*

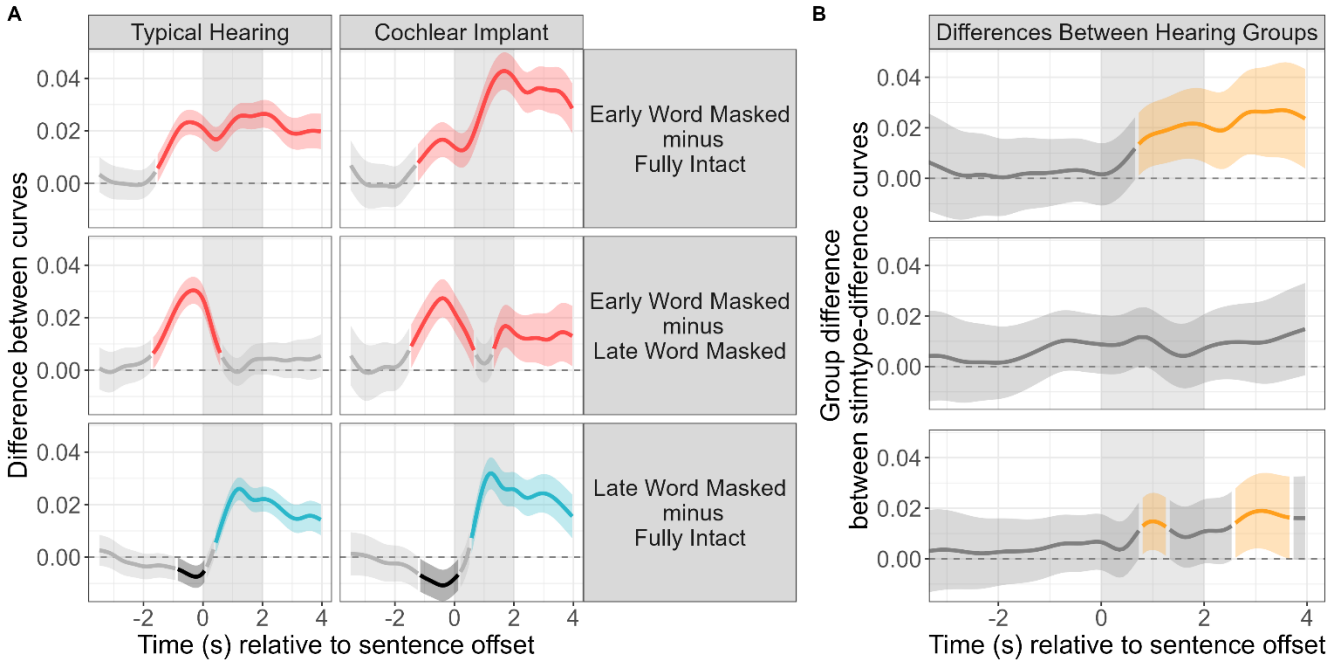


Figure 7: Difference curves from the GAMMs results, illustrating differences between curves from Figure 6, and also illustrating differences between groups. A) The curve represents the modeled difference between stimulus types, with meaningful stretches of time above or below zero shown in color corresponding to which sentence type was greater for the comparison.

Colored regions are the same stretches of time shown as color bars in Figure 6. B) Modeled effect of hearing groups on the difference curves. Each row represents the modeled difference of pupil responses across groups for each stimulus type comparison as shown in column A. Colored regions show stretches of time where the CI listeners had statistically larger differences between curves compared to TH listeners (color online).

## Effect of repetition accuracy on pupil responses

Incorrect responses tend to result in a larger or more sustained increase in pupil dilation (Winn et al., 2015; Zhang et al., 2021), and the difference in performance scores across stimulus types invites analysis of intelligibility effects on the current data from CI listeners (there were not enough incorrect trials for TH listeners to analyze). Pupil responses for the CI listener group for correct and incorrect trials for each sentence type are shown in Figure 8. The elevation in pupil size observed when sentences demand repair is sustained for a longer amount of time when the repair was not fully successful (i.e. when there was still a mistake in the response), compared to when the word was correctly inferred. For sentences with an early-masked words, incorrect responses led to increased pupil dilation from 0.99 to 4 seconds relative to sentence offset, which was a longer duration than the corresponding effect for errors in sentences with later-masked words (1.33 to 4 seconds relative to sentence offset). This result is consistent with generally larger effects of early mistakes in semantically coherent sentences (Gianakas et al., 2022; Winn & Teece, 2021). For sentences that were presented fully intact, the pattern of sustained pupil dilation for incorrect responses was only briefly different than when the response was correct (2.03 to 3.13 seconds relative to sentence offset), although fewer errors were made for those stimuli overall (n = 49) compared to sentences with an early-masked word (n= 150) or a late-masked word (n = 116).

