Is repairing speech errors an automatic or a controlled process?

Insights from the relationship between error and repair probabilities in English and Spanish

Nazbanou Nozari[a,b]*, Clara D. Martin[c,d], and Nicholas McCloskey[e]

[a]Department of Neurology, Johns Hopkins University, Baltimore, USA; [b]Department of Cognitive Science, Johns Hopkins University, Baltimore, USA; [c]Basque Center on Cognition, Brain and Language, San Sebastian, Spain; [d]Ikerbasque - Basque Foundation for Science, Bilbao, Spain; [e]Center for Substance Abuse Research, Lewis Katz School of Medicine, Temple University, Philadelphia, USA.

Correspondences should be addressed to:

Nazbanou Nozari, MD, PhD

Address: 1629 Thames Street, Suite 350, Baltimore, MD 21231, USA

Phone: 443-287-1712

Fax: 410-955-0188

E-mail: nozari@jhu.edu

Abstract

Speakers can correct their speech errors, but the mechanisms behind repairs are still unclear. Some findings, such as the speed of repairs and speakers' occasional unawareness of them, point to an automatic repair process. This paper reports a finding that challenges a purely automatic repair process. Specifically, we show that as error rate increases, so does the proportion of repairs. Twenty highly-proficient English-Spanish bilinguals described dynamic visual events in real time (e.g., "*The blue bottle disappears behind the brown curtain")* in English and Spanish blocks. Both error rates and proportion of corrected errors were higher on (a) noun phrase (NP)2 vs. NP1, and (b) word1 (adjective in English and noun in Spanish) vs. word2 within the NP. These results show a consistent relationship between error and repair probabilities, disentangled from position, compatible with a model in which greater control is recruited in error-prone situations to enhance the effectiveness of repair.

Keywords: speech errors, monitoring, repair, cognitive control, bilingualism, sentence production, semantic interference, incremental planning

Introduction

Speakers detect and correct their speech errors from an early age. Children show evidence of spontaneous error correction almost as soon as they start producing speech (Karmiloff-Smith, 1986), and this ability increases steadily with age (Hanley, Cortis, Budd, & Nozari, 2016) until adulthood. While the mechanisms of error *detection* have received much attention in the literature (see Nozari & Novick, 2017 for a review), relatively little is known about the mechanisms underlying error *correction*. This is in part due to the sparsity of empirical reports on the properties of repairs, which are necessary for proposing a formal model of repairing errors in speech production. This paper investigates one of the key properties of repairs, namely whether they are products of an automatic or a controlled process.

Repair as an automatic process

The question of automaticity is one of the oldest and most debated issues in psychology, dating back to William James in the 19th century (James, 1890). Various criteria, as well as different approaches such as feature-based vs. construct-based views, have been proposed to define automatic processes, and even feature-based approaches have created disagreement on whether judgments of automaticity should hinge on all or a subset of critical features, giving rise to all-or-none vs. decompositional views, respectively (see Moors & De Houwer, 2006 for a comprehensive review). Part of this divergence is due to the large diversity in the scopes and domains to which applying definitions of automaticity has been attempted: for example, automaticity has been investigated in the context of tasks as simple as retrieving the name of a picture to much more complex tasks such as driving from home to work in a crowded city. Despite the differences, however, there is at least some agreement that certain criteria are highly relevant to the question of automaticity, and have also been useful in evaluating automaticity in the language processing system. Some of these criteria include being *unintentional, goal-independent,*

stimulus-driven, fast, efficient (i.e., effortless), and unconscious. As pointed out by Moors and De Houwer (2006), even partial presence of these features is useful for placing a task closer to the automatic end of the automatic vs. controlled spectrum.

In language processing, Fodor's seminal essay on "The modularity of mind" presented a strong defense of automaticity in language comprehension by arguing for its fast, efficient, and mostly unintentional and unconscious nature (as in overhearing others without planning to do so; Fodor, 1983). Interestingly, there were only occasional references to language *production* in that essay, and those few remarks indicated that Fodor viewed language production to be fundamentally different from comprehension in that the former stemmed from thought. The link to thought processes, in Fodor's view, puts language production in the domain of conscious planning, as opposed to automatic processing which he held responsible for comprehension, although he briefly mentioned that the motor part of production may indeed behave more similarly to comprehension in terms of automaticity (Fodor, 1983, p. 42). Fodor's dilemma with language production reflects a larger problem of evaluating the automaticity of production in light of the standard criteria for automaticity. The most important caveat is that language production is, on the one hand, a clearly intentional and goal-oriented task, and such goals affect the inner processes of production such as the placement of a selection criterion (Nozari & Hepner, 2018). On the other hand, many production processes that are carried out to map meaning onto sound, such as spreading activation, do meet many criteria of automaticity such as being fast, efficient and largely subconscious (Dell, 1986). Still some aspects of the same operations, such as resolution of conflict between competing alternatives during both lexical and syntactic processing, seem to require control (see Hartsuiker & Moors, 2016 and Nozari, 2018 for comprehensive reviews).

In the current context, we are less concerned with an ontological classification of repairs as automatic or controlled. The goal, instead, is to define properties of repairs that would bring us closer to developing a functional model of error repairs in speech production. To this end, we will employ some

of the criteria previously proposed to distinguish automatic vs. controlled processes, but we are not committing to any specific theory of automaticity.

In the non-linguistic domain, fast, efficient and sometimes unconscious repairs have been reported in tasks such as the anti-saccade task. In such tasks participants must suppress the natural urge to saccade towards a pop-up stimulus appearing on one side of the visual field and instead move their eyes to the opposite side. Several studies have shown that participants sometimes correct a saccade (e.g., move their eyes to the left, if they had originally performed a saccade to the right), without awareness of having committed and corrected an error (Endrass, Reuter, & Kathmann, 2007; Nieuwenhuis, Ridderinkhof, Blom, Band, & Kok, 2001; Wessel, Danielmeier, & Ullsperger, 2011).

Although not directly speaking to the repair process per se, findings on post-error adjustments also imply that these processes do not depend heavily on conscious processing. One such adjustment is post-error slowing (e.g., Notebaert et al., 2009): participants' tendency to respond more slowly on the trial after an error trial, which has also been shown in the context of speech errors (Freund & Nozari, 2018). In non-verbal tasks, post-error slowing has been observed on trials where participants were not necessarily aware of the error (Hester, Foxe, Molholm, Shpaner, & Garavan, 2005). In a clever study, Hester, Simoes-Franklin, and Garavan (2007) showed that cocaine users, despite having significantly poorer error awareness than non-users, showed nevertheless similar post-slowing behavior.

These and similar pieces of evidence from the non-verbal tasks suggest that post-error adjustments, including repairs, may be at least partially automatic. The tasks from which these findings were recovered, however, are inarguably much simpler than language production, and often have a very limited response set. It is thus important to consider whether an automatic model of repair is even viable in language production.

Evidence pointing to an automatic repair process in language production

At the global task level (i.e., producing the correct word from meaning), it is clear that production has an intention and a goal. On many occasions, there is no external "stimulus" triggering production, so the production process per se cannot be defined as a purely stimulus-driven process. It is thus reasonable to ask whether a repair process that does not meet the explicit criterion of intentionality is even viable. The answer is that although the overall production process has a goal, one could envision a repair process that is partially unintentional. An example of such a process would be simply selecting the next most highly activated item once the produced item is deemed to be an error (see Nooteboom & Quené, 2019, for a similar view). In such a case, the repair process did not set and follow a new goal. Instead, a simple operation, i.e., selection, was triggered when an error was signaled. It is thus theoretically possible to have a repair model that is at least partially unintentional and in some sense "stimulus-driven", i.e., driven by a fixed prior event which in this case is an error signal. In light of this, one can examine whether speech error repairs meet some of the other criteria for automaticity.

Repairing speech errors certainly meets the criteria of fast and efficient processing: errors can be repaired very quickly with very little temporal gap between the point of error interruption and the start of the repair (sometimes as short as 0 ms; Blackmer & Mitton, 1991; Hartsuiker & Kolk, 2001; Nooteboom & Quené, 2019). While not incompatible with a very efficient controlled process, fast operations are often taken to indicate some degree of automaticity (Fodor, 1983; Nozari, 2018). Complementing these findings, the nature of repairs in certain individuals with post-stroke aphasia is also relevant to this debate: some such individuals, despite having difficulty with primary production processes, apply numerous fast repairs to their errors. In conduction aphasia this is called conduite d'approche and is often observed in phonological errors (e.g., Target = igloo: response = /aj-, aj-, ajk-, ajgpl, ajpg-, ajglu, ej, iglu, ajglu, rgglu, glu, o, ajglu, IrJglu, li-, gli-, ajglu/, **igloo,** /iglu/, igloo; (Kohn, 1984)). The interesting finding here is that repair attempts do not always bring the response closer to the target. Sometimes the affected individuals even produce the correct response in the course of their multiple

attempts (the bold and underlined response in the example given here), but move right along to producing other incorrect responses. Similar correction behavior is sometimes observed in non-conduction aphasics attempting to correct their lexical errors. For example, Nozari (2019) reported an individual with intact comprehension and predominantly semantic errors in production who frequently attempted to correct his errors, but sometimes unwittingly passed over a correct repair (e.g., Target = orange; response = apple, pineapple, pumpkin, orange, pineapple, peach?). This pattern is exactly what is predicted by a repair process which automatically outputs the activated alternatives without tight control over matching the response to the target.

