Neuropsychologia 84 (2016) 272–281

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Does naming accuracy improve through self-monitoring of errors?

Myrna F. Schwartz*, Erica L. Middleton, Adelyn Brecher, Maureen Gagliardi, Kelly Garvey

Moss Rehabilitation Research Institute, Elkins Park, PA, USA

ARTICLE INFO

Article history:
Received 18 June 2015
Received in revised form
24 November 2015
Accepted 21 January 2016
Available online 7 February 2016

Keywords:
Speech self-monitoring
Aphasia
Learning
Feedback
Naming
Repair

ABSTRACT

This study examined spontaneous self-monitoring of picture naming in people with aphasia. Of primary interest was whether spontaneous detection or repair of an error constitutes an error signal or other feedback that tunes the production system to the desired outcome. In other words, do acts of monitoring cause adaptive change in the language system? A second possibility, not incompatible with the first, is that monitoring is indicative of an item's representational strength, and strength is a causal factor in language change. Twelve PWA performed a 615-item naming test twice, in separate sessions, without extrinsic feedback. At each timepoint, we scored the first complete response for accuracy and error type and the remainder of the trial for verbalizations consistent with detection (e.g., "no, not that") and successful repair (i.e., correction). Data analysis centered on: (a) how often an item that was misnamed at one timepoint changed to correct at the other timepoint, as a function of monitoring; and (b) how monitoring impacted change scores in the Forward (Time 1 to Time 2) compared to Backward (Time 2 to Time 1) direction. The Strength hypothesis predicts significant effects of monitoring in both directions. The Learning hypothesis predicts greater effects in the Forward direction. These predictions were evaluated for three types of errors – Semantic errors, Phonological errors, and Fragments – using mixed-effects regression modeling with crossed random effects. Support for the Strength hypothesis was found for all three error types. Support for the Learning hypothesis was found for Semantic errors. All effects were due to error *repair*, not error detection. We discuss the theoretical and clinical implications of these novel findings.

1. Introduction

Speech monitoring is a complex cognitive skill that operates largely beneath the surface of awareness. Intuition suggests that auditory comprehension is important for self-monitoring, and evidence shows this to be so. For example, monitoring suffers when healthy speakers are asked to detect their errors in the presence of noise (Lackner and Tuller, 1979; Oomen et al., 2001; Postma and Noordanus, 1996). On the other hand, the linguistic signatures of monitoring – self-interruption, editing terms ("uh-", "no"), and repairs – often happen too rapidly for auditory feedback to have plausibly played a role (Levelt, 1983). Either the comprehension system monitors speech before, as well as after, articulation (inner-speech monitoring; Hartsuiker and Kolk, 2001; Levelt, 1983), or the mechanisms of monitoring are internal to the production system (Nozari et al., 2011; Postma, 2000).

From a functional communication perspective, speech self-monitoring plays an important role in keeping speech errors in check and the dialogue on track (Pickering and Garrod, 2004). We wondered if it might also play a role in use-dependent, incremental language learning (e.g., Damian and Als, 2005; Oppenheim et al., 2010). That is, might error detection or repair constitute an error signal or other feedback that tunes the production system to the desired outcome? We explored this novel hypothesis through an analysis of spontaneously monitored naming errors in participants with aphasia.

1.1. Speech monitoring in aphasia

Generally speaking, people with aphasia (PWA) produce higher than normal rates of error in speech and naming and show less evidence of monitoring (e.g., Schlenck et al., 1987). PWA who routinely fail to monitor their speech errors tend to carry a more severe diagnosis, (e.g., jargon aphasia or Wernicke's aphasia) and have poorer therapy outcomes (Fillingham et al., 2006; Marshall et al., 1994; Marshall and Tompkins, 1982; Wepman, 1958). A link between monitoring and recovery was demonstrated 20 years ago by Marshall and colleagues (Marshall et al., 1994). They studied 30 PWA 1 – 6 months post onset and just prior to a 3-month program of general aphasia therapy. Before and during therapy, the participants performed a 40-item picture naming test, which was scored for spontaneous "self-correction effort" (what we here call

* Corresponding author.
E-mail address: mschwart@einstein.edu (M.F. Schwartz).

http://dx.doi.org/10.1016/j.neuropsychologia.2016.01.027
0028-3932/© 2016 Elsevier Ltd. All rights reserved.

error detection) and "self-correction success" (here, *error repair).* A key finding was that comprehension and production outcomes of therapy correlated positively with the rate at which naming errors were detected, though not with the rate of repair. This important study does not appear to have generated much follow-up. This may be because the findings left open the possibility that monitoring plays no causal role in recovery but simply indexes a language system that is less impaired and thus more likely to recover. Our study addressed that possibility along with the more interesting possibility that acts of monitoring actively promote adaptive change, i.e., learning, in the damaged system.

1.2. *Cognitive accounts of monitoring and monitoring deficits*

The monitoring deficit in aphasia has been investigated in relation to cognitive models of naming. Comprehension-based monitoring models predict a positive correlation between PWA's ability to comprehend and their ability to monitor. Contrary to this, Nickels and Howard (1995) reported that the rate at which PWA detected their phonological naming errors did not correlate with their scores on auditory speech processing measures. They also observed no correlation between phenomena indicative of pre-articulatory monitoring (self-interrupted naming attempts, such as/bi-/for *banana)* and tasks requiring the parsing of inner-speech (e.g., selecting two homophones from a triad of pictures). Several other investigations into the relationship between monitoring and auditory comprehension status in PWA were similarly negative (Marshall et al., 1998, 1985; Nozari et al., 2011; Schlenck et al., 1987; see also Oomen et al. (2001)).