*[[ Figure 8: CI Pupil all vs correct ]]*

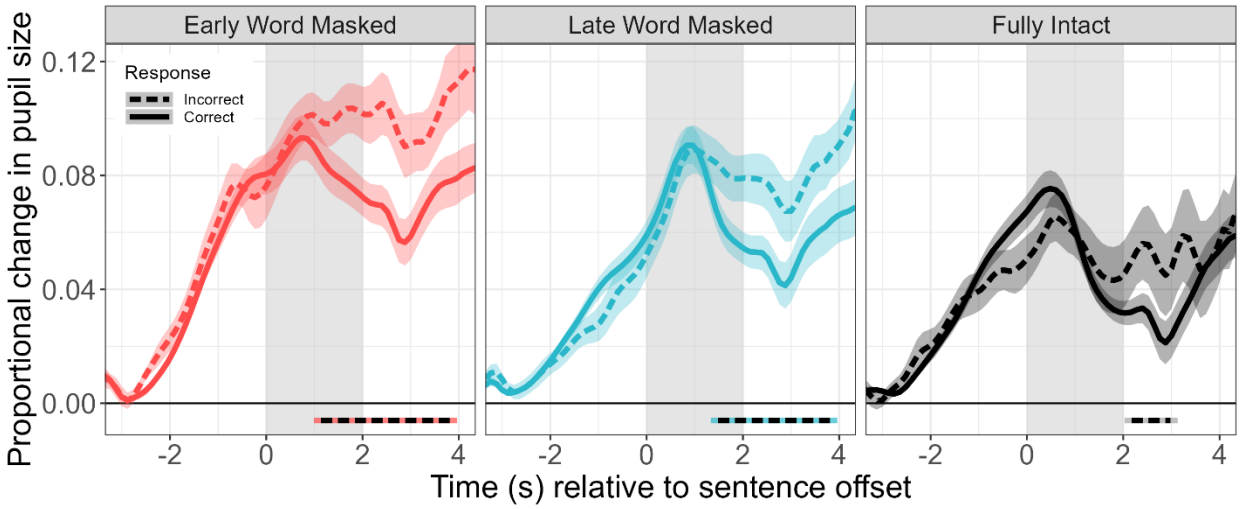


Figure 8: Effect of repetition errors on pupil responses, for CI listeners only. Pupil responses for correct trials are shown in solid lines, with responses from incorrect trials shown in dashed lines. Stretches of time where there was a statistical difference between the two pupil curves are designated by a dashed line below the curves.

## Discussion

The present study aimed to address the question of how the position of a misperceived word impacts listening effort and intelligibility. By prospectively designing sentences with missing words at different word positions within a sentence, we could ensure the mental repair process happened at specific moments, even though it can normally be uncontrollable or undetectable simply based on the participant response. Crucially, the stimuli with early- and late-position words were drawn from the same set of sentences, toward the goal of comparisons that were unavailable in previous studies.

Consistent with previous work, listeners exerted more cognitive resources to disambiguate sentences with missing words, as indicated by increases in pupil dilation that were larger and more rapid compared to when all the words in a sentence were available (Figure 7). The main novel result was that sentences with earlier-masked words had larger increase and duration of extra pupil dilation compared to sentences with late-masked words. This is most likely because the late-masked words have the advantage of preceding disambiguating information before the missing word that could help resolve any linguistic ambiguity in advance, whereas stimuli with early-masked words forced the listener to hold some uncertainty until gathering sufficient contextual clues later on. These results suggest that a mere tally of the number (or percent) of errors in a sentence loses valuable information about the unequal impact of making perceptual mistakes earlier or later in an utterance.

Compared to listeners with TH, CI listeners showed increased duration of increased pupil size when disambiguating missing words, suggesting more time needed to recover from the process of mentally repairing missing words. This result is consistent with CI listeners being less likely to be able to take advantage of sentence context as it unfolds in real time (Winn 2016), which may stem from contextual information being degraded by the device itself. Consistent with previous literature (Winn et al., 2015; Zhang et al., 2021), incorrect responses in the current study produced greater pupil dilation in the moments after the sentence ended. The larger increase in pupil dilation for incorrect responses could be an indication that the listener is still grappling with some unresolved linguistic ambiguity created by the missing word, resulting in lingering effort after the sentence.

