Do speakers and listeners remember the speech errors or the repairs in communications?

Nikhil V. Lakhani (nlakhani@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

Nazbanou Nozari (bnozari@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

Abstract

Conversations sometimes include speech errors that are repaired. But what do speakers and listeners remember, the error, the repair, or both? In three experiments, we investigated this question by having speakers give instructions for clicking on pictures (Exp 1) or by having listeners follow those instructions by clicking on the referenced pictures (Exps 2 and 3), followed by a surprise recognition test for the spoken words. Results of Exps 1 and 2 showed that both speakers and listeners have better memory for errors than repairs. Exp 3 managed to reverse this pattern by preventing listeners from clicking on the objects that were the referents of speech errors. Collectively, these results suggest superior memory for errors, not when they are simply perceived, but when they are tied to an action.

Keywords: language production; speech errors; repairs; memory

Introduction

The ability to detect and repair speech errors is critical to maintaining effective communication. But what do speakers and listeners retain from an errorful communication? Do they remember the error, the repair or both? The emphasis on the "communicative goal" in language processing (Clark & Brennan, 1991) could suggest that interlocutors should cognitively overwrite speech errors, and any memory thereof, as they clearly do not contribute to the communicative goal and even distract from it. On the other hand, errors are prominent cognitive events. They trigger special mechanisms that alert us to a possible need for regulation in the form of applying greater control, and they help optimize performance through error-based learning (e.g., Maier, Yeung, & Steinhauser, 2011).

In language production, error-based learning drives learning new representations and structures (Branigan & Messenger, 2016; Fazekas, Jessop, Pine, & Rowland, 2020), but it also has a few dark sides; it can lead to the unlearning of other information (Oppenheim, Dell, & Schwartz, 2010), cause lexical perseverations (Hepner & Nozari, 2020), and even lead to the learning of the error itself (Humphreys, Menzies, & Lake, 2010). Such effects imply lingering effects of speech errors in speakers' cognitive systems. In keeping with this, Tydgat, Diependaele, Hartsuiker, and Pickering (2012) had participants name pictures which sometimes unexpectedly switched to a different picture, forcing the speakers to apply a repair. The findings showed a modulation of the repair latencies as a function of the relationship between the error and the repair, with generally shorter and longer latencies for semantically and phonologically related errors and repairs.

In language comprehension, evidence shows that listeners take repairs to replace the error. In an eye-tracking study, Corley (2010) presented participants with conjoined verbs (*eat and move*), repairs (*eat- uh*, *move*), or control utterances (*eat/move*) in sentences like "The boy will [verb(s)] the cake". Eye fixations patterned closely with plain "eat" for conjoined verbs, but with plain "move" for the repair condition, showing that listeners had treated the repair as overwriting the error. Interestingly, upon hearing the theme, the repair and plain "eat" conditions showed similar fixation proportions to "cake", which was taken by the author to imply a lingering effect of the error.

In summary, holding on to errors in a conversation is detrimental for achieving the communicative goal, which is why speakers apply repairs and listeners process them effectively and sometimes even engage in anticipating them based on contextual cues (Lowder & Ferreira, 2019). Nevertheless, there is also evidence that errors may not be completely overwritten by repairs. To our knowledge, the consequences of the possible lingering effects of errors on memory have not been systematically studied. The literature on misinformation and refusal to update beliefs even after correction (e.g., Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012) points to a strong and persistent memory for errors over repairs, but it differs from the issue at hand in two ways. First, in the misinformation literature, the error and repair are often separated by longs gaps in time, and second, the effect is often colored by various emotional, political, and social factors that bias belief systems in complex ways, making it difficult to pinpoint the principled way in which memories are formed for errors followed by immediate repairs in conversations.