A recent study (Nozari et al., 2011) explored the relationship between error monitoring and *production* abilities in aphasia, with more promising results. It has long been appreciated that some PWA selectively monitor their semantic errors, others their phonological errors (Alajouanine and Lhermitte, 1973; Marshall et al., 1985; Stark, 1988). Nozari et al. analyzed PWA's error monitoring in relation to the semantic-phonological version of Dell's interactive two-step model of naming (Foygel and Dell, 2000; Schwartz et al., 2006). The model postulates that the proximal cause of semantic errors in PWA is heightened conflict between the target and related words, due to lesion-induced weakness in the semantic (s) weights (see Fig. 1). The proximal cause of phonological (nonword) errors is heightened conflict among phonemes, due to weak phonological (p) weights. Nozari et al. (2011) ran a simulation study showing that in the normal (unlesioned) model, conflict at either the word or phoneme level reliably predicted errors at that level. They argued for a general-purpose conflict-monitoring system (Yeung et al., 2004) that reacts to such conflict within the production system by signaling the occurrence of an error.

Now consider aphasia: a basic premise of Dell's model is that lesion-induced weakness in s- or p-weights heightens conflict at that level, resulting in errors. Nozari et al. (2011) hypothesized that a further consequence of heightened conflict at a given level might be to lessen the reliability of the conflict signal, causing error monitoring to suffer. They tested this hypothesis with naming and monitoring data from 29 PWA. First, each individual's naming response distributions (proportions of correct responses and several error types) were entered into the model for the fitting of s- and p-weights for that individual (Foygel and Dell, 2000). Then, the naming data were analyzed trial-by-trial for evidence that an error was detected, and the individual's rate of detection for semantic errors and phonological (nonword) errors was calculated. In the final step, participants' semantic and phonological error detection rates were correlated with their model-fitted s-and p-weights, and, for comparison purposes, with 8 measures of auditory input processing. Positive correlations were found between semantic error detection and strength of the s-weights (r=.59, p=.001) and between phonological error detection and strength of the p-weights (r=.43, p=.021). Reversing the pairings (e.g., semantic error detection with fitted p-weights) yielded negative correlations. Thus, as predicted, lower weights at a particular level of the production system correlated with lower detection rates for errors only at that level. In contrast, detection rates and auditory input measures were uncorrelated across the group and doubly dissociated at the level of individual participants (see also Marshall et al. (1998)).

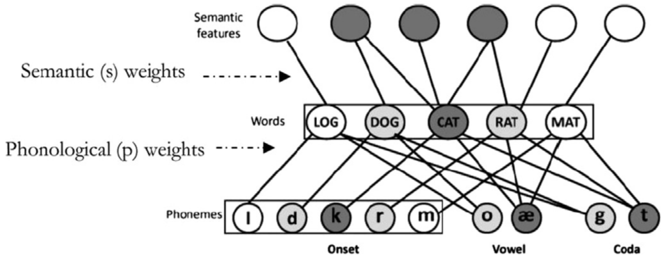

Fig. 1. The interactive two-step model of word production. Boxes indicate the places where conflict was measured during simulated normal naming: the word level at the end of step 1; the phoneme level at the end of step 2 (reprinted from Nozari et al. (2011)).

Nozari et al.'s (2011) correlations and modeling data support a production-based, conflict-centered account of semantic and phonological error detection in aphasia. Their study did not deal with error repair. However, assumptions common to most monitoring accounts – whether comprehension- or production-based – is that repair of speech errors takes place under central control and typically involves restarting or re-programming the speech act (Levelt, 1983; Postma, 2000).[1]

1.3. *The present study*

Like Nozari et al. (2011), the present study analyzed the spontaneous self-monitoring of a heterogeneous group of PWA during performance of a naming task. As noted, our goal was to explore a possible link between monitoring and incremental learning, and to this end, we measured the impact of monitoring *at the level of individual items*. We also analyzed detected and repaired items separate from those detected without repair.

The specific question we asked was whether targets of errors that are spontaneously monitored are at an advantage on retest, relative to unmonitored errors. There are at least two possible reasons why this might be the case. The first is that detecting or repairing an error constitutes a *learning event*, which causally impacts later performance on that item (Learning hypothesis). The second possibility is that detection and repair simply index item strength, and stronger items have a higher probability of being named correctly on another occasion (Strength hypothesis). It is important to note that the Strength and Learning hypotheses are not mutually exclusive; both can be true.

Table 1 shows how the present study went about testing the Strength and Learning hypotheses. In broad outlines, we administered a very large naming test twice, in separate sessions, without extrinsic feedback. On both administrations, each trial was scored for the type of naming response (correct, error) and whether the accompanying verbalizations qualified as evidence of

[1] Postma (2000) suggests that central control is not necessarily involved in the repair of response execution errors, such as those involving faulty placement of the tongue or jaw during articulation. Such errors may be corrected through online peripheral adaptations and need not involve reprogramming. The typical speech error is an error of *selection* (lemma or phoneme), and these, as a class, are repaired under central control.

Table 1
Study design.