## On the importance of measuring the timing (not just the magnitude) of effort

The observation of lingering effort in cases of successful and unsuccessful mental repair of missing words invites concern about the potential implications for perceiving continuous

speech, which typically lacks extended moments of silence that the listener can use to reevaluate and repair previous perceptions. Listeners with severe-profound hearing impairment have suggested there is a significant time lag between hearing and understanding, and feelings of being “behind” due to the extra effort needed to follow the conversation (Hughes et al., 2018). Recent work verifies that misperceiving one word has down-stream consequences for the accurate perception of later sentences if the listener cannot quickly resolve the mistake (Winn, 2024). Testing with two full sentences reveals that some listeners experience severe reduction in performance that would not have been evident by testing one sentence (Svirsky et al., 2024), validating the notion that lingering effort in single-sentence stimuli might overlook difficulties that have implications for real-world interaction.

## A caveat on interpreting the use of context and sentence coherence

Taking advantage of sentence context as a compensatory listening strategy is one potential approach to explore how language processing interacts with listening effort. A common method for evaluating the influence of sentence context on perception is to have high- or low-probability sentences (Bilger et al., 1984), or to have sentences that are either semantically coherent or incoherent (O’Neill et al., 2020; Signoret et al., 2018; Van Engen & Peelle, 2014). In several recent studies involving CI listeners, incoherent responses are shown to elicit larger signatures of effort compared to other types of responses or planned stimulus variations (Winn, 2024; Winn & Teece, 2021, 2022). However, a crucial caveat that must be considered when interpreting these studies is the increase in effort resulting from incoherent perceptions might hinge on the listener’s expectation that the sentences *should be* coherent. This caveat might explain why observed the unexpected result of larger pupil responses for a *clear* speaking style compared to a conversational style. All of the stimuli in that study were contextually incoherent, which might have resulted in the clear speaking style highlighting the unnaturalness of the anomalous sentence content. The influence of the listener’s expectation for sentence coherence

(or expectation of any other reliable pattern) could alter their approach to listening and their allocation of effort in the task. This idea could be explored by a study that directly compares responses from a random mix of stimulus types against results from a blocked design where the listener has clear expectations for stimulus type. The current study can also be contextualized by this idea; perhaps the increased signs of effort for repaired sentences would be diminished if the listener expected to repair every sentence, and increased if the repaired stimuli were less frequent (i.e. more surprising).

## Conclusion

Mentally repairing a misperceived word elicits increased effort, particularly when that word occurred earlier in the sentence, and especially when the repair process was unsuccessful. Elevated listening effort lingers longer after the sentence for CI listeners, especially when needing to repair an earlier missing word. These patterns suggest that not all words should be weighted equally when assessing a listener's perceptual accuracy for words within a sentence. When CI listeners repair missing words, they are also more likely to make mistakes on words elsewhere in the sentence (both earlier and later), even though those words were presented in the clear. These patterns further highlight how participant responses do not reflect the perceptual accuracy itself, but rather the processing of the entire utterance, which builds a foundation at the beginning of the sentence and continues to solidify by the end of the sentence.

# Acknowledgements

Participant recruitment and data collection were assisted by Katherine Teece, Emily Hugo, Tereza Krogseng, Miski Mohamed, and Lexi Olson. Statistical analysis was aided by input from Stefanie Kuchinsky, Nick Pandža, and Michael Johns. The experiment design was assisted by our late colleague Akira Omaki. This research was supported by NIH-NIDCD F32 DC021076 (Smith) and R01 DC017114 (Winn).

# References


Ayasse, N. D., Hodson, A. J., & Wingfield, A. (2021). The Principle of Least Effort and Comprehension of Spoken Sentences by Younger and Older Adults. *Frontiers in Psychology*, *12*, 629464. https://doi.org/10.3389/fpsyg.2021.629464

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. N. (2016). *Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models* (arXiv:1601.02043). arXiv. http://arxiv.org/abs/1601.02043

Başkent, D., Clarke, J., Pals, C., Benard, M. R., Bhargava, P., Saija, J., Sarampalis, A., Wagner, A., & Gaudrain, E. (2016). Cognitive Compensation of Speech Perception With Hearing Impairment, Cochlear Implants, and Aging: How and to What Degree Can It Be Achieved? *Trends in Hearing*, *20*, 233121651667027. https://doi.org/10.1177/2331216516670279

Beatty, J. (1982). *Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources*.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research*, *27*(1), 32–48. https://doi.org/10.1044/jshr.2701.32

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670. https://doi.org/10.3758/BRM.42.3.665