Current study

The current study investigated memory for errors and repairs in a communicative setting, free of the influence of belief systems, for speakers (Exp 1) and listeners (Exps 2 and 3). An "error" was defined as a word that, given the communicative context, was not the final intention of the speaker. A "repair", on the other hand, was the word that replaced the error and reflected the final intention of the speaker. The task was adapted from Tydgat et al. (2012) with modifications. Participants believed that they were completing a referential communication task with an online partner. They viewed scenes with three pictures (Fig. 1) and were to either give instructions (speakers) or follow instructions (listeners) to click on a certain picture. In Exp 1, a picture was cued, and speakers had to name it. However, they were told that on some of the trials the cue could unexpectedly change to a different picture. It was the new picture that was the intended target for their partner to click on, and they were instructed to quickly repair their error. Speakers were encouraged to name the pictures as fast as possible, leading to the production of two words on these trials: an *error* (the first name), and a *repair* (the second name). On other trials, no repair was necessary (*unchanged*). A surprise recognition test was administered at the end, comparing memory for errors, repairs, and unchanged words that they spoke.

The next two experiments tested the same in listeners. Participants believed that they were taking part in a referential communication task with another online participant, i.e., a speaker. The actual recordings from Exp 1 were used for this purpose. Listeners were told that they must identify and click on the speakers' intended picture, but that speakers sometimes make speech errors and repair them, and that they should quickly click on the new target picture. As in Exp 1, speed was emphasized in this experiment, leading participants to often click on both error and repair. Exp 3 was similar to Exp 2, but instead of emphasizing speed, encouraged an accurate choice based on the speakers' final intentions, leading them to withhold selection based on speech errors. Listeners in both Exps 2 and 3 received a surprise recognition memory test at the end of the experiment, comparing their memories for errors, repairs, and unchanged words that they heard.

Predictions. If memory is guided by communicative intentions, speech errors and their intended targets should be overwritten in memory. This predicts that a qualitatively similar pattern is expected across all three experiments: both speakers and listeners should remember the repair (i.e., the ultimate intended target) better than the error (i.e., the unintended target), perhaps with larger effects in Exp 3 compared to Exp 2 for limiting action to the intended item. Alternatively, errors may have a special effect on memory by being prominent cognitive events that trigger error-based learning to avoid repeating the error in the future, leading to superior memory for errors compared to repairs—and perhaps even unchanged items—in speakers and listeners. This can manifest in two forms: if *error perception* is sufficient to produce the effect, we would expect superior memory for errors compared to repairs across the three experiments, with stronger effects in Exps 1 and 2 (participants both perceive an error and act upon it) compared to Exp 3 (participants perceive an error but do not act upon it). If, on the other hand, *error commission* is necessary for producing the effect, we would expect superior memory for errors than repairs in Exps 1 and 2, but not for Exp 3. Finally, it is possible that speakers and listeners retain different memories from conversations. For example, error-based learning may be particularly prominent for the speakers, leading to better memory for the spoken errors in Exp 1, while listeners' focus on the communicative goal may predict better memory for repairs in Exps 2 and 3.

Experiment 1

Exp 1 investigated memory for errors and repairs in speakers.

Participants

Sixty native speakers of American English (M_{age} = 21.3; SD_{age} = 2.1; 85% female) were recruited online through Prolific, an online platform for research study recruitment.

Materials and Methods

Three experimental lists of 20 words were created, matched for mean syllable length, phoneme length, character length, and lexical frequency based on the $SUBTLEX_{US}$ corpus (Brysbaert & New, 2009). Additionally, the three lists were matched in semantic similarity between list items calculated as Resnik scores using WordNet (Resnik, 1995), and in phonological similarity using the position-independent phonological overlap metric (Goldrick, Folk, & Rapp, 2010). A control list comprising 60 words was also created, matched to the experimental lists in the indices named above. A total of 120 colored images with white backgrounds from Google Images or from the Bank of Standardized Stimuli (BOSS) was chosen corresponding to the 120 unique words, with name agreement higher than 70% established by a pilot study, and were resized to 22.5% of the display window on side lengths (to be adaptable to all screen sizes for the online study). The three experimental lists were rotated between error, repair, and unchanged item types, generating six versions of the task. Control items remained the same across all versions.