Item	Time 1		Time 2	
	Naming resp	Detection	Naming resp	Detection
1	E	No	E	No
2	E	Yes	E	Yes
3	E	No	C	No
4	C	No	C	No
5	C	No	E	Yes
6	E	Yes	C	No
7	C	No	E	No
8	E	Yes	C	No
9	E	Yes	E	No
10	C	No	C	No
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Type of analysis	Variables examined	
	Predictor variable	Outcome variable
Forward (F)	Error detection, time 1	Naming success, time 2
Backward (B)	Error detection, time 2	Naming success, time 1

error detection (yes, no). In the *forward analysis*, we tested whether across subjects and across items, the probability of naming success at Time 2 was predicted by the status of error detection (i.e., detected versus not detected) at Time 1. In the *backward analysis,* we tested whether the probability of success at Time 1 was predicted by the status of error detection at Time 2.

The Strength Hypothesis predicts a relationship between error detection and naming success in both the forward and backward analyses. This prediction is based on two widely accepted assumptions about the naming problem in post-stroke aphasia. The first assumption is that at the item level, naming success is probabilistic rather than deterministic; it is common for PWA to name an item correctly on one occasion and incorrectly on another (Jefferies and Lambon Ralph, 2006). The second assumption is that for a given patient, the relative consistency with which a particular item is named is diagnostic of the strength of that item; given two administrations, items named twice are stronger than those named once, and those named once are stronger than those never named. This assumption is often on display in single-subject treatment experiments, when a baseline naming test is administered multiple times and only items that are misnamed a certain number of times are selected to be treated. The Strength hypothesis goes a step farther and takes detection status into consideration. Consider two items *x* and *y*. Both are named incorrectly, but the error on *x* is detected, whereas the one on *y* is not. The Strength hypothesis takes this as evidence that *x* is stronger and thus more likely to be named correctly on a second occasion. Does it matter whether that second occasion comes after or before the detection measurement? Not according to the Strength hypothesis; it predicts that detected errors will be associated with a higher rate of correct responding at the other timepoint over errors that are not detected, in both the forward and backwards analyses.

Now, the Learning hypothesis posits that detecting an error *alters the strength* of the affected item. As learning can only exert its effects on future behavior, if learning is a factor in addition to the Strength hypothesis, we expect the magnitude of the detection effect will be greater in the forward than in the backward analysis (Learning hypothesis). If the detection effect is significant but equivalent in the forward and backward analyses, this will provide support for the Strength hypothesis only.

The actual experiment was somewhat more complex than this overview suggests. For one thing, we subdivided the errors into a variety of types and tested the Strength and Learning hypotheses in each type, separately. Additionally, inspired by clinical evidence that detection rates and repair rates pattern differently in aphasia (Marshall and Tompkins, 1982; Marshall et al., 1994), we subclassified detected errors for whether or not they were repaired in the same trial; and we tested the Strength and Learning hypotheses in detected errors with and without repair. Finally, and critically, we measured strength and learning effects in errors that were neither detected nor repaired and compared monitoring-related effects to this unmonitored error baseline. This baseline controls for the many factors outside of monitoring that can potentially influence item performance on retest, including familiarity, priming, and, most interestingly, error learning. When a speaker makes a naming error, there is the possibility that the association between picture and errorful response will be strengthened through a process of Hebbian learning (Humphreys et al., 2010; Middleton and Schwartz, 2013), thereby increasing the probability of error on retest. To the extent that error learning is a factor in this experiment, its influence should be captured by scores on the unmonitored-error baseline and neutralized by comparing the experimental conditions to this baseline.

2. Methods

2.1. Participants

Participants were 12 right-handed individuals who had incurred aphasia as a result of left hemisphere stroke. At time of testing (4–109 months post onset) they presented with mild–to-moderate aphasia, good comprehension, and moderate-to-mild naming impairment (Table 2). This was a convenience sample. In a separate study, each had performed a 615-item naming test twice, without feedback, and had achieved accuracy levels between 30% and 70%. The present study used those naming data to investigate how the participants monitored their errors. All data were collected under research protocols approved by the Einstein Healthcare Network Institutional Review Board. Participants gave written consent and were paid for their participation.

2.2. Data collection

The 615-item naming test was created from pictures of common objects collected from published picture corpora (Szekely

Table 2
Participant characteristics.

Participant	Gender	Age	Edu	Mos. Post	Aphasia quotient	Aphasia subtype	Naming % correct
1	M	54	16	4	76	TCM	59
2	F	73	14	15	83	Conduction	65
3	M	48	14	11	65	broca	55
4	M	59	16	81	57	broca	59
5	F	63	12	20	93	Anomic	78
6	F	73	12	24	73	TCM	77
7	M	58	18	109	56	broca	78
8	F	60	10	5	89	Anomic	79
9	M	53	12	61	71	Conduction	70
10	M	62	19	6	85	Anomic	83
11	M	47	14	23	74	Anomic	83
12	M	55	16	10	77	Anomic	64
Mn.		58.8	14.4	30.8	74.8		70.8
Min,Max		47,73	10,19	4,109	56,93		55,83

Notes: Aphasia Quotient and Aphasia Subtype: *Western Aphasia Battery* (Kertesz, 1982); Naming: *Philadelphia Naming Test* (Roach et al., 1996), percentage of the 175 trials named correctly.

Table 3
Characteristics of the 615-Item Picture Corpus.