Boswijk, V., Loerts, H., & Hilton, N. H. (2020). Salience is in the eye of the beholder: Increased pupil size reflects acoustically salient variables. *Ampersand*, *7*, 100061. https://doi.org/10.1016/j.amper.2020.100061

Dingemanse, G., & Goedegebure, A. (2022). Listening Effort in Cochlear Implant Users: The Effect of Speech Intelligibility, Noise Reduction Processing, and Working Memory Capacity on the Pupil Dilation Response. *Journal of Speech, Language, and Hearing Research*, *65*(1), 392–404. https://doi.org/10.1044/2021_JSLHR-21-00230

Engelhardt, P. E., Ferreira, F., & Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology (2006)*, *63*(4), 639–645. https://doi.org/10.1080/17470210903469864

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(1), 308–327. https://doi.org/10.1037/a0034353

Federmeier, K. D. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Gabay, S., Pertzov, Y., & Henik, A. (2011). Orienting of attention, pupil size, and the norepinephrine system. *Attention, Perception, & Psychophysics*, *73*(1), 123–129. https://doi.org/10.3758/s13414-010-0015-4

Gianakas, S. P., Fitzgerald, M. B., & Winn, M. B. (2022). Identifying Listeners Whose Speech Intelligibility Depends on a Quiet Extra Moment After a Sentence. *Journal of Speech, Language, and Hearing Research*, *65*(12), 4852–4865. https://doi.org/10.1044/2022_JSLHR-21-00622

Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C. M., & Boisvert, I. (2018). Social Connectedness and Perceived Listening Effort in Adult Cochlear Implant Users: A Grounded Theory to Establish Content Validity for a New Patient-Reported Outcome Measure. *Ear & Hearing*, *39*(5), 922–934. https://doi.org/10.1097/AUD.0000000000000553

Hunter, C. R. (2021). Dual-Task Accuracy and Response Time Index Effects of Spoken Sentence Predictability and Cognitive Load on Listening Effort. *Trends in Hearing*, *25*, 233121652110180. https://doi.org/10.1177/23312165211018092

Hunter, C. R., & Humes, L. E. (2022). Predictive Sentence Context Reduces Listening Effort in Older Adults With and Without Hearing Loss and With High and Low Working Memory Capacity. *Ear and Hearing*, *43*(4), 1164. https://doi.org/10.1097/AUD.0000000000001192

Johns, M. A., Calloway, R. C., Karunathilake, I. M. D., Decruy, L. P., Anderson, S., Simon, J. Z., & Kuchinsky, S. E. (2024). Attention Mobilization as a Modulator of Listening Effort: Evidence From Pupillometry. *Trends in Hearing*, *28*, 23312165241245240. https://doi.org/10.1177/23312165241245240

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science (New York, N.Y.)*, *154*(3756), 1583–1585. https://doi.org/10.1126/science.154.3756.1583

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*(5947), 161–163. https://doi.org/10.1038/307161a0

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., Tolias, A. S., Cardin, J. A., & McCormick, D. A. (2015). Waking State: Rapid Variations Modulate Neural and Behavioral Responses. *Neuron*, *87*(6), 1143–1161. https://doi.org/10.1016/j.neuron.2015.09.012

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. *Cognition*, *169*, 147–164. https://doi.org/10.1016/j.cognition.2017.08.013

O'Neill, E. R., Parke, M. N., Kreft, H. A., & Oxenham, A. J. (2020). Development and Validation of Sentences Without Semantic Context to Complement the Basic English Lexicon Sentences. *Journal of Speech, Language, and Hearing Research: JSLHR*, *63*(11), 3847–3854. https://doi.org/10.1044/2020_JSLHR-20-00174

O'Neill, E. R., Parke, M. N., Kreft, H. A., & Oxenham, A. J. (2021). Role of semantic context and talker variability in speech perception of cochlear-implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *149*(2), 1224–1239. https://doi.org/10.1121/10.0003532

Patro, C., & Mendel, L. L. (2016). Role of contextual cues on the perception of spectrally reduced interrupted speech. *The Journal of the Acoustical Society of America*, *140*(2), 1336. https://doi.org/10.1121/1.4961450

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37 Suppl 1*, 5S-27S. https://doi.org/10.1097/AUD.0000000000000312

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608. https://doi.org/10.1121/1.412282

Porretta, V., & Tucker, B. V. (2019). Eyes Wide Open: Pupillary Response to a Foreign Accent Varying in Intelligibility. *Frontiers in Communication*, *4*. https://www.frontiersin.org/articles/10.3389/fcomm.2019.00008

R Core Team. (2021). *R: The R Project for Statistical Computing*. https://www.r-project.org/

RStudio Team. (2020). *RStudio: Integrated Development Environment for R*.