The experiment was coded in jsPsych (De Leeuw, 2015) and hosted on a server running JATOS (Lange, Kühn, & Filevich, 2015). Browser and audio checks were implemented for more uniformity. In each version, participants were presented with 20 trials in the experimental and 20 trials in the baseline conditions. Experimental trials consisted of two critical items—an error and a repair from the experimental lists—and a lure from the control list (Fig. 1). Baseline trials consisted of one critical item—an unchanged item from the experimental list—and two lure items from the control list. Generation of the trial triplets was pseudorandomized with the following constraints across the six versions: each lure appeared once in the experimental and twice in baseline trials, and the mean frequency of lures was matched between the experimental and baseline trials. Two items were paired in a given combination only once across all lists, and semantic and phonological similarities were minimized between items in a given trial (Resnik score of < 5.0; phonological overlap score < 0.5). A participant viewed a given item only once throughout the experiment.

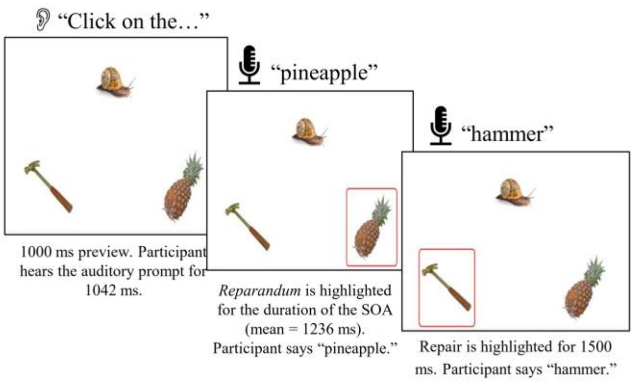

Figure 1. Example of an experimental trial in Exp 1. A baseline trial was similar in structure, but only one item (unchanged) was highlighted.

The production task. Participants were given a cover story that they were playing an online game with a live partner, who would click on pictures on their screen based on the instructions delivered by the participant. A trial began with a 1000 ms preview period (Fig. 1), followed by an audio prompt with a duration of 1042 ms stating "Click on the…". In the baseline condition, immediately upon the end of the audio prompt, a red square framed one of the three pictures (the unchanged item) for 1500 ms. Participants were instructed to name the highlighted picture as quickly and accurately as possible and were told that speed earned them additional points in the game. Next, a blank screen with a central fixation cross appeared for 500 ms before the next trial began. The experimental trials started similarly, but the frame disappeared from the first picture (error) and appeared around a different picture (repair) with a jittered SOA drawn from a Gaussian distribution (M = 1236 ms; SD = 316 ms; Min. = 880 ms; Max. = 1944 ms; Fig 1). Participants were instructed to name the highlighted pictures as quickly and accurately as possible and to repair their response if they noticed a change. They were reminded that speed earned them more points and were discouraged from waiting to detect a change before they started naming the highlighted picture. The trial ended after 1500 ms, at which time a blank screen with a central fixation cross appeared for 500 ms before the next trial began.

Participants watched a tutorial video followed by two practice trials before the experiment began. They then completed four blocks of five experimental and five baseline trials, with breaks in between. After each block, a made-up score was demonstrated along with a reminder to respond as fast and accurately as possible. Order of trial presentation and location of images corresponding to different item types was randomized for each participant. Verbal responses were recorded for offline transcription.

The surprise recognition task. After completing the production task, all 120 words corresponding to the 120 pictures that participants saw in the experiment were presented to them, one at a time, in randomized order, in black on a white background in the center of the screen in 54px Open Sans font, and participants were asked to indicate whether they had spoken the word during the experiment by pressing one of the two buttons as quickly and accurately as possible. They received video instruction on how to place the index and middle fingers of their dominant hand on the 'g' and 'h' keys of the keyboard, counterbalanced to map on to 'yes' and 'no' responses across participants, and completed 12 practice trials with feedback to learn the mapping. A trial ended as soon as a key press was detected or after 3000 ms. A blank screen with a central fixation cross was demonstrated for 500 ms before the next trial began. Keypresses and response times (RTs) were registered for analysis.