Variable	Mean (SD)
Name agreement	.93 (.06)
Log frequency/million	1.06 (0.57)
Visual complexity[a]	2.68 (0.76)
# of phonemes	5.20 (1.99)
# of morphemes	1.17 (0.38)
# of syllables	1.84 (0.85)

[a] Five-point scale where 1 = "image is very simple" to 5 = "image is very complex".

et al., 2004) and Internet sources. Properties of the pictures and target names are summarized in Table 3. Values for visual complexity and name agreement (i.e., proportion of trials where the dominant name was produced) were taken from published corpora when available. Otherwise, these values were obtained in normative studies, with a minimum of 40 responses per item. Frequency values for all names were taken from the SUBTLEX$_{US}$ project (Brysbaert and New, 2009).

Each participant performed the naming test twice over the course of two weeks. A single administration of the naming test generally required two sessions in a week, with an average of 1.5 days (SD = 2 days) between sessions. The interval between the last session of the first administration and first session of the second administration was on average 4.3 days (SD = 1.7 days). In each administration, all 615-pictures were presented in random order on a desktop or laptop computer. Naming responses were digitally recorded. Participants were asked to name the pictures as best they could. They were instructed to let the experimenter know when they had given their final answer by pointing to a paper with a "thumbs up" graphic, after which the experimenter advanced the trial. This procedure was instituted in order to avoid experimenter-provided feedback of any kind. If the participant did not indicate they had given their final answer within 20 s, the trial ended automatically. An intervening "Ready?" screen separated trials, and the experimenter initiated each trial when the participant was ready.

2.3. *Response coding*

The procedure resulted in 14,760 naming trials, though 6 trials were removed that were associated with an experimenter error in administration. Trained experts transcribed and checked each response word-by-word from the session audiotaped recordings, including dysfluencies, pauses, tangents, and comments. Fragments and nonwords (i.e., neologisms) were transcribed in IPA. For the purpose of this study, an original coding system was applied to the transcripts. Based loosely on rules developed to score errors (Roach et al., 1996) and spontaneous monitoring (Nozari et al., 2011) on the *Philadelphia Naming Test* (PNT), the present system adopted modifications necessitated by the use of a much larger set of targets with a wider range of name agreement and a sizeable proportion of compound words.

2.3.1. *Coding naming attempts*

We coded the first naming attempt per trial; subsequent responses were considered indicative of monitoring, or were ignored. Trials with no naming attempts were not analyzed further. To qualify as a naming attempt, the response could be an isolated noun, compound noun, or head of a multiword noun phrase. It could also be a nonword or word (of any grammatical category) whose phonological overlap with the target was .50 or higher, according to the formula below (from Lecours and Lhermite (1969)). These naming attempts would later be categorized as phonological errors. We set the overlap criterion to a relatively high .50 in order to minimize the contribution of neologisms that were not true phonological errors but by chance had some small overlap with the target.

$$\text{Phonological overlap} = \frac{\text{\# shared phonemes in target and response} \times 2}{\text{sum of phonemes in target and response}}$$

Each naming attempt was assigned one of 13 naming attempt codes. Four of these account for the preponderance of the data, and we focused the analyses on these:

2.3.1.1. Correct (C). A monomorphemic or compound noun that matched the target, with allowance for incorrect number marking (e.g., Target: *mouse* → Response: "mice"; *cheerleaders* → "cheerleader"). Accompanying modifiers, quantifiers/classifiers, or following prepositional phrase were noted, but their accuracy was not considered and this material was not further analyzed.

2.3.1.2. Semantic errors (Sem). Monomorphemic or compound noun that conveyed a conceptual mismatch in the form of category coordinate (*trumpet* → "tuba"), thematic associate (*pirate* → "treasure"), or incorrect but related superordinate or subordinate (*apple* → "vegetable"; *shoe* → "slipper"). Responses that met these criteria and also resembled the target phonologically (e.g., *lemon* → "lime") were included in this category. Questionable semantic errors, inclusive of near-synonyms (*notebook* → "copybook"), common confusions (*alligator* → "crocodile"), and subordinate responses that correctly characterized the depiction (*flower* → "daffodil") were initially analyzed as a separate category, comprising 450 responses, 3% of the data. In considering whether to treat these as semantic errors or acceptable alternatives, we reasoned that as acceptable alternatives, their rate of error detection should be as low as unambiguously correct responses. In fact, the detection rate for these questionable responses was 10%, 10 times higher than the 1% (false) detection rate for correct responses ($p = .03$). We therefore included them as semantic errors. Correct superordinate responses (*apple* → "fruit") have a similarly ambiguous status. Because these were very rare in the corpus (less than 1% of the data), we excluded them from the analysis.

2.3.1.3. Phonological errors (Phon). A naming attempt with ≥ .50 phonological overlap with the target that did not meet the criteria for semantic error. Phonological errors included nonwords (*banana* → "/**ənæn**/") as well as real-word nouns, adjectives, adverbs or verbs (e.g., *stool* → "sit"; *chair* → "care", *flower* → "/**sauər**/ (sour)".

2.3.1.4. Fragments (Frag). A self-interrupted response comprising minimally a consonant + vowel (CV) or vowel + consonant (VC) sequence. In some cases, the fragment was the only response produced; but more often, it preceded another, complete response (correct or error). To avoid over-inflating the fragment category with simple stuttering or groping behaviors, we required fragments to be *non-repetitious* with the subsequent response (Ex. 1). In cases where the following response repeated the fragment (Ex. 2), the fragment was ignored and the subsequent response was considered the target attempt. All coded fragments were considered errors.