Direct evidence for unconscious repair of speech errors is hard to obtain in speech production of neurotypical adult speakers, because single word production (e.g., picture naming) rarely leads to errors to begin with, and failed awareness of errors and repairs in multi-word utterances may reflect memory lapses as opposed to genuine lack of awareness at the time of repair. Nevertheless, apart from anecdotal evidence of unconscious repairs (Laver, 1973), there is some evidence that repairs without full consciousness over the errors may be possible. For example, Postma and Noordanus (1996) asked participants to press a button whenever they detected an error in their speech. They found that occasionally self-repairs were unaccompanied by a button press, although this might simply mean that the participant temporarily forgot the task instructions. In a just-completed study in our lab, participants heard single words and typed them under time pressure without immediately seeing the results on the screen. After each trial, they were asked whether they made a mistake. On 359 occasions, participants denied having made an error, even though they had made an error and had repaired it. These reports were not driven by the outcome (i.e., participants did not deny having made an error more often if the final response was correct), and were observed in the majority of participants (80% out of 60). Still, this effect may be exclusive to typing and not extendable to oral production. Finally, it has been reported

that children as young as 2 years of age, who lack awareness over what went wrong in their speech, can nevertheless repair their errors (Clark, 1978; Karmiloff-Smith, 1986).

To summarize, although direct evidence for an automatic repair process is hard to obtain for language production, various pieces of evidence point to the possibility of such a process.

Evidence pointing to a controlled repair process in language production

There is also indirect evidence for a potential influence of attentional and control processes on error correction. For one thing, the percentage of corrected errors in speech is relatively low, which could suggest that only errors that are attended to are corrected. For example, Nooteboom's (1980) analysis of Meringer's speech corpus revealed a 75% correction rate for phonological errors and only a 53% correction rate for semantic errors. Lapses in attention could explain missed repairs, but these may also result from an automatic process that does not have a high hit rate. A more convincing piece of evidence for the attentional account is that when instructed to correct their errors in the experimental setting, speakers' repair rate often goes up, showing that they can strategically allocate more resources to increase the repair rate. But even under these circumstances, the percentage of repaired errors often remains below 80%, showing some form of resource limitation, generally aligned with the claims of attentional effects on repair processing (Levelt, 1983; Oomen, Postma, & Kolk, 2005), although resource limitation is not a unique feature of controlled processes (Moors & De Houwer, 2006). In a similar vein, error detection performance in children increases with age, with 5, 6, and 8 year olds correcting on average 40%, 50% and 70% of their errors in a sentence production task (Hanley et al., 2016). This is the age when children's attentional and executive control abilities also develop fast (Weighall, 2008), but the influence of other maturation processes cannot be ruled out.

The most convincing evidence for the influence of attention on repairs is the finding that the percentage of detected errors increases towards the end of the utterance. Levelt (1983) had participants

describe colored objects in a network and investigated the proportion of corrected errors on colors that were 0, 1, 2, and 3 syllables away from the end of the phrase (e.g., "And then you come to the *blue*" = 0 syllable gap; "There is a *yellow* node" = 1 syllable gap, etc.). He found that the correction rate for phrase-final errors (i.e., gap 0) was 57%. In comparison, less than 20% of errors in longer gaps were corrected. Levelt (1989) interpreted this finding as an influence of attention on monitoring: attentional resources are used for primary production processes (e.g., message planning, selecting the lexical items, selecting a syntactic frame, etc.) at earlier parts of the sentence, but are shifted to monitoring processes towards the end of the phrase, when they are no longer needed for primary production processes. This explanation is certainly plausible, but there is another striking finding there. The total number of errors in the gap 0 position is 329, 4.5 times greater than the number of errors in gap 1 (n = 73).

The much greater rate of errors towards the end of the utterance suggests a three-way relationship between position, error rate, and correction rate: Correction rate is highest at the end position, which also happens to be the most error-prone position. There are thus two possibilities: a) that the higher correction rate is directly related to position and indirectly related to error rates. This account maintains that speakers strategically shift their attentional processes from primary production to monitoring processes from the beginning to the end of the sentence, as the more primary production processes have been completed towards the end of the sentence. According to this account, the higher error rate in the final position is either irrelevant to repair rates, or may be a result of the strategic shift of attentional processes towards monitoring and repair instead of primary production processes. Either way, it does not view the higher probability of errors as the cause of the higher probability of repairs. Empirically, this account predicts that later positions in the sentence should be associated with a higher probability of repairs, as more primary production processes have been completed in later compared to earlier positions in the sentence.

b) The second possibility is that the repair rate is directly related to the error rate, and only accidentally related to position. The greater proportion of repairs in more error-prone situations could be explained as follows: when production becomes more difficult, the increased need for control leads to the deployment of additional control resources (e.g., Freund & Nozari, 2018). If unsuccessful in preventing errors from surfacing in overt speech, these newly deployed resources could at least catch them after they are uttered. Therefore this account predicts that the probability of repairs may be higher in earlier than later positions in the sentence, if those earlier positions are more error-prone.

Current study

While the characteristics of phonological repairs have been studied in some detail (Nooteboom & Quené, 2013a,b; 2015; 2017; 2019), much less attention has been paid to studying lexical repairs. There are, however, good reasons for studying lexical repairs separately from segmental repairs, as the two show different properties. For example, a study of spontaneous repairs in picture naming in aphasia found that phonological repairs were started faster than lexical repairs, even after controlling for the degree of phonological overlap in lexical and segmental error-repair pairs (Schuchard, Middleton, & Schwartz, 2017). This difference might in turn point to at least partially different mechanisms for detecting and repairing lexical and segmental errors. For example, forward models (e.g., Hickok, 2012), which are highly plausible for monitoring segmental errors, are much less readily applicable to the detection of lexical errors, while other mechanisms such as conflict detection can be easily applied to lexical error detection/repair (Nozari & Novick, 2017).

One notable exception in the relatively understudied field of lexical repairs are studies in which participants were externally prompted to change their response, e.g., because the picture they are naming suddenly changed while they were producing its name (e.g., Hartsuiker, Catchpole, de Jong, & Pickering, 2008; Hartsuiker, Pickering, & De Jong, 2005; Tydgat, Diependaele, Hartsuiker, & Pickering,

2012; Tydgat, Stevens, Hartsuiker, & Pickering, 2011; Van Wijk & Kempen, 1987). While valuable in evaluating many aspects of the repair process including, but not limited to, potential interactions between externally induced control processes and internal dynamics of the language production system (Nozari, Freund, Breining, Rapp, & Gordon, 2016), such paradigms do not capture other aspects of the monitoring and repair processes at work during everyday speech production. Most importantly, they overlook a key feature of repairs, namely that the repair does not just spring into the speaker's consciousness after an error is detected, but in many cases has been actively competing with the error for some time before the error has surfaced. The fact that the repair is already highly activated in the production system by the time of error detection is of vital importance for quick and subconscious repairs. This, in turn, calls for studies that capture lexical repairs under conditions in which co-activation of the reparandum and the repair is a natural consequence of planning an utterance from a conceptual message (e.g., Levelt, 1983). The current study was designed with this goal in mind. Specifically, the design aimed to establish whether the probability of lexical repairs fluctuated with changing control demands and resources, as a function of error position in the sentence or error probability.

As discussed earlier, several pieces of evidence suggest some degree of automaticity in repairing speech errors. Performance in pure automatic processes that are not subject to regulation by control processes such as attention is usually stable within the same individual. For example, if the repair process entails a mechanism of "replacing the rejected response with the next most highly activated response", there is no reason that this mechanism should be implemented more frequently at the end of the sentence than at its beginning. The fact that the proportion of repaired errors is different in different positions (Levelt, 1983, 1989) points to a process that is at least to some degree subject to fluctuations of control. To better understand the relationship between control and repairs, we must first disentangle the relationship between position, error probability, and repair probability. To this end, we conducted an experiment which allowed us to a) attempt to replicate Levelt's (1983) findings of the

three-way relationship between position, error and repair probability, and b) to investigate whether the probability of repair follows position or the probability of errors, and in which direction.