Results and Discussion

In around two-thirds of the produced utterances, a repair immediately followed the error. The remaining one-third contained natural disfluencies, including filled pauses (*uh, um*) and interjections (*I mean*), giving speech the natural variability that is observed in the timing of error detection and correction (Nooteboom & Quené, 2017). All items from trials with no responses, incomplete responses, or a different response than the target word were excluded from the recognition analyses (10% of the data, distributed equally among errors, repairs, and unchanged items). The remaining items (1044 errors, 1045 repairs, and 1121 unchanged responses) were analyzed.

On the recognition test, average d′ across all conditions was 2.23 (SD = 0.55), showing good discriminability. The correct rejection for lures was high (91%; SE = 1%). Figure 2 shows the accuracy (Fig. 2a) and RT (Fig. 2b) patterns for the errors, repairs, and unchanged items.

Data were analyzed using multilevel models with mixed random effects in R v.4.0.2 using LmertTest 3.1-3. A logistic link was used for the accuracy data. P-values were calculated using the multcomp package v. 1.4-17, which applies Tukey correction to three pairwise comparisons. Recognition accuracy and RTs were the dependent variables (DVs) and item type (error, repair, unchanged) the independent variable (IV), with random intercepts of subjects and items as the random effect structure. Random slopes were initially included in keeping with recommendations of Barr, Levy, Scheepers, and Tily (2013), but due to the lack of convergence in some models, we report uniform model structures across the three experiments without random slopes. For the RT analyses, values greater than 3SD from the participant's mean RTs were excluded, and the data were log transformed to approximate a normal distribution.

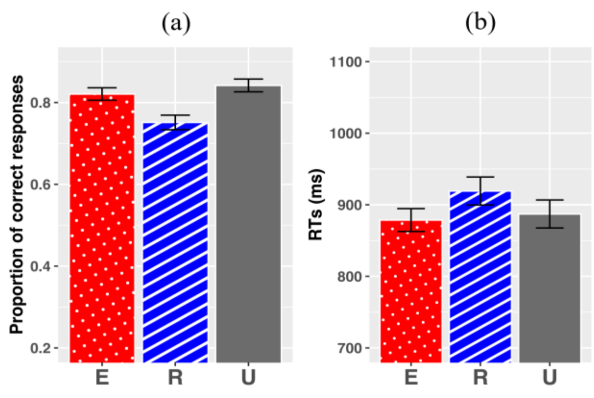

Figure 2. Accuracy (a) and RTs ± SE (b) in Exp 1.
E = Error, R = Repair, U = Unchanged

Model's results on accuracy with Tukey-corrected p-values revealed that recognition accuracy was not significantly different between error and unchanged items ($\beta = 0.15$, $z = 1.25$, $p = 0.21$), but was significantly lower for repairs compared to both unchanged items ($\beta = 0.63$, $z = 5.61$, $p < .001$) and errors ($\beta = 0.49$, $z = 4.31$, $p < .001$). Model's results on RT data mirrored this pattern. Pairwise comparisons with Tukey-corrected p-values showed no significant differences between errors and unchanged items ($\beta = .003$, $z = 0.34$, $p = 0.74$), but RTs were significantly longer for repairs compared to both unchanged items ($\beta = .025$, $z = 2.51$, $p = .024$) and errors ($\beta = .029$, $z = 2.79$, $p = .016$).

In summary, speakers were both slower and less accurate in remembering their repairs compared to both errors and unchanged items.

Experiment 2

Exp 2 investigated the same effect in listeners by using the recordings from participants in Exp 1 to mimic a live referential communication task. Listeners were instructed to be speedy in identifying the speakers' intended picture and clicking on it.