(1) *shoe* → "/**ʃə-**/, slipper".
(2) *goose* → "/**dʌ-**/, duck".

2.3.2. *Coding detection*

Every trial, incorrect or correct, was scored for presence/absence of utterances indicative of detection (maximum detection score per trial = 1). (Because the data were extracted from audiotapes, non-verbal gestures could not be scored.) Any one of the following, occurring after the naming attempt and within the trial, earned a 1 for detection (Note: detection utterances are underlined.)

- Overt rejection or negation of the naming attempt: *shoe* → "slipper, no" (or, "that's not it," "it's not that," "nope", etc.)
- Changed naming attempt: *duck* → "goose, duck"; *shoe* → "boot, slipper"*; banana* → "/bi-/, banana"
- Response elaboration – change through affixation of a bound or free morpheme (other than the plural morpheme): *cheerleaders* → "cheer, cheering" (or "cheering squad", etc.).

Expressions of doubt or uncertainty ("Is that it?") were not considered evidence of detection. Neither was questioning intonation, unfilled or filled pauses, simple repetition, or elaboration through modification (squirrel → "chipmunk, with a bushy tail"). It bears repeating that for purposes of detection coding, accuracy of the naming attempt and appropriateness of the detection behavior were irrelevant. For example, a changed naming attempt counted as detection whether the change was to the correct response or another error.

2.3.3. *Coding repairs*

In the final step, detected errors were classified as *repaired* if the correct response was produced immediately after the error, or as the last of multiple attempts. All other detected errors were classified as *detected without repair.*

2.4. *Reliability*

The coding system was developed through multiple passes through randomly selected subsets of the study data; in each pass, the data were subjected to independent coding by multiple scorers, followed by group resolution and rule refinement. The reliability analyses were conducted on data from 3 participants that had been reserved for this purpose and were therefore unfamiliar to the scorers. In the initial phase of the analysis, 20% of the reserved data (245 naming trials from each of 3 PWA) was assigned one of the 14 possible codes (13 naming attempt codes plus non-naming attempt) by each of 4 trained scorers, working independently. Point-to-point percent agreement was calculated for all scorer pairs on each of the 3 data sets. All discrepancies were reconciled prior to the next phase of the analysis, in which 2 of the scorers independently coded all identified naming attempts (i.e., all trials minus those coded as non-naming attempt) for presence/absence of detection and presence/absence of repair. This analysis was performed on a total of 598 trials (185, 200, and 213 from data sets 1, 2, and 3, respectively); and point-to-point inter-scorer agreement percentages were calculated for each dataset.

2.5. *Data analysis*

The statistical analyses focused on all items that elicited a semantic error, phonological error, or fragment at either administration of the naming test. The dependent variable was change in accuracy (i.e., *change score*) per item, coded as 0 (at the other administration, the participant failed to produce a correct response for that item) or 1 (at the other administration, the item was correctly named). Change score was calculated from Time 1 to Time 2 for items in the forward analysis, and from Time 2 to Time 1 for items in the backward analysis. The data were analyzed using mixed-effects logit regression (Jaeger, 2008; Quene and Van den Bergh, 2008), where the logit (log odds) of change score was modeled as a function of fixed factors and random effects (for an introduction to mixed-effects models, see Baayen et al. (2008)). The regressions were conducted using the glmer function from the lme4 package in R version 3.1.2 (R Core Team., 2014). All models described below included random intercepts for participants and items to capture the correlation among observations that can arise from multiple participants giving responses to the same set of items (i.e., crossed random effects). Random slopes for key design variables entered as fixed effects were included if they improved model fit by chi-square deviance in model log likelihoods (Baayen et al., 2008). However, random slopes entered into any model never changed the statistical significance (alpha = .05) of the design variables.

Table 4
Overall counts (proportion) of trials per response type and naming test administration as a function of Detection and Repair.

			Administration	
Response type	Detection	Repair	Time 1	Time 2
Correct	Detected	Repaired	23(.01)	15(.00)
	Detected	Not repaired	23(.01)	15(.00)
	All Detected		46(.01)	30(.01)
	Not detected	–	3909(.99)	4129(.99)
Semantic error	Detected	Repaired	123(.13)	93(.10)
	Detected	Not repaired	177(.18)	139(.15)
	All Detected		300(.31)	232(.26)
	Not detected	–	659(.69)	669(.74)
Phonological error	Detected	Repaired	94(.20)	90(.20)
	Detected	Not repaired	90(.20)	85(.18)
	All Detected		184(.40)	175(.38)
	Not detected	–	276(.60)	285(.62)
Fragment error	Detected	Repaired	281(.52)	222(.51)
	Detected	Not repaired	193 (.36)	166 (.38)
	All Detected		474 (.88)	388 (.89)
	Not detected	–	63 (.12)	46 (.11)

Table 5
Mean (standard error) change score per error type by direction and detection/repair.