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., & Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *NeuroImage*, *37*(3), 1017–1031. https://doi.org/10.1016/j.neuroimage.2007.04.066

Signoret, C., Johnsrude, I., Classon, E., & Rudner, M. (2018). Combined effects of form- and meaning-based predictability on perceived clarity of speech. *Journal of Experimental Psychology. Human Perception and Performance*, *44*(2), 277–285. https://doi.org/10.1037/xhp0000442

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews. Cognitive Science*, *5*(6), 679–692. https://doi.org/10.1002/wcs.1323

Steinhauer, S. R., Bradley, M. M., Siegle, G. J., Roecklein, K. A., & Dix, A. (2022). Publication guidelines and recommendations for pupillary measurement in psychophysiological studies. *Psychophysiology*, *59*(4), e14035. https://doi.org/10.1111/psyp.14035

Svirsky, M. A., Neukam, J. D., Capach, N. H., Amichetti, N. M., Lavender, A., & Wingfield, A. (2024). Communication Under Sharply Degraded Auditory Input and the "2-Sentence" Problem. *Ear and Hearing*, 10.1097/AUD.0000000000001500. https://doi.org/10.1097/AUD.0000000000001500

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25*(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Van Engen, K. J., & Peelle, J. E. (2014). Listening effort and accented speech. *Frontiers in Human Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnhum.2014.00577

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the Time Course of Pupillometric Data. *Trends in Hearing*, *23*, 2331216519832483. https://doi.org/10.1177/2331216519832483

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2022). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs* (2.4.1) [Computer software]. https://cran.r-project.org/web/packages/itsadug/index.html

Vickery, B., Fogerty, D., & Dubno, J. R. (2022). Phonological and semantic similarity of misperceived words in babble: Effects of sentence context, age, and hearing loss. *The Journal of the Acoustical Society of America*, *151*(1), 650–662. https://doi.org/10.1121/10.0009367

Warren R. M. (1970). Perceptual restoration of missing speech sounds. Science (New York, N.Y.), 167(3917), 392–393. https://doi.org/10.1126/science.167.3917.392

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, *369*, 67–78. https://doi.org/10.1016/j.heares.2018.05.006

Winn, M. B. (2016). Rapid Release From Listening Effort Resulting From Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, *20*, 233121651666972. https://doi.org/10.1177/2331216516669723

Winn, M. B. (2023). Time Scales and Moments of Listening Effort Revealed in Pupillometry. *Seminars in Hearing*, *44*(2), 106–123. https://doi.org/10.1055/s-0043-1767741

Winn, M. B. (2024). The Effort of Repairing a Misperceived Word Can Impair Perception of Following Words, Especially for Listeners With Cochlear Implants. *Ear and Hearing*, 10.1097/AUD.0000000000001537. https://doi.org/10.1097/AUD.0000000000001537

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The Impact of Auditory Spectral Resolution on Listening Effort Revealed by Pupil Dilation. *Ear and Hearing*, *36*(4), e153-165. https://doi.org/10.1097/AUD.0000000000000145

Winn, M. B., & Moore, A. N. (2018). Pupillometry Reveals That Context Benefit in Speech Perception Can Be Disrupted by Later-Occurring Sounds, Especially in Listeners With Cochlear Implants. *Trends in Hearing*, *22*, 233121651880896. https://doi.org/10.1177/2331216518808962

Winn, M. B., & Teece, K. H. (2021). Listening Effort Is Not the Same as Speech Intelligibility Score. *Trends in Hearing*, *25*, 233121652110276. https://doi.org/10.1177/23312165211027688

Winn, M. B., & Teece, K. H. (2022). Effortful Listening Despite Correct Responses: The Cost of Mental Repair in Sentence Recognition by Listeners With Cochlear Implants. *Journal of Speech, Language, and Hearing Research : JSLHR*, *65*(10), 3966–3980. https://doi.org/10.1044/2022_JSLHR-21-00631

Wood, S. (2023). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation* (1.9-0) [Computer software]. https://cran.r-project.org/web/packages/mgcv/index.html

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge. *Trends in Hearing*, *22*, 233121651877717. https://doi.org/10.1177/2331216518777174

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing*, *31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

Zhang, Y., Lehmann, A., & Deroche, M. (2021). Disentangling listening effort and memory load beyond behavioural evidence: Pupillary response to listening effort during a concurrent memory task. *PLOS ONE*, *16*(3), e0233251. https://doi.org/10.1371/journal.pone.0233251