Participants

Sixty native speakers of American English ($M_{age} = 20.6$; $SD_{age} = 2.2$; 63% female) who had not participated in Exp 1 were recruited online through Prolific and similar online recruitment resources.

Materials and Methods

Materials were identical to Exp 1. Audio recordings of participants from Exp 1 were used in Exp 2.

The comprehension task. Participants were told that they were playing an online game with a live participant, who would deliver instructions to them on which pictures to click on. They were asked to follow the instructions as quickly and accurately as they could by clicking on the corresponding pictures. They were warned that speakers may make mistakes and change their answers and were told not to wait for the final answer before they clicked on a picture.

Each participant in Exp 2 was yoked to a participant from Exp 1. Participants viewed the trials in the same order as the yoked participant and heard their instructions verbatim. Arrangement of the pictures on the screen was identical to Exp 1, except no red frame appeared around pictures in Exp 2. Instead, upon hearing the names, participants clicked on the corresponding images. All trials, including the baseline trials, were extended by 2000ms from Exp 1 to ensure that listeners had time to click on both pictures. Yoking kept the timing identical to the original speakers' timing, preserving the natural timeline of the communicative task.

The surprise recognition task. This task was identical to Exp 1.

Results and Discussion

One participant was excluded because of failing to follow instructions. Trials that were excluded from Exp 1 (no responses and responses different from the intended targets) were also excluded from Exp 2. Additionally, a trial was excluded from the recognition task if the participant failed to click on all the target pictures (corresponding to the unchanged item in the baseline trials and both error and repair in the experimental trials; 3% of the data). The remaining items (964 errors, 965 repairs, and 1082 unchanged responses) were analyzed. On the recognition test, average d′ across all conditions was 1.79 (SD = 0.59). Correct rejection of lures was 90% (SE = 1%). Figure 3 shows the accuracy (Fig. 3a) and RTs (Fig. 3b). The same model structure as Exp 1 was used to analyze the data, with Tukey-corrected pairwise comparisons. Recognition accuracy was marginally lower for errors than unchanged items ($\beta = 0.17$, $z = 1.70$, $p = .088$), and significantly lower for repairs compared to both unchanged items ($\beta = 0.41$, $z = 4.15$, $p < .001$) and errors ($\beta = 0.24$, $z = 2.39$, $p = .034$). RTs were not significantly different in any of the comparisons.

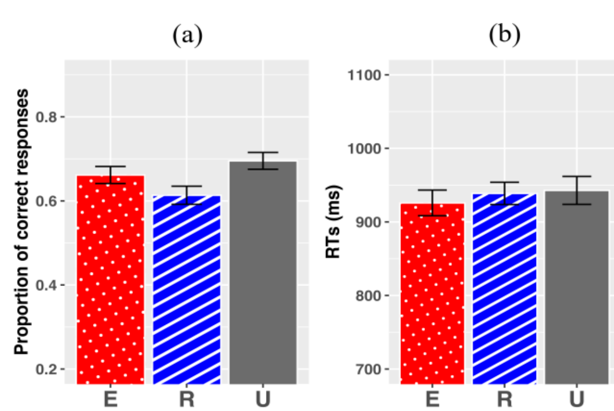

Figure 3. Accuracy (a) and RTs ± SE (b) in Exp 2.
E = Error , R = Repair, U = Unchanged

In summary, the accuracy findings were similar to Exp 1. Listeners remembered repairs less accurately than both errors and unchanged items. Comparable RTs in all conditions ruled out a possibility of speed-accuracy trade-off, although the absence of RT differences in addition to accuracy differences may suggest less robust differences in listeners than speakers.

Experiment 3

The results of Exps 1 and 2 found support for the hypothesis that errors are remembered better than repairs. But in both of those experiments, participants were acting upon errors: speakers in Exp 1 spoke the error words and listeners in Exp 2 clicked on the images corresponding to the errors. Exp 3 tested whether simply perceiving an error was sufficient to create the memory advantage or whether such an advantage was driven by action, by removing the action tied to the error item. A new group of participants heard the same recordings from Exp 1 but were instructed to identify the speaker's final intention before clicking on the corresponding picture, thus discouraging them from clicking on the picture corresponding to the error. Therefore, participants in Exps 2 and 3 were identical in *perceiving* the error, but different in *acting* upon the error.