			Direction	
Error type	Detection	Repair	Forward (time 1 to time 2) *M* (*SE*)	Backward (time 2 to time 1) *M* (*SE*)
Semantic	Detected	Repaired	.66 (.05)	.37 (.07)
	Detected	Not repaired	.22 (.05)	.18 (.04)
		Average of Detected	.43 (.04)	.24 (.05)
	Not detected	–	.25 (.03)	.25 (.03)
Phonological	Detected	Repaired	.68 (.06)	.57 (.07)
	Detected	Not repaired	.33 (.06)	.35 (.10)
		Average of Detected	.51 (.05)	.40 (.08)
	Not detected	–	.44 (.07)	.40 (.07)
Fragment	Detected	Repaired	.58 (.05)	.58 (.03)
	Detected	Not repaired	.29 (.04)	.30 (.05)
		Average of Detected	.44 (.03)	.44 (.03)
	Not detected	–	.36 (.10)	.33 (.11)

To evaluate the Strength hypothesis, change score was modeled

Table 6
Mixed Regression Results: Effects of Detection by Error Type and Direction.

				Semantic Errors					
	Forward Direction					Backward Direction			
Fixed Effect	*Coef*	*SE*	*Z*	*p*	Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-1.16	0.17			Intercept	-1.20	0.14		
Effect of Detection					*Effect of Detection*				
Detected[a]	0.73	0.17	4.23	< .001	Detected[a]	0.26	0.20	1.27	.20
Random Effect	s^2				Random Effect	s^2			
Items	0.27				Items	0.79			
Participants	0.22				Participants	0.05			
				Phonological Errors					
	Forward Direction					Backward Direction			
Fixed Effect	*Coef*	*SE*	*Z*	*p*	Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-0.29	0.22			Intercept	-0.58	0.20		
Effect of Detection					*Effect of Detection*				
Detected[a]	0.32	0.23	1.38	.17	Detected[a]	0.29	0.24	1.21	.23
Random Effect	s^2				Random Effect	s^2			
Items	0.18				Items	0.29			
Participants	0.31				Participants	0.24			
				Fragments					
	Forward Direction					Backward Direction			
Fixed Effect	*Coef*	*SE*	*Z*	*p*	Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-0.77	0.56			Intercept	-1.11	0.38		
Effect of Detection					*Effect of Detection*				
Detected[a]	0.53	0.50	1.06	.29	Detected[a]	0.85	0.39	2.18	.03
Random Effect	s^2				Random Effect	s^2			
Items	0.26				Items	0.46			
Participants	2.19				Participants	0.02			
Detection by Participants	1.35								

Note. Excluding the intercepts, *Coef*=model estimation of difference in change score (in log odds) from the reference category for each fixed effect; *SE*=standard error of the estimate; *Z*=Wald Z test statistic; s^2=Random effect variance.

[a] Reference is Not Detected condition.

as a function of a two-level detection factor (levels: Detected; Not Detected), with separate regressions conducted for each of the three error types in each direction (forward, backward). Next, parallel models were constructed except change score was modeled with a three-level factor to separate out the effects of repair from detection (levels: Detected-Not Repaired; Detected-Repaired; Not Detected). The Learning hypothesis was evaluated for each error type by calculating the interaction of direction (forward/backward) and the three-level detection-repair factor. In all models, the Not Detected category provided a baseline for magnitude of performance change between the two naming test administrations that was unrelated to detection or repair.

3. Results

3.1. Reliability

The first phase of the analysis, examining inter-scorer agreement in the assignment of naming attempt codes, yielded the following mean pairwise agreement percentages: Set 1: 88% (range 85–90%); Set 2: 92% (90–95%); Set 3: 92% (91–94%), for an overall agreement score of 91%. In the second phase analysis, point-to-point agreement on detection (yes/no) was 96%, 99%, and 98% and on repair (yes/no), 98%, 100%, and 97%, for the 3 datasets, respectively.

3.2. Descriptive and statistical analyses

Table 4 reports the overall number of observations for the major response types at each test administration as a function of detection/repair category. Detection behavior after correct responses was extremely rare. This is to be expected, as detection in the context of a correct response qualifies as a false alarm, where the participant negated or changed the response even though the response was correct. Detection behaviors were considerably more frequent after errors. Collapsed across repair event and time of test, the rate of detection for semantic errors was .29, for phonological errors, .39, and for fragments, .89.

Table 5 reports mean change score as a function of error type and detection/repair categories in the forward and backward directions. A striking result in this table is that change scores for Detected-Repaired are consistently higher than those for Detected-Not Repaired. For example, fragments errors in the Detected-Repaired category have a mean change score of .58 in both forward and backward directions, whereas in the Detected-Not Repaired category, the means are .29 (forward) and .30 (backward).

The first mixed logit regression analyses collapsed across the Repaired and Not Repaired subcategories to enable the simple comparison of change scores for Detected versus Not Detected categories. Table 6 shows the results for all three error types, in both forward and backward directions. Consistent with the Strength hypothesis, the change score for Detected exceeded the Not Detected baseline in the forward analysis for the semantic errors and in the backward analysis for fragments. None of the

Table 7
Mixed Regression Results: Effects of Detection With/Without Repair by Error Type and Direction.