To capture the cognitive processes involved in producing sentences from meaning, with the purpose of a communicative message, we used a paradigm similar to Levelt's (1983). In the "Haunted Hotel" paradigm, participants were presented with objects in a hotel room (e.g., window, curtain, suitcase, telephone). They then learned that they would be viewing scenes of a haunted room, where these objects moved around and interacted with one another. They were instructed to describe what was going on in the room to a confederate (the experimenter) under time pressure. Each event consisted of two objects performing one action, e.g., *"The blue bottle disappears behind the brown curtain*", and could always be described using a [NP1 (Noun Phrase) + verb + NP2] structure. NPs always consisted of a determiner, an adjective, and a noun. The choice of this type of event (and its corresponding structure) allows us to conduct two sets of analyses.

Analysis 1 aimed to replicate the findings of Levelt (1983), by comparing repair rates on NP2 vs. NP1. The NPs were semantically related in two ways: (a) colors which form a taxonomically related category, and (b) nouns that are made to be semantically related in the experiment by creating the common theme of "objects found in the haunted hotel room" which has been introduced to the participants in the orientation phase. Abdel Rahman and Melinger (2011) showed that creating such ad hoc categories is an effective way to induce semantic interference. Because the NPs are semantically related and are named repeatedly, we expected significantly more errors on NP2 compared to NP1, because of the well-established semantic blocking effect (Belke, Meyer, & Damian, 2005; Belke & Stielow, 2013; Damian & Als, 2005; Damian & Bowers, 2003; Nozari et al., 2016; Schnur, Schwartz, Brecher, & Hodgson, 2006; Schnur et al., 2009). This effect, which we suspect also underlies the significantly larger number of errors in the final position in Levelt (1983), would predict higher error rates in the later compared to earlier

positions in the sentence. If the relationship between position, error rate, and probability of repairs observed in Levelt (1983) is robust, analysis 1 should be able to replicate those findings.

Analysis 2 then aimed to disentangle the effect of position and probability of errors on the probability of repairs, by comparing error rates on the two elements of the NP (color and noun). Findings from similar paradigms used in previous experiments suggest that the time pressure of describing dynamic events encourages incremental planning (Arnold & Nozari, 2017). Under these circumstances, there is always less time to plan the earlier element of the NP, compared to the later element (which can be planned as the earlier element is being articulated). For example, Nozari, Arnold, & Thompson-Schill (2014) found that error rates were comparable on adjectives and nouns in NPs such as "red trapezoid" even though the lexical frequency of the noun was much lower than the adjective, which may indicate that the retrieval of the low-frequency noun could have benefitted from the additional planning time provided during the articulation of the adjective. Similarly, anodal stimulation of the left prefrontal cortex was less successful in decreasing the error rates on adjectives compared to nouns, pointing to the limited planning time to produce the early element of the NP as a serious obstacle in producing the correct target word. Based on these findings we predicted that when incremental planning is encouraged by the task (e.g., describing dynamic events under time pressure), earlier elements of the NP would be more error prone than later elements. Therefore, in a sentence like *"The blue bottle disappears behind the brown curtain*", we would expect more errors on "blue" and "brown" (*word1*) compared to "bottle" and "curtain" (*word2*).

If position is the determining factor, the probability of repair should be higher on word2. If, on the other hand, error rate is the critical factor, repair probability should be higher on word1. But there is a problem here: in English adjectives come before nouns, thus the position within NP is systematically confounded with part of speech. It is thus impossible to study the true effect of position on the proportion of repairs unless we can disentangle position and part of speech. The solution to this

problem is to use a language in which the word order within the NP is reversed. For example, the Spanish equivalent of the sentence above would be “La botella azul desaparece por detrás de la cortina marrón.” where “botella” and “cortina” are now word1, and “azul” and “marrón” are word2. If our assumption about the higher probability of error on earlier NP elements holds regardless of part of speech, English and Spanish should show very similar pattern of errors based on position. We can then test the effect of position on repairs without the confound of part of speech.

Adding Spanish has two additional advantages: 1) it provides an opportunity for replication across both analyses. 2) It allows us to examine whether the effects are language-specific or not. As far as predictions go, there is no clear reason to expect that one account would hold in one language but not the other, at least as long as speakers are more or less equally proficient in producing sentences in both languages. To this end, we recruited 20 highly proficient Spanish-English bilingual speakers who still used Spanish with at least some family members and friends, although they lived and worked in an English-speaking environment (Baltimore, Maryland, United States). They completed four blocks of the same task, two in English, and two in Spanish in counterbalanced order, generating a corpus of 500+ speech errors on NPs. Note that our primary interest in this experiment is not to compare monitoring in L1 and L2. As far as proficiency and fluency are comparable for the task at hand, we would expect fairly comparable baseline performance in terms of number of errors and repairs across the two languages. Thus, in this case, the use of two languages is aimed at replicating the same findings in two languages, and ruling out the confound of part of speech in order to have a clean test of the two hypotheses of main interest.

Below is a summary of the three possible outcomes of the experiment and accounts supported by each:

1) No relationship between error rates, position, and repair proportions (i.e., a failure to replicate Levelt’s 1983 findings). This scenario would be most compatible with an account in which repair

processes are not sensitive to change in control demands and resources, i.e., a *purely automatic account.*

2) An increase in repair proportions for later —as opposed to earlier— parts of the utterance (i.e., NP2 > NP1 and W2 > W1). This finding would support Levelt's (1983) account of controlled repairs by suggesting the reallocation of resources from primary production processes to monitoring and repair processes in later parts of the utterance. We call this the *controlled position-based* account.

3) A change to the probability of repairs covarying with the probability of errors within speakers, irrespective of the position of the error in the sentence. This pattern would support a *controlled error-based account*, and could itself show one of the following directions:

3a- A decrease in repair proportions with increasing error probability. This would point to the sensitivity of repairs to control demands, and would specifically support a model in which primary production and repair processes share a limited pool of resource. When primary production conditions become more difficult, error rates increase. Since repair processes share the same resources, without further adjustments in the system, proportion of repairs would decrease. We call this the *fixed-resource account.*

3b- An increase in repair proportions with increasing error probability. Similar to scenario 3a, this scenario would point to the sensitivity of the repair process to fluctuations of control. However, in this case, when primary production processes become more difficult, the system adjustably recruits more resources. Since these resources may not be adequate to prevent the problem to begin with, they may manifest as increased repair rates. We call this the *adjustable-resource account.*

Methods

In order to elicit errors and observe corrections, we strived to capture the natural production and monitoring processes as much as possible in an experimental environment. To this end, we chose

paradigms that require producing an utterance with a communicative intention from a conceptual (visual) event (e.g., Arnold & Nozari, 2017; Nozari & Omaki, 2018; Nozari, Arnold, & Thompson-Schill, 2014).

Participants

Twenty English-Spanish bilingual speakers (12 females; M_{age} = 21, SD_{age} = 2.4 years) were recruited from Johns Hopkins community and participated in the study in exchange for payment. The number of participants was fixed before data collection without sample size calculation, because there are no prior effect sizes available for many effects of interest in the study. All participants gave informed consent under a protocol approved by the Institutional Review Board of Johns Hopkins School of Medicine. Participants were all highly proficient in Spanish (self-reported proficiency on a 1-10 scale = 9.3, *SD* = 0.9) and English (proficiency = 9.9, *SD* = 0.4), and acquired both languages early in life ($AoA_{Spanish}$ = 0.0, *SD* = 0; $AoA_{English}$ = 3.4, *SD* = 2.7). Besides their high proficiency in both languages, they were dominant in English: Eighteen of them reported English as their dominant language, one of them reported Spanish, and one reported equal dominance in both. Their dominance in English was also revealed by a vocabulary test (picture naming in each of the languages), in which participants produced significantly more words in English (63.2, *SD* = 1.1 out of 65) than in Spanish (52.2, *SD* = 8.2).

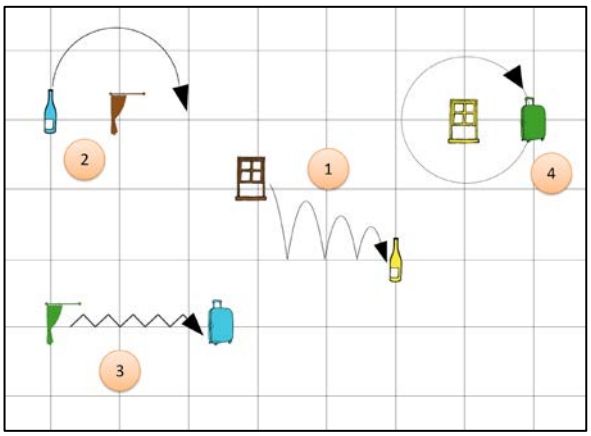

Figure 1. Example of a slide with four events that unfold in the following sequential order: 1. Bouncing towards, 2. Jumping over, 3. Zigzagging towards, and 4. Looping around. The lines show the motion path and the arrows the direction of the movement. The experiment contained 56 slides, half in English and half in Spanish blocks for a total of 224 events.