Participants

Sixty native speakers of American English (M_{age} = 21.4; SD_{age} = 2.0; 77% female) who had not participated in Exps 1 or 2 were recruited online through Prolific.

Materials and Methods

The comprehension task. The same materials and audio recordings as Exps 1 and 2 were used in Exp 3.

The comprehension task. Like Exp 2, participants were told that they were playing an online game with a live participant who would deliver instructions to them on which pictures to click on. But this time, the instructions were changed: instead of encouraging participants to click on images as quickly as they could, they were told that they should identify the speaker's real intention and click on the corresponding picture. They were warned that speakers may make mistakes and change their answers, and to wait for the final answer. To emphasize accuracy over speed, the deadline was extended by 1000 ms on all trials, including the baseline trials. Yoking was done in a similar manner to Exp 2 and timing was identical to that experiment.

The surprise recognition task. This task was identical to Exps 1 and 2.

Results and Discussion

One participant was excluded due to not following the task instructions. Trials excluded from Exp 1 were also excluded from Exp 3. In addition, a trial was excluded from the recognition task if the participant clicked on the error or failed to click on the repair or an unchanged item (6% of the data). The remaining items (936 errors, 939 repairs, and 1057 unchanged responses) were analyzed. On the recognition test, average d′ was 1.61 (SD = 0.57). Correct rejection of lures was 91% (SE = 1%). Figure 4 shows the accuracy (Fig. 4a) and RTs (Fig. 4b). The same statistical models as the previous experiments with Tukey-corrected pairwise comparisons were used. Recognition accuracy was significantly lower for errors than unchanged items (β = 0.57, z = 5.83, $p < .001$), but significantly higher for repairs compared to both unchanged items (β = -0.35, z = -3.49, p = <.001) and errors (β = -0.92, z = -8.90, $p < .001$). Analysis of RTs mirrored a similar pattern. RTs for responding to errors were marginally longer than unchanged items (β = 0.03, z = 2.06, p = 0.08), but significantly shorter for repairs compared to both unchanged items (β = -0.023, z = -1.88, p = 0.08) and errors (β = -0.05, z = -3.84, $p < .001$).

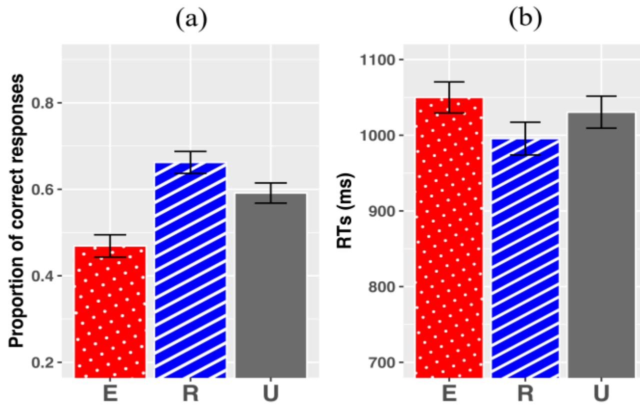

Figure 4. Accuracy (a) and RTs ± SE (b) in Exp 3.
E = Error, R = Repair, U = Unchanged

In summary, preventing action on the perceived errors caused substantially poorer memory for errors in favor of repairs. In fact, a post-hoc one-sample Wilcoxon test demonstrated that the 47% (SE = 3%) accuracy for errors in this experiment was not significantly different from chance (p = 0.31), showing no reliable memory trace for errors at the group level. It is important to point out that this pattern was not a simple consequence of the longer deadlines in Exp 3. Participants did take longer to click on the repair in Exp 3 compared to Exp 2 (5487, SE = 36 ms vs. 5131, SE = 29 ms), but they also took longer to click on the unchanged items (5149, SE = 40 ms vs. 3860, SE = 26 ms; all RTs calculated from the beginning of a trial). Yet the longer RTs had opposite effects on memory for repairs and unchanged items: memory was *better* for repairs (66% ± 3% vs. 61% ± 2%) but *worse* for unchanged items (59% ± 2% vs. 70% ± 2%) in Exp 3 vs. Exp 2, ruling out the extended deadline of Exp 3 as the driving force behind the observed differences.