Semantic Errors									
Forward Direction					Backward Direction				
Fixed Effect	*Coef*	*SE*	*Z*	*p*	Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-1.15	0.16			Intercept	-1.21	0.13		
Effect of Detection With/Without Repair					*Effect of Detection With/Without Repair*				
Detected Repaired [a]	1.73	0.24	7.31	< .001	Detected Repaired[a]	0.92	0.27	3.36	< .001
Detected Not Repaired [a]	-0.19	0.23	-0.84	.40	Detected Not Repaired[a]	-0.29	0.26	-1.09	.28
Random Effect	s^2				Random Effect	s^2			
Items	0.25				Items	0.78			
Participants	0.18				Participants	0.04			
Phonological Errors									
Forward Direction					Backward Direction				
Fixed Effect	*Coef*	*SE*	*Z*	*p*	Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-0.30	0.20			Intercept	-0.60	0.19		
Effect of Detection With/Without Repair					*Effect of Detection With/Without Repair*				
Detected Repaired [a]	.74	0.28	2.65	.008	Detected Repaired[a]	0.76	0.29	2.63	.009
Detected Not Repaired [a]	-0.13	0.28	-0.46	.64	Detected Not Repaired[a]	-0.25	0.31	-0.81	.42
Random Effect	s^2				Random Effect	s^2			
Items	0.13				Items	0.32			
Participants	0.24				Participants	0.20			
Fragments									
Forward Direction					Backward Direction				
Fixed Effect	*Coef*	*SE*	*Z*	*p*	Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-0.61	0.29			Intercept	-1.11	0.37		
Effect of Detection With/Without Repair					*Effect of Detection With/Without Repair*				
Detected Repaired [a]	.94	0.31	3.06	.002	Detected Repaired[a]	1.42	0.40	3.58	< .001
Detected Not Repaired [a]	-0.40	0.32	-1.24	.21	Detected Not Repaired[a]	-0.02	0.40	-0.04	.97
Random Effect	s^2				Random Effect	s^2			
Items	0.11				Items	0.39			
Participants	0.09				Participants	0.00			

Note. Excluding the intercepts, *Coef*=model estimation of difference in change score (in log odds) from the reference category for each fixed effect; *SE*=standard error of the estimate; *Z*=Wald Z test statistic; s^2=Random effect variance.

[a] Reference is Not Detected condition.

other effects in Table 6 were significant.

Follow-up analyses separating effects of detection with and without repair confirmed the importance of repair evidenced in Table 5. As can be seen in Table 7, change scores for Detected-Repaired exceeded those for Not Detected in both directions for all three error types (all $ps < .01$). In contrast, change scores for Detected-Not Repaired tended to be (non-significantly) lower than the Not Detected baseline. These results support the Strength hypothesis, with the important qualification that error monitoring in the form of error correction is an indicator of item strength, not error detection alone.

Table 8 reports results of the interaction analyses that test the Learning hypothesis. For semantic errors, the predicted interaction between detection category and direction was confirmed for the category Detected Repaired (see Detected Repaired/Backward fixed effect under Semantic Errors). With the reference set to the Not Detected condition in the Forward direction, the interaction tested (in log odds) was: (Not Detected/Forward – Not Detected/Backward) – (Detected Repaired/Forward – Detected Repaired/Backward). The interaction coefficient of −0.79 means the second difference was greater than the first. In contrast, the coefficient for the Detected Not Repaired/Backward fixed effect (−0.04) indicates the difference in change score between Detected-Not Repaired and Not Detected baseline was the same in the forward and backward directions. There was no evidence in favor of the Learning hypothesis in the analyses on the other two error types (Table 7 and 8; for detected-repaired/backward and detected-not repaired/backward fixed effects, all $ps > .10$). Fig. 2 depicts the data, in the form of proportions, for semantic errors, phonological errors, and fragments. Only semantic errors exhibited the predicted interaction.

4. Discussion

The central question posed in this study is whether the targets of errors that are spontaneously monitored are at an advantage on retest, relative to unmonitored errors. We considered two not incompatible reasons why monitoring might confer such an advantage. The Strength hypothesis posits that the targets of monitored errors are stronger than the targets of undetected errors. The Learning hypothesis posits that targets of monitored errors are strengthened through learning.

We tested these hypotheses by administering a 615-item naming test twice and measuring how often an item that was misnamed on one occasion changed to a correct response on the other, as a function of the speaker's detection and/or repair of the error. We analyzed data from 12 PWA, focusing on change scores for three error types – semantic errors, phonological errors, and fragments. Mixed-effects regression modeling with crossed random effects showed that for each of the error types, in both the forward and backward direction, detection with repair was associated with higher change scores compared to a baseline of not detected errors. These results constitute strong support for the Strength hypothesis, with the qualification that the monitoring-related index of item strength is repair, not detection. The Learning hypothesis also was confirmed, though only for the Semantic errors. For these errors, the monitoring-associated increase in change scores was greater in the forward than the backward direction. Here, again, the analyses isolated error *repair* as the critical element.

To our knowledge, only one other study has attempted to tease apart effects of error detection versus repair in PWA. As described in the Introduction, Marshall et al. (1994) found that the aspect of

Table 8
Mixed Regression Results: Interaction of Detection With/Without Repair by Error Type and Direction.