Materials

The "Haunted Hotel" paradigm consisted of 224 events. Each event entailed an interaction between two colored objects (see Table 1 for the list of

possible objects, colors, and actions, and Figure 1 for pictorial examples) in a PowerPoint slide. For example, participants saw a blue bottle disappearing behind a brown curtain, corresponding to the English sentence “The blue bottle disappears behind the brown curtain” or the Spanish sentence “La botella azul desaparece por detrás de la cortina marrón”. The events could always be described using a [NP1 + verb + NP2] structure, and NPs always consisted of a determiner, an adjective and a noun. Two sets of objects were used to create more diversity in utterances (sets 1 and 2 in Table 1). Each set (112 events) was divided into two blocks (a total of four blocks per participant), one to be described in English and one to be described in Spanish. The order of blocks and events were counterbalanced between participants, such that each participant started one set of two blocks in English and another set of two blocks in Spanish.

Spanish nouns in each set were either all feminine (set 1) or all masculine (set 2) to minimize the difficulty of choosing the correct determiner and make the English and Spanish blocks as close to each other as possible. For the same reason, to the extent possible (3 out of 4) we chose adjectives whose form did not change based on the noun’s gender in Spanish. Half of the nouns and half of the verbs were phonologically-related in English and Spanish, but this manipulation is not relevant to the current goals and will not be further discussed in this paper. Nouns, adjectives, and verbs were matched in frequency and length (in letter but not in phonemes) in English and

Table 1. The linguistic materials.

Nouns	
Set 1	
English	**Spanish**
bottle	(la) botella
curtain	(la) cortina
window	(la) ventana
suitcase	(la) maleta
Set 2	
telephone	(el) teléfono
package	(el) paquete
mirror	(el) espejo
newspaper	(el) periódico
adjectives	
English	**Spanish**
green	verde
brown	marrón
yellow	amarillo
blue	azul
verbs	
English	**Spanish**
disappears (behind)	desaparecer (por detrás)
pass (behind)	pasar (por detrás)
produce	producir
zigzag (towards)	zigzaguear (hacia)
jump (over)	saltar (por encima)
loop (around)	rodear
bounce (towards)	brincar (hacia)
bump (into)	chocar (con)

Spanish. Nouns in the two sets were matched in frequency, as well as letter, phoneme and syllable length in both languages (see Appendix A).

Each of the four blocks was divided into 14 slides, each containing four consequent events (for a total of 56 events per block, and 224 per participant). The four events on each slide each took between 2 and 4 seconds and were separated from one another by a 1.5-second interval. Several aspects of the design aimed to elicit incremental planning as the optimal strategy for completing the task: For one thing, half of the actions were ambiguous before the completion of the event. For example, "disappearing behind" and "passing behind" started as identical events, with the difference that in the former case the first object never emerged from behind the second object, but in the latter case it did. This ambiguity forced participants to wait until the action was completed to plan later parts of the sentence. Moreover, in some cases, the second object was not clear when the action started. For example, an object may move towards the general direction of two objects before bumping into one of them. These two manipulations would make planning of the entire sentence at the onset of the event impossible in many cases. At the same time, the timing of events had been carefully set using extensive pilot testing such that speakers would run out of time if they held off on planning the first NP until the action had been finished. Collectively, these sets of constraints strongly encouraged an incremental mode of planning: if participants planned chunks of the sentence as actions unfolded in real time, they could speak at a normal rate without falling behind. This mode of speaking makes substantial overlap between planning of NP1 and NP2 unlikely.

Procedures

Participants completed two sessions. The first session comprised language proficiency tests in English and Spanish. They completed the Haunted Hotel task in the second session. At the beginning of the second session, participants completed a short orientation, introducing the objects and actions to be

seen in the experiments. They were told that all the objects were found in a haunted hotel room (the common theme to make the semantic relations clear), that these objects would move around in the room, and that their motions needed to be described to the experimenter. Participants named all objects, colors, and actions in both English and Spanish and were corrected if necessary. They then practiced describing each action using slides with only two objects on the screen until they could produce the descriptions fluently. Finally, they completed two practice slides each with four events, similar to the experimental slides. Practice was repeated if necessary, until participants could describe the actions easily at the normal speech rate. Half the participants started the first set with the English and the other half with the Spanish block. Since the second set consisted of new objects, the orientation and practice were repeated with the new objects. All instructions were delivered in the language corresponding to the first block of the set. The order of the blocks in the second set was then switched, such that participants who had started the first set with an English block, started the second set with a Spanish block and vice versa. Correspondingly, the language of the instructions, orientation, and practice was changed in the second set, such that each participant was instructed once in English and once in Spanish throughout the experiment.

During the experiment, participants produced four sentences corresponding to the four events of one slide without taking a break in between (except for the 1.5 constant gap between the events). After each slide, however, they could take a break, and advance to the next slide whenever they were ready. This rhythm created sentence sets of four, before the flow of speech was interrupted.

Statistical analyses

All analyses are conducted once using non-parametric tests and once using multi-level mixed models (MLMs; e.g., Jaeger, 2008) in R version 3.4.3, using the lmerTest package version 3.0-1 (Kuznetsova, Brockhoff, & Christensen, 2017). Non-parametric tests have the advantage of being simple and requiring few assumptions, thus removing worries about the results being unreliable due to complexities in the

model structure (as in MLMs) or violation of the model's basic assumptions by the data. MLMs on the other hand, have the advantage of capturing a much more complex structure, e.g., items nested under subjects and crossed subject-item dataframes. They can model random effects of subjects and items, and are suitable for testing main effects and interactions in the same model. This complexity, however, comes with issues of overfitting, lack of convergence, and sometimes violations of assumptions. We, therefore, report the results of both methods for the analyses of interest, and look for general convergence between the results obtained from the two methods. In fitting the MLMs, we have aimed for the maximal random effect structure the model can handle, in keeping with the recommendations of Barr, Levy, Scheepers, and Tily (2013). This structure was simplified for post-hoc models where the inclusion of the full random effect caused overfitting. The random effect structure was kept consistent between comparable models, e.g., post-hoc tests in English and Spanish. Key effects are reported in the text. Full tables reporting the details of the MLM analyses can be found in Appendix B.

Results

For the purposes of this paper, we focus only on the NPs. Determiner errors were rare and are not counted in this coding scheme. The rate of misses (i.e., where no description was provided for an event) was less than 1% for English sentences and 2.5% for the Spanish sentences. This difference was not significant (z = 1.23, p = .22). Any deviations from the target utterance, complete (e.g., "yellow" for "green") or incomplete (e.g., "yell-" for green) was counted as an error. We collected 573 errors on NPs in total. Out of these, 54% were color and 46% noun errors. Participants made a total of 264 errors in English (M = 12.75; SE = 1.99) and 309 (M = 15.7; SE = 2.38) in Spanish. The difference was not significant (z = 1.64, p = .10). Corrections were coded as attempts to revise an uttered word, whether complete (e.g., "yellow, no, green"), or incomplete ("yell-green"). The majority of errors (91%) were corrected immediately, i.e., with no intervening words between the reparandum and the repair, and two thirds of

these errors were corrected before the reparandum was completed. The overall repair rate was 79% in English and 71% in Spanish, also not significantly different from one another (z = 0.14, p = .89). These general findings suggest that there were no obvious differences in baseline performance in English and Spanish in our bilingual group, allowing us to compare the pattern of errors and corrections across the two languages without worrying about large imbalances in performance potentially affecting these patterns.

Analysis of NP1 vs. NP2

First, we examined the error rates on NP1 and NP2. The left panel in Figure 2 shows these data. Non-parametric tests revealed significantly more errors on NP2 compared to NP1 collapsed over both languages (z = 3.06, p = .002), and post-hoc tests showed the significance of the effects independently in English (z = 2.55, p = .011) and in Spanish (z = 2.15, p = .03). The results of the MLM analyses converged: there was a main effect of NP position (z = 2.07, p = .038), and no interaction between NP position and language (z = 0.19, p = .85; Table B1). Post-hoc MLMs revealed significant effects of NP position in both English (z = 2.71, p = .007; Table B2) and Spanish (z = 3.026, p = .002; Table B3). An analysis of sentence position within a set (sentences 1, 2, 3 and 4 spoken consecutively without a break) also returned a significant effect of NP position, consistent with the prior analysis (z = 3.05, p = .002), as well as an effect of sentence position, such that error rates increased from sentence 1 to 4 (z = 2.69, p = .007), and a marginal interaction between the two (z = -1.71, p = .09).

Next, we examined the proportion of corrected errors on NP1 and NP2. The right panel of Figure 2 shows these data. Reflecting a similar pattern to that found for errors, non-parametric tests revealed a significantly higher proportion of corrected errors on NP2 compared to NP1 collapsed over both languages (z = 2.77, p = .006), and post-hoc tests showed the significance of the effects independently in English (z = 2.02, p = .028) and in Spanish (z = 2.62, p = .009). The MLM analysis also showed a main effect of NP position (z = 1.98, p = .048), with no interaction between NP position and language (z = -.45,

p = .65; Table B4). Post-hoc MLMs revealed a significant effect of NP position in English (*z* = 2.45, p = .014; Table B5) and Spanish (*z* = 1.99, *p* = .047; Table B6).

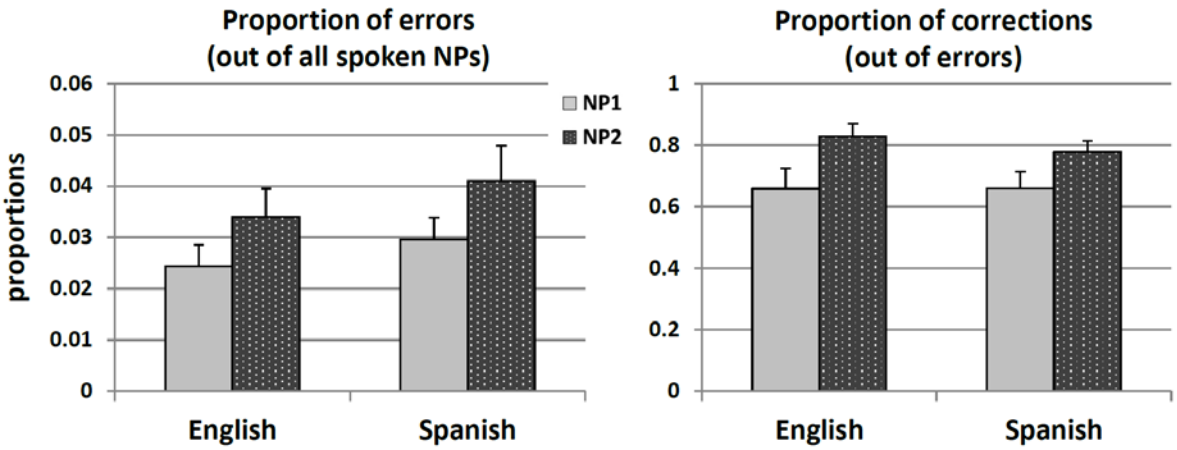

Figure 2. Proportion of errors (left) and corrections (right) on NP1 (solid light gray bars) and NP2 (polka dot black bars) in English and Spanish. Height of the bars reflects the mean of subject means and the error bars are SEs.

Discussion

The results of the error analyses provided robust evidence for the increase in the error rates from NP1 to NP2 in both languages, as expected by the semantic blocking effect. Moreover, the results showed an increase in error rates from sentence 1 to 4, consistent with a cumulative effect of semantic interference. The findings are also in agreement with Levelt's (1989) demonstration of increased error rates towards the end of the utterance, which, in both cases, can be explained by increased semantic interference. The main question of interest was whether the correction process is sensitive to the probability of committing an error. The results suggest that when error rates increase, so does the proportion of corrected errors. These findings provide an independent replication of Levelt's (1989) findings, and show a three-way relationship between position, error, and proportion of corrections. Similar to that experiment, however, it is not possible to disentangle the effect of position and proportion of errors on the proportion of corrections. The next analysis aims to do that.

Analysis of the first vs. second word within the NP

This analysis focuses on the errors and corrections on word1 and word2 collapsed over NP1 and NP2. The left panel in Figure 3 shows the error data. Non-parametric tests revealed significantly more errors on word1 compared to word2 collapsed over both languages (z = 3.50, p <.001). Post-hoc tests showed that the effect was significant in English (z = 3.27, p = .001) and marginal in Spanish (z = 1.92, p = .055). MLM analyses confirmed these results: there was a significant main effect of word position (z = -3.47, p<.001), and no main effect of language (z = -.046, p = .96). There was also an interaction between word

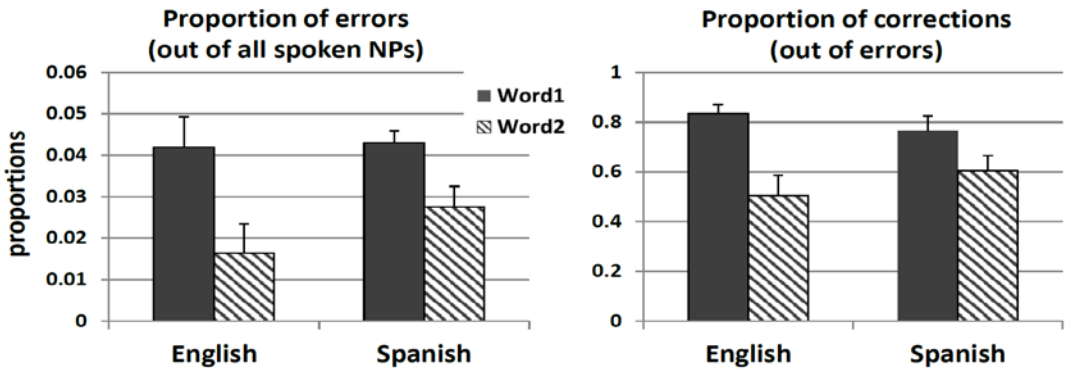

Figure 3. Proportion of errors (left) and corrections (right) on word 1 within the NP (solid black bars) and word 2 (shaded bars) in English and Spanish. Height of the bars reflects the mean of subject means and the error bars are SEs.

position and language, suggesting that the difference between error rates on word1 and word2 was more prominent in English than in Spanish (z = 2.02, p = .044; Table B7). Post-hoc MLM tests revealed a significant effect of word position in English (z = -6.95, p<.001; Table B8), and a marginal effect in Spanish (z = -1.88, p = .06; Table B9).

Next, we examined the proportion of corrected errors on word1 and word2 collapsed over the two NPs. The right panel of Figure 3 shows these data. In keeping with the results of the NP analysis, corrections followed a very similar pattern as errors. Non-parametric tests revealed significantly more corrections on word1 compared to word 2 (z = 3.70, p <.001). Post-hoc tests showed this effect to be significant in English (z = 3.21, p = .001) and marginal in Spanish (z = 1.78, p = .078). MLM analyses also showed a significant main effect of word position (z = -4.61, p<.001), a marginal effect of language (z = -

1.70, p =.089), and no interaction between word position and language (z = 1.29, p = .197; Table B10). Post-hoc analyses using the MLM revealed significant effects in both English (z = -4.40, p<.001; Table B11) and in Spanish (z = -3.25, p = .001; Table B12).

Discussion

As predicted, there were more errors on the first compared to the second element of the noun phrase because of the time pressure during the NP planning. The interaction between word position and language suggests that the difference between error rates on word1 and word2 is greater in English than Spanish. As can be seen in Figure 3, this difference is driven by the greater reduction of errors on word2 in English compared to Spanish. This means that participants were better able to plan the second element of the NP while articulating the first element in the dominant (English) language. Important for the purpose of this study, the greater likelihood of committing an error on the first compared to the second element of the NP held in both English and Spanish, despite the fact that word1 and word2 comprised different parts of speech in these languages. The results, thus, show that later positions in the phrase are not necessarily associated with higher error rates, i.e., that error rates do not monotonically increase towards the end of the utterance.

We can now ask whether the likelihood of correction is sensitive to the position of the reparandum in the phrase (in which case proportion of corrections should be higher on word2 compared to word1) or to the error rates (in which case proportion of corrections should be higher on word1 compared to word2). The results supported the second possibility. Probability of correction closely followed the probability of errors, suggesting that it is indeed the likelihood of committing an error, and not the position in the utterance which guides corrective behavior.

General Discussion

The study investigated repair to lexical errors in a task in which bilingual participants produced sentences by describing simple visual events. While the goal of the current study was not to compare monitoring in L1 vs. L2, it is a useful addition to the sparse empirical reports on bilingual monitoring in that it reports on lexical monitoring in a homogenous, highly proficient bilingual group of speakers who produce a large number of errors and repairs. A few studies that have looked at this issue are either low on statistical power (e.g., Van Hest, 1996), or have focused on phonological errors and repairs (e.g., Broos, Duyck, & Hartsuiker, in press). As expected, overall performance in this group was comparable in English and Spanish. One difference was a greater reduction of errors on word2 within the NP in English compared to Spanish. This finding suggests better ability for parallel planning of word2 during the articulation of word1 in the dominant language (Hanulová, Davidson, & Indefrey, 2011; Kroll, Bobb, Misra, & Guo, 2008). The overall pattern of errors was also very similar in the two languages (Fig. 4, upper panel): NP2 was significantly more error-prone than NP1, reflecting the semantic blocking effect (Belke et al., 2005; Belke & Stielow, 2013; Damian & Als, 2005; Nozari et al., 2016; Schnur et al., 2006; 2009) at the sentence production level. Moreover, error rates on sentences 1 to 4 within a set showed a monotonic increase, pointing to the cumulating nature of interference (see Belke & Stielow, 2013 for a discussion). While semantic interference is often observed on response latencies in neurotypical speakers, the effect manifested in error rates here because the current design requires incremental planning and thus shifts the decision point on lexical retrieval towards speedy but error-prone processing (Nozari & Hepner, 2018).

The pattern of within-NP errors was also similar in the two languages: there were always more errors on word1 compared to word2, regardless of whether word1 was an adjective (English) or a noun (Spanish; a marginal effect here). Therefore our basic assumptions about the distribution of error probabilities across the sentence were verified in both languages. Importantly, the greater probability of errors on word1 vs. word2 within the NP broke the monotonic relationship between error rates and

later positions in the sentence, allowing us to disentangle the effect of position from error probability. Three possible outcomes were evaluated: 1) a *purely automatic account* in which repair probability is insensitive to changes in control demands either related to position or error probability. 2) A *position-based controlled account* in which repair probability increases monotonically for later positions as more primary production processes have been completed. 3) An *error-based controlled account* in which repair probability changes with error probability. The *inflexible-resource account* version of the error-based controlled account predicts that repair probability would decrease with increasing error probability. Conversely, the *adjustable-resource account* version predicts an increase in the repair probability with increasing error rates.

Critical findings

We replicated the finding of Levelt (1983) by showing greater proportion of detected errors on the later and more error-prone NP2 compared to NP1. This finding refutes the *purely automatic account*. The results of the analysis of within-NP repairs, however, showed that it is the probability of errors, and not the position in the sentence, that determines the probability of repair. This can be seen clearly in Figure 4, which shows that the pattern of repairs (lower panel) is strikingly similar to the pattern of errors (upper panel); when error rates rise, so do the proportion of corrected errors. Position, on the other hand, is only predictive of higher repair rates when it is also associated with higher error rates, for example, the later NP2 is also more error-prone than the earlier NP1. When later position is not associated with higher error rates, as in the case of word2 vs. word1 within the NP, the probability of repair tracks error rates, and not position. These findings rule out the *controlled position-based account,* in favor of the *controlled error-based account*. More specifically, they support the *adjustable-resource account.*

It is worth noting that even though the constraints on repairing lexical and segmental errors may be different because of the different nature of representations and the different time-constraints of lexical selection and phonological encoding processes (Nooteboom, 2005); the current findings in lexical repairs have support in phonological repairs as well. Consonant errors are more likely in the onset position than in any other position in the word (e.g., Nooteboom & Quené, 2015; Shattuck-Hufnagel, 1992) and onset position also happens to have higher repair rates, as well as shorter cutoff-to-repair times, than other positions (Nooteboom & Quené, 2019). These findings mirror those reported in this experiment for lexical errors, and point to more efficient repair processes where the error probability is the highest in both lexical and segmental domains.

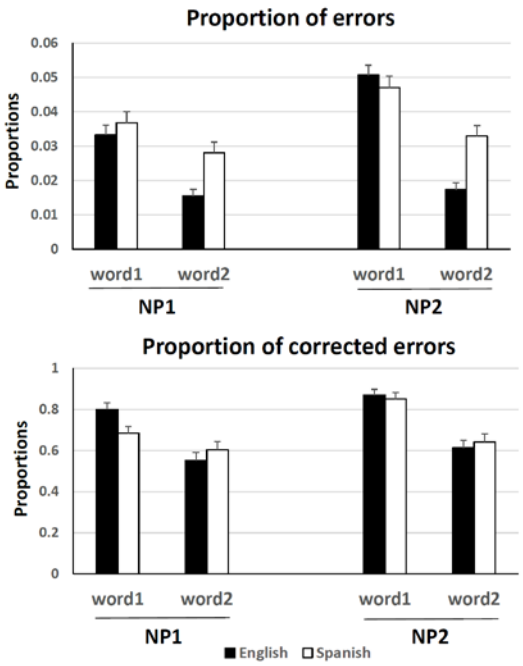

Figure 4. Summary of all the data. Proportion of errors (upper panel) and corrected errors (lower panel) for word1 and word2 in NP1 and NP2 in English (black bars) and Spanish (white bars). The figure shows the similarity between the fluctuations of error rates (decrease from word1 to word2, increase from NP1 to NP2) in the two languages, and the similarity between this pattern and the pattern of corrections.

Before we discuss an account of error repair that accommodates these findings, we must address what may seem like a discrepancy with prior reports. We have previously reported evidence for a relationship between the quality of the production system and the quality of error detection/repair. In what was summarized as the conflict-based account of monitoring, we showed that damaged (Nozari, Dell, & Schwartz, 2011) or immature (Hanley et al., 2016) production systems are associated with higher levels of conflict between the target and the competitors. This high conflict gives rise to higher error rates, and also poorer distinction between error and correct trials (see Nozari & Novick, 2017, for a short review). The conflict-based model, thus, predicts a reverse relationship between the number of errors

(in so far as they represent the quality of the production system) and the ability to detect/repair those errors, the exact opposite of what we have reported here.

The critical distinction here is between the general state of the production system (i.e., that which cannot be much improved by implementation of control) and the temporary states (i.e., those which can benefit from the implementation of control). For example, if semantic-lexical mapping has been damaged (e.g., after stroke), the amount of conflict between target and non-target lexical representations during each naming attempt cannot be decreased more than a certain amount no matter how much control is exerted. In other words, high conflict has become a "trait" of the individual. On the other hand, a healthy system may temporarily be in a "state" of high conflict (e.g., because of time pressure during planning), which can be effectively resolved by deployment of control processes. This trait/state difference is important in making predictions regarding the relationship between error rates and proportion of repaired errors: when high conflict is a "trait", more errors and worse detection/correction performance is expected. This means that *across individuals*, one would expect a negative correlation between the probability of errors and the probability of repairs. In keeping with this prediction, we observed a significant negative correlation between the proportion of errors they commit in Spanish and the ability to detect/correct those errors (r = -.48, p = .034)[1] across the individuals, adding to the results of previous studies showing a similar correspondence between the quality of production and the quality of monitoring/repair in individuals with aphasia and children (Hanley et al., 2016; Nozari et al., 2011). *Within the same individuals*, however, fluctuations of conflict as a "state" could still lead to better recruitment of control, hence the pattern of covariance between error and repair probabilities reported in this study.

An account of repair in speech production

[1] This correlation is also negative in English (r = -.27) but does not reach significance (p = .25) most likely because of the smaller variability in the number of errors made in English across subjects.

Generally-speaking, two classes of accounts have been proposed for repairing speech errors. The first type views the repair process as re-initiating the production process from scratch, albeit with some degree of priming of the repair by the reparandum (Hartsuiker & Kolk, 2001). This presumably requires recommitting to the goal or defining a new goal to restart the production process. The second class views the repair process as a form of revising the previous plan (e.g., Boland, Hartsuiker, Pickering, & Postma, 2005). A version of this account was introduced in the Introduction: detection of an error triggers a repair process which simply entails selecting the most highly activated item that was *not* produced (see also Nooteboom & Quené, 2019). Note that this process does not require resetting or recommitting to a goal, as it picks up production from the middle of the process. This account is reasonable, because while there is a debate on whether activation of non-target responses slows down the production of target (e.g., Abdel Rahman & Melinger, 2009; Mahon, Costa, Peterson, Vargas, & Caramazza, 2007; Nozari & Hepner, 2018), it is agreed by all production accounts that similar target and non-target representations can and do become activated simultaneously. Moreover, this simple mechanism could lead to efficient repair behavior at least in fully developed neurotypical systems, because the target response is often highly activated even on trials where a non-target response has been erroneously produced. Support for this claim comes from studies reporting responses that seem to be a blend of the correct and error responses suggesting the coactivation ---and the tendency for selection--- of both in close temporal vicinity (Goldrick & Blumstein, 2006; McMillan & Corley, 2010; Nooteboom & Quené, 2008). In weaker production systems, e.g., those of children and individuals with aphasia, this mechanism still leads to repair attempts, although the success rate of such repairs would be lower since competitors have comparatively high levels of activations and may be selected mistakenly as the repair. This kind of process can account well for findings such as conduite d'approche discussed in the earlier sections, and why speakers may inadvertently switch a correct response to a

wrong one, as repairs are done fairly automatically, as long as alternatives with high activation are available (e.g., Nozari, 2019).

This purely automatic account, however, cannot by itself explain the relationship observed between the error rates and proportion of corrected errors found here, and also reported in Levelt (1983) and Nooteboom and Quené (2019). The most straightforward prediction is that repair attempts in the purely automatic account should remain fairly insensitive to error probabilities. If anything, one would expect a decrease in the accuracy of repair outcomes, because although an automatic process may initiate lots of repairs in error-prone situations (as in conduite d'approche), more competitors have activation levels comparable to the target and can be selected as potential repairs by an automatic system. It is thus unlikely that correct repairs are more prevalent under error-prone conditions.

An automatic mechanism can, however, be augmented by cognitive control. The increase in the rate of repairs when speakers make more errors is compatible with greater recruitment of control resources in these situations. More precisely, when competition is high, the monitor detects the need for greater control, signals this need to control centers, which in turn deploy the necessary control to the part of the system where such control is needed to resolve the conflict between competing responses. This control is primarily put towards preventing the generation of an error (Nozari & Hepner, 2018); however, since speakers are under time pressure, sometimes the control cannot resolve the conflict in a timely manner, and an overt error is generated. Under such circumstances, the control system exerts its influence to repair the errors, e.g., by suppressing the activation of the recently produced error or by boosting the activation of the next most highly activated response so that it could overtake the error. Since error-prone conditions are precisely those conditions that recruit the highest amount of control, a controlled process of repair can explain the direct relationship between the probabilities of error commission and repair.

To summarize, the empirical findings on repairing speech errors point to an underlying automatic process that provides "quick and dirty" repairs. This mechanism can explain many findings such as the quick timeline of repairs, as well as a rapid sequence of repair attempts in individuals with brain damage, and potentially cases of unconscious repairs in production. The automatic account, however, needs to be augmented with control processes to explain effects such as increased probability of error correction as a function of increased error rates. Such control may be implemented strategically, i.e., speakers may become aware of the difficult parts of production and consciously divert attentional control to those parts. This possibility shifts the locus of the effect from *correction* to *detection*. An alternative —non-mutually-exclusive— explanation is that control is implicitly assessed and implemented when necessary through incremental learning processes which help regulate production (e.g., Freund & Nozari, 2018). It is a key critical question for future research to determine the nature of control processes that contribute to repairs and the circumstances under which a certain type of control is triggered. The answer to this question may also shed more light on why some repairs seem to be readily available after error detection, while others take a while to be planned (Nooteboom & Quené, 2017).

Another critical question for future studies concerns the exact mechanisms by which control may enhance the repair behavior in error-prone situations. One mechanism could be the triggering of an activation boost (Gauvin & Hartsuiker, under review). An alternative could be stronger suppression of the reparandum to give the repair a better chance of production, or a combined activation-suppression account.

Acknowledgment

We thank Michael Freund for his help with data collection. This work was supported in part by the NSF grant 1631993 awarded to N.N., and in part by the Therapeutic Cognitive Neuroscience Fund endowed to the Cognitive Neurology division of the Neurology Department at Johns Hopkins University. C.D. Martin was supported by the Spanish Ministry of Economy and Competitiveness (SEV-2015-490; PSI2017-82941-P; Europa-Excelencia ERC2018-092833) and the Basque Government (PIBA18-29).

Disclosure of interest

The authors declare no conflict of interest.

Data availability statement

The data corpus used in this paper is currently being analyzed for (a) testing the predictions of an implemented computational model of speech repairs, and (b) the effects directly related to bilingualism. Once those analyses are completed and their results are submitted for publication, the data will be made publicly available. In the meantime, they are available upon request.

References

Abdel Rahman, R., & Melinger, A. (2009). Dismissing lexical competition does not make speaking any easier: A rejoinder to Mahon and Caramazza (2009). *Language and Cognitive Processes*, *24*(5), 749–760. http://dx.doi.org/10.1080/01690960802648491

Abdel Rahman, R., & Melinger, A. (2011). The dynamic microstructure of speech production: Semantic interference built on the fly. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 149. https://doi.org/10.1037/a0021208

Arnold, J. E., & Nozari, N. (2017). The effects of utterance timing and stimulation of left prefrontal cortex on the production of referential expressions. *Cognition*, *160*, 127–144. http://dx.doi.org/10.1016/j.cognition.2016.12.008

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. http://dx.doi.org/10.1016/j.jml.2012.11.001

Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology*, *58*(4), 667–692. http://dx.doi.org/10.1080/02724980443000142

Belke, E., & Stielow, A. (2013). Cumulative and non-cumulative semantic interference in object naming: Evidence from blocked and continuous manipulations of semantic context. *The Quarterly Journal of Experimental Psychology*, *66*(11), 2135–2160. http://dx.doi.org/10.1080/17470218.2013.775318

Blackmer, E. R., & Mitton, J.(1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*, 173–194. http://dx.doi.org/10.1016/0010-0277(91)90052-6

Boland, H. T., Hartsuiker, R. J., Pickering, M. J., & Postma, A. (2005). Repairing inappropriately specified utterances: Revision or restart?. *Psychonomic bulletin & review*, *12*(3), 472-477. http://dx.doi.org/10.3758/BF03193790

Broos, W., Duyck, W., & Hartsuiker, R. (in press). Monitoring Speech Production and Comprehension: Where is the Second-Language Delay? *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021818807447

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,* 977–990. http://dx.doi.org/10.3758/BRM.41.4.977

Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In *The child's conception of language* (pp. 17-43). Heidelberg, Germany: Springer. https://doi.org/10.1007/978-3-642-67155-5_2

Damian, M. F., & Als, L. C. (2005). Long-lasting semantic context effects in the spoken production of object names. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1372. https://doi.org/10.1037/0278-7393.31.6.1372

Damian, M. F., & Bowers, J. S. (2003). Locus of semantic interference in picture-word interference tasks. *Psychonomic Bulletin & Review*, *10*(1), 111–117. http://dx.doi.org/10.3758/BF03196474

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283. http://dx.doi.org/10.1037/0033-295X.93.3.283

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., Carreiras, M. (2013). EsPal: One-stop Shopping for Spanish Word Properties. *Behavior Research Methods*, 45, 1246-1258. http://dx.doi.org/10.3758/s13428-013-0326-1

Endrass, T., Reuter, B., & Kathmann, N. (2007). ERP correlates of conscious error recognition: aware and unaware errors in an antisaccade task. *European Journal of Neuroscience*, *26*(6), 1714–1720. http://dx.doi.org/10.1111/j.1460-9568.2007.05785.x

Fodor, J.A. (1983). The modularity of mind. Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/4737.001.0001

Freund, M., & Nozari, N. (2018). Is adaptive control in language production mediated by learning? *Cognition*, *176*, 107–130. https://doi.org/10.1016/j.cognition.2018.03.009

Gauvin, H., & Hartsuiker, R. (under revision). Towards a new model of verbal monitoring. Retrieved from https://osf.io/fraz9/

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649-683. https://doi.org/10.1080/01690960500181332

Hanley, J. R., Cortis, C., Budd, M.-J., & Nozari, N. (2016). Did I say dog or cat? A study of semantic error detection and correction in children. *Journal of Experimental Child Psychology*, *142*, 36–47. https://doi.org/10.1016/j.jecp.2015.09.008

Hanulová, J., Davidson, D. J., & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Psycholinguistic and neurocognitive evidence on second language word production. *Language and Cognitive Processes*, *26*(7), 902–934. http://dx.doi.org/10.1080/01690965.2010.509946

Hartsuiker, R. J., Catchpole, C. M., de Jong, N. H., & Pickering, M. J. (2008). Concurrent processing of words and their replacements during speech. *Cognition*, *108*(3), 601–607. https://doi.org/10.1016/j.cognition.2008.04.005

Hartsuiker, R. J., & Kolk, H. H. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive psychology*, *42*(2), 113-157. https://doi.org/10.1006/cogp.2000.0744

Hartsuiker, R. J. & Moors, A. (2016). On the automaticity of language processing. In: H.J. Schmid (Ed.), Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning (pp. 201-225). Berlin: De Gruyter. http://doi.org/10.1037/15969-010

Hartsuiker, R. J., Pickering, M. J., & De Jong, N. H. (2005). Semantic and phonological context effects in speech error repair. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 921. http://dx.doi.org/10.1037/0278-7393.31.5.921

Hester, R., Foxe, J. J., Molholm, S., Shpaner, M., & Garavan, H. (2005). Neural mechanisms involved in error processing: a comparison of errors made with and without awareness. *Neuroimage*, *27*(3), 602–608. https://doi.org/10.1016/j.neuroimage.2005.04.035

Hester, R., Simoes-Franklin, C., & Garavan, H. (2007). Post-error behavior in active cocaine users: poor awareness of errors in the presence of intact performance adjustments. *Neuropsychopharmacology*, *32*(9), 1974. https://doi.org/10.1038/sj.npp.1301326

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135. https://doi.org/10.1038/nrn3158

James, W. (1890). The principles of psychology. New York: Holt, Rinehart & Winston. http://dx.doi.org/10.1037/11059-000

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446. https://dx.doi.org/10.1016%2Fj.jml.2007.11.007

Karmiloff-Smith, A. (1986). From meta-processes to conscious access: Evidence from children's metalinguistic and repair data. https://doi.org/10.1016/0010-0277(86)90040-5

Kohn, S. E. (1984). The nature of the phonological disorder in conduction aphasia. *Brain and Language*, *23*(1), 97–115. http://dx.doi.org/10.1016/0093-934X(84)90009-9

Kroll, J. F., Bobb, S. C., Misra, M., & Guo, T. (2008). Language selection in bilingual speech: Evidence for inhibitory processes. *Acta Psychologica*, *128*(3), 416–430. https://doi.org/10.1016/j.actpsy.2008.02.001

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). http://dx.doi.org/10.18637/jss.v082.i13

Laver, J. D. (1973). The detection and correction of slips of the tongue. *Speech Errors as Linguistic Evidence*, 132–143. https://doi.org/10.1515/9783110888423.132

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104. http://dx.doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J. M. (1989). Speaking: From intention to articulation. Cambridge, MA: MIT Press. http://doi.org/10.2307/1423219

McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. . *Cognition*, *117*(3), 243-260. http://doi.org/10.1016/j.cognition.2010.08.019

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503. https://doi.org/10.1037/0278-7393.33.3.503

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological bulletin*, *132*(2), 297-326. https://doi.org/10.1037/0033-2909.132.2.297

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. *Psychophysiology*, *38*(5), 752–760. http://dx.doi.org/10.1111/1469-8986.3850752

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. *Errors in Linguistic Performance*. http://doi.org/10.2307/413970

Nooteboom, S. G. (2005). 10 Listening to oneself: Monitoring speech production. *Phonological Encoding and Monitoring in Normal and Pathological Speech*, 167. http://doi.org/ 10.4324/9780203506196

Nooteboom, S., & Quené, H. (2008). Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language*, *58*(3), 837-861. http://dx.doi.org/10.1016/j.jml.2007.05.003

Nooteboom, S. G., & Quené, H. (2013a). Heft lemisphere: Exchanges predominate in segmental speech errors. *Journal of Memory and Language*, *68*(1), 26–38. http://dx.doi.org/10.1016/j.jml.2012.08.004

Nooteboom, S. G., & Quené, H. (2013b). Parallels between self-monitoring for speech errors and identification of the misspoken segments. *Journal of Memory and Language*, *69*(3), 417–428. http://doi.org/ 10.1016/j.jml.2013.04.006

Nooteboom, S. G., & Quené, H. (2015). Word onsets and speech errors. Explaining relative frequencies of segmental substitutions. *Journal of Memory and Language*, *78*, 33–46. http://dx.doi.org/10.1016/j.jml.2014.10.001

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. http://dx.doi.org/10.1016/j.jml.2017.01.007

Nooteboom, S. G., & Quené, H. (2019). Temporal aspects of self-monitoring for speech errors. *Journal of Memory and Language, 105,* 43-59. http://dx.doi.org/10.1016/j.jml.2018.11.002

Notebaert, W., Houtman, F., Van Opstal, F., Gevers, W., Fias, W., & Verguts, T. (2009). Post-error slowing: an orienting account. *Cognition*, *111*(2), 275–279. https://doi.org/10.1016/j.cognition.2009.02.002

Nozari, N. (2019). The dual origin of semantic errors in access deficit: activation vs. inhibition deficit. *Cognitive Neuropsychology*, online March, 2019. https://doi.org/10.1080/02643294.2019.1587397

Nozari, N. (2018). How special is language production? Perspectives from monitoring and control. In K. Federmeier and D. Watson (Eds.), *Psychology of Learning and Motivation: Current Topics in Language* (Vol. 68, pp. 179-23). Cambridge, MA: Academic Press. https://doi.org/10.1016/bs.plm.2018.08.006

Nozari, N., Arnold, J. E., & Thompson-Schill, S. L. (2014). The effects of anodal stimulation of the left prefrontal cortex on sentence production. *Brain Stimulation*, *7*(6), 784–792. http://dx.doi.org/10.1016/j.brs.2014.07.035

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. https://doi.org/10.1016/j.cogpsych.2011.05.001

Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, Cognition and Neuroscience*, *31*(7), 886–903. http://dx.doi.org/10.1080/23273798.2016.1157194

Nozari, N., & Hepner, C. R. (2018). To select or to wait? The importance of criterion setting in debates of competitive lexical selection. *Cognitive Neuropsychology*, 1–15. http://dx.doi.org/10.1080/02643294.2018.1476335

Nozari, N., & Novick, J. (2017). Monitoring and control in language production. *Current Directions in Psychological Science*, *26*(5), 403–410. http://dx.doi.org/10.1177/0963721417702419

Nozari, N., & Omaki, A. (2018). Syntactic production is not independent of inhibitory control: Evidence from agreement attraction errors. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society Austin, TX: Cognitive Science Society*.

Oomen, C. E., Postma, A., & Kolk, H. H. J. (2005). Speech monitoring in aphasia: Error detection and repair behaviour in a patient with Broca's aphasia. In R. J. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), Phonological encoding and monitoring in normal and pathological speech. Hove, UK: Psychology Press.

Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech*, *39*(4), 375–392. https://doi.org/10.1177%2F002383099603900403

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199–227. http://dx.doi.org/10.1016/j.jml.2005.10.002

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2009). Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proceedings of the National Academy of Sciences*, *106*(1), 322–327. https://doi.org/10.1073/pnas.0805874106

Schuchard, J., Middleton, E. L., & Schwartz, M. F. (2017). The timing of spontaneous detection and repair of naming errors in aphasia. *Cortex*, *93*, 79–91. https://doi.org/10.1016/j.cortex.2017.05.008

Shattuck-Hufnagel, S. (1992). The role of word structure in segmental serial ordering. *Cognition*, *42*(1–3), 213–259. http://dx.doi.org/10.1016/0010-0277(92)90044-I

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. http://dx.doi.org/10.1016/j.actpsy.2012.02.004

Tydgat, I., Stevens, M., Hartsuiker, R. J., & Pickering, M. J. (2011). Deciding where to stop speaking. *Journal of Memory and Language*, *64*(4), 359–380. https://doi.org/10.1016/j.jml.2011.02.002

Van Hest, E. (1996). *Self-repair in L1 and L2 production*. Tilburg: Tilburg University Press. https://doi.org/10.1075/itl.117-118.05van

Van Wijk, C., & Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology*, *19*(4), 403–440. http://dx.doi.org/10.1016/0010-0285(87)90014-4

Weighall, A. R. (2008). The kindergarten path effect revisited: Children's use of context in processing structural ambiguities. *Journal of Experimental Child Psychology*, *99*(2), 75–95. https://doi.org/10.1016/j.jecp.2007.10.004

Wessel, J. R., Danielmeier, C., & Ullsperger, M. (2011). Error awareness revisited: accumulation of multimodal evidence from central and autonomic nervous systems. *Journal of Cognitive Neuroscience*, *23*(10), 3021–3036. https://doi.org/10.1162/jocn.2011.21635

Tables

Table 1. The linguistic materials.

Nouns	
Set 1	
English	**Spanish**
bottle	(la) botella
curtain	(la) cortina
window	(la) ventana
suitcase	(la) maleta
Set 2	
telephone	(el) teléfono
package	(el) paquete
mirror	(el) espejo
newspaper	(el) periódico
adjectives	
English	**Spanish**
green	verde
brown	marrón
yellow	amarillo
blue	azul
verbs	
English	**Spanish**
disappears (behind)	desaparecer (por detrás)

pass (behind)	pasar (por detrás)
produce	producir
zigzag (towards)	zigzaguear (hacia)
jump (over)	saltar (por encima)
loop (around)	rodear
bounce (towards)	brincar (hacia)
bump (into)	chocar (con)

Figure Captions

Figure 1. Example of a slide with four events that unfold in the following sequential order: 1. Bouncing towards, 2. Jumping over, 3. Zigzagging towards, and 4. Looping around. The lines show the motion path and the arrows the direction of the movement. The experiment contained 56 slides, half in English and half in Spanish blocks for a total of 224 events.

Figure 2. Proportion of errors (left) and corrections (right) on NP1 (solid light gray bars) and NP2 (polka dot black bars) in English and Spanish. Height of the bars reflects the mean of subject means and the error bars are SEs.

Figure 3. Proportion of errors (left) and corrections (right) on word 1 within the NP (solid black bars) and word 2 (shaded bars) in English and Spanish. Height of the bars reflects the mean of subject means and the error bars are SEs.

Figure 4. Summary of all the data. Proportion of errors (upper panel) and corrected errors (lower panel) for word1 and word2 in NP1 and NP2 in English (black bars) and Spanish (white bars). The figure shows the similarity between the fluctuations of error rates (decrease from word1 to word2, increase from NP1 to NP2) in the two languages, and the similarity between this pattern and the pattern of corrections.