General Discussion

In three experiments, we evaluated whether speakers and listeners had better memory for errors or repairs in communications. Our first finding was that when errors were tied to actions (speaking the error for the speaker and clicking on the corresponding picture for the listener) both speakers and listeners had better memory for errors than for repairs. These findings do not support the communicative goal hypothesis, which predicts worse memory for errors as words that do not contribute to the communicative goal. Instead, they support an account that considers errors prominent events for memory. The overall higher accuracy and the presence of the effects in both accuracy and RT measures in speakers compared to listeners is aligned with the "production effect", the finding that producing a word substantially strengthens its memory trace (MacLeod & Bodner, 2017). But interestingly, the pattern was qualitatively similar in listeners, suggesting that producing the erroneous word was not necessary for giving it an advantage over the repair. The memory benefit cannot be attributed to timing. Listeners actually had *more* time to select and encode the repair than the error. It is also difficult to account for the pattern by appealing to differential interference. Both proactive and retroactive interference have robust effects on free recall (Unsworth, Brewer, & Spillers, 2013) and recognition (Bowles & Glanzer, 1983). Interestingly, neither speakers nor listeners showed a marked decrease in recognition accuracy for errors compared to the unchanged items, suggesting that the memory formed for the item first acted upon was almost as robust and accurate as the single-item baseline trials. This finding is aligned with studies that suggest a persistent trace of error in the cognitive systems of speakers and listeners (e.g., Corley, 2010; Humphreys et al., 2010; Tydgat et al., 2012), and extends their findings to explicit recognition of such items in later memory probes.

The second main finding of this study was the reversal of this pattern when listeners were prevented from acting upon errors. From a theoretical perspective, this finding shows that simply perceiving an error is not enough to produce a strong memory trace. Rather, it is the action tied to an error that generates the strong memory trace. Once warned against acting upon errors, recognition rate for such errors was at chance. This mimics "directed forgetting" (Basden & Basden, 2013; MacLeod, 1998), the finding that participants have poorer later memory for items that they were instructed to forget. Critically, the current experiment did not include any mention of memorizing or forgetting. Moreover, participants closely monitored the utterance for errors and repairs, and thus must have clearly perceived the errors when they were spoken. Nevertheless, since they were not related to the task goal, they left no reliable trace in memory at the group level. This finding partially supports the communicative goal hypothesis by demonstrating that highlighting the communication goal, at least when directly paired with goal-oriented action, can shift memory towards repairs and nearly efface memory for errors.

Despite the different timelines, the first finding of the study, namely the superior memory for errors than repairs, fits in beautifully with the misinformation literature. The neutral materials of the current study induce little social bias, and yet first utterances are remembered more accurately even after correction, suggesting that the persistence of false beliefs might be, at least in part, due to the basic cognitive processes underlying how utterances are remembered. Our second finding suggests that the problem may be mitigated by discouraging immediate actions related to the error and encouraging vigilant monitoring for repairs, although this may not be easy in practice.

Conclusion

Our results suggest a strong and persistent trace of speech errors in memory, in the absence of any instructions to memorize. This effect, however, can be substantially modulated by the communicative goal, at least for the listener. Collectively, these results suggest a hybrid model, in which memory for conversations is not entirely driven by the conversational goals but can be substantially altered by highlighting such goals.