Semantic Errors				
Interaction of Direction and Detection With/Without Repair				
Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-1.13	0.16		
Detected Repaired [a]	1.74	0.24	7.10	< .001
Detected Not Repaired [a]	-0.15	0.24	-0.64	.52
Backward [b]	0.03	0.14	0.22	.82
Detected Repaired/Backward [c]	-0.79	0.35	-2.22	.03
Detected Not Repaired/Backward [c]	-0.04	0.34	-0.11	.91
Random Effect	s^2			
Items	0.88			
Participants	0.16			
Phonological Errors				
Interaction of Direction and Detection With/Without Repair				
Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-0.23	0.22		
Detected Repaired [a]	0.87	0.30	2.84	.004
Detected Not Repaired [a]	0.11	0.31	0.36	.72
Backward [b]	-0.26	0.20	-1.29	.20
Detected Repaired/Backward [c]	-0.32	0.41	-0.77	.44
Detected Not Repaired/Backward [c]	-0.51	0.43	-1.17	.24
Random Effect	s^2			
Items	0.87			
Participants	0.26			
Fragments				
Interaction of Direction and Detection With/Without Repair				
Fixed Effect	*Coef*	*SE*	*Z*	*p*
Intercept	-0.62	0.32		
Detected Repaired [a]	1.02	0.34	3.04	.002
Detected Not Repaired [a]	-0.42	0.35	-1.19	.23
Backward [b]	-0.47	0.48	-0.97	.33
Detected Repaired/Backward [c]	0.44	0.53	0.83	.41
Detected Not Repaired/Backward [c]	0.45	0.56	0.81	.42
Random Effect	s^2			
Items	0.61			
Participants	0.10			

Note. Excluding the intercepts, *Coef*=model estimation of difference in change score (in log odds) from the reference category for each fixed effect; *SE*=standard error of the estimate; *Z*=Wald Z test statistic; s^2=Random effect variance.

[a] Reference is Not Detected condition;
[b] Reference is Forward condition;
[c] Reference is Not Detected condition in the Forward direction.

error monitoring that correlated with clinical therapy outcomes was detection and not repair. Of the many methodological differences between that study and the present one, perhaps the most important is the item-level analysis we conducted. Different from the simple correlations that Marshall et al. (1994) reported, our item-level analysis affords confidence that the documented relation between change scores and monitoring is causal, and not mediated by a third clinical variable, such as aphasia severity.

4.1. Error detection

Overall rates of error detection (irrespective of repair) were .29 and .39 for semantic and phonological errors, respectively. (Fragment errors were detected at a much higher rate (.89), due in large part to the way these errors were defined (see Section 2.3.1)). These detection rates for semantic and phonological errors are lower than might have been anticipated, given the relatively mild level of aphasia in our participants. Our participants did not have the severe production and comprehension deficits commonly associated with defective monitoring (e.g., Wernicke's jagon aphasia). The observed detection rates are also low in comparison with past studies. For example, Marshall et al.'s, (1994) study, which included more acute and more severe participants, reported a mean detection rate of .54 (SD .26). The study by Nozari et al. (2011) was more comparable to ours in participant characteristics[2] and scoring methods, yet their participants averaged .65 and .57 detection for semantic and phonological errors, respectively, much higher than the present rates of .29 and .39. The major difference between that study and the present one is the difficulty level of the naming stimuli. Nozari et al. (2011) used the 175-item PNT, which averages 1.37 in log frequency, 4.27 phonemes in length, and .97 name agreement (Kittredge et al., 2008). The test we used was harder by all these measures (see Table 3) and by the inclusion of many compound names. It is reasonable to assume that the greater difficulty of our naming test encouraged the participants to be more conservative in rejecting responses of which they were uncertain. The very low incidence of false alarms (1%) supports this.

We found no evidence that detection alone, unaccompanied by repair, predicted or promoted retest accuracy, relative to errors that go undetected. This might be because detection alone really is of no consequence, or because its impact was too small to be measured by the methods we employed. Our analyses rested on the comparison of monitoring conditions against an undetected-error baseline. This controlled for the many variables apart from monitoring that can influence change scores (e.g., test familiarization, stimulus priming, and error learning). However, the variance contributed by these extraneous variables probably made it harder to discern small effects of monitoring.

4.2. Error repair

Approximately .50 of the phonological and fragment errors that were detected were also repaired (see Table 4). For semantic errors, the repair proportion was around .40. Repair accounted for all positive effects in this study. Repaired errors (relative to the not detected baseline) had significantly higher change scores in both the forward and backward direction, for all three error types. From this, we conclude that as general rule, the targets of repaired errors have greater strength. Additionally, repaired errors of the semantic type exhibited the directional asymmetry (forward > backward) predicted by the Learning hypothesis. This indicates that through the process of repairing semantic errors, their targets become further strengthened through learning. We discuss the strength and learning effects in turn.

4.2.1. Repair as an index of item strength

Within the framework of connectionist naming models (e.g., Fig. 1), item strength often is instantiated in connection weights. Item *a* is stronger (more retrievable) than *b* because *a*'s connections are stronger, presumably as a function of incremental, use-dependent learning. Models that implement frequency-related variation in item strength do so by adjusting weights (e.g., Middleton et al., 2015; Nozari et al., 2010). From this perspective, the assertion that repaired targets are stronger than other errorful targets equates to saying that their connection weights are closer to normal.

We gain further insight into why repair indexes strong weights when we recall that speech error repair is generally thought to involve the restart or reprogramming of computational processes with the production system (Hartsuiker and Kolk, 2001; Levelt, 1983; Postma, 2000). In the case of a naming error, the central

[2] For example, on the *PNT*, fitted s- and p-weights were .027 and .024 in their study, compared to .028 and .022 in ours; average performance on the *Pyramid and Palm Trees* test was 87.6 in their study, 90.8 in ours.