Limitations and future directions

In the current study, we defined "repairs" and "errors" as utterances that did or did not express the final intention of the speaker, respectively. This is reasonable for testing the predictions of a communicative perspective. Also, the source of error in the speaker's cognitive system is always obscure to the listener, thus from the listeners' standpoint these appear as true semantic errors. But for speakers, the current manipulation represents a special occasion, in which the external world forced a change, not only in the lexical item, but also in the concept of the referent. This is not always—or often—the case with speech errors. Conversely, speech errors and repairs are often produced with the same correct concept in mind (Nozari, Dell, & Schwartz, 2011). Moreover, errors can arise at different levels of the production system. Some may entail the choice of the wrong word, others, the choice of the wrong phonemes (Dell, 1986), and the two are likely to have different consequences for the memory for words. We thus acknowledge that the current results for speakers represent a special situation that may not be extrapolated to other kinds of speech errors. But these results provide a solid basis for comparison with future studies testing memory for errors elicited in other ways.

One might also suggest that the results of Exps 1 and 2 actually demonstrate better memory for the first item than the second item, as opposed to errors and repairs. This is a possibility, but not detrimental to our claim. The reason is that, by definition, errors *always* precede repairs. Therefore, if memory is systematically poorer for the second item, it will inevitably be the repair. Finally, Exp 1 was designed to elicit full utterances as errors. But sometimes errors are interrupted before being fully spoken (Nooteboom & Quené, 2017; Nozari, Martin, & McCloskey, 2019), and this may have different consequences for memory, e.g., by diminishing the influence of the "production effect" (MacLeod & Bodner, 2017). Future research can shed light on these issues.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Basden, B. H., & Basden, D. R. (2013). Directed forgetting: A contrast of methods and interpretations. In *Intentional forgetting* (pp. 151–184). Psychology Press.

Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition*, *11*(3), 307–315.

Branigan, H. P., & Messenger, K. (2016). Consistent and cumulative effects of syntactic experience in children's sentence production: Evidence for error-based implicit learning. *Cognition*, *157*, 250–256.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association.

Corley, M. (2010). Making predictions from speech with repairs: Evidence from eye movements. *Language and Cognitive Processes*, *25*(5), 706–727.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*, *7*(11), 180877.

Goldrick, M., Folk, J. R., & Rapp, B. (2010). Mrs. Malaprop's neighborhood: Using word errors to reveal neighborhood structure. *Journal of Memory and Language*, *62*(2), 113–134.

Hepner, C. R., & Nozari, N. (2020). The dual origin of lexical perseverations in aphasia: Residual activation and incremental learning. *Neuropsychologia*, *147*, 107603.

Humphreys, K. R., Menzies, H., & Lake, J. K. (2010). Repeated speech errors: Evidence for learning. *Cognition*, *117*(2), 151–165.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PloS One*, *10*(6), e0130834.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131.

Lowder, M. W., & Ferreira, F. (2019). I see what you meant to say: Anticipating speech errors during online sentence processing. *Journal of Experimental Psychology: General*, *148*(10), 1849–1858.

MacLeod, C. M. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Lawrence Erlbaum Associates Publishers.

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, *26*(4), 390–395.

Maier, M. E., Yeung, N., & Steinhauser, M. (2011). Error-related brain activity and adjustments of selective attention following errors. *Neuroimage*, *56*(4), 2339–2347.

Nooteboom, S. G., & Quené, H. (2017). Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language*, *95*, 19–35.

Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, *63*(1), 1–33.

Nozari, N., Martin, C. D., & McCloskey, N. (2019). Is repairing speech errors an automatic or a controlled process? Insights from the relationship between error and repair probabilities in English and Spanish. *Language, Cognition and Neuroscience*, *0*(0), 1–16.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *ArXiv Preprint Cmp-Lg/9511007*.

Tydgat, I., Diependaele, K., Hartsuiker, R. J., & Pickering, M. J. (2012). How lingering representations of abandoned context words affect speech production. *Acta Psychologica*, *140*(3), 218–229.

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2013). Focusing the search: Proactive and retroactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(6), 1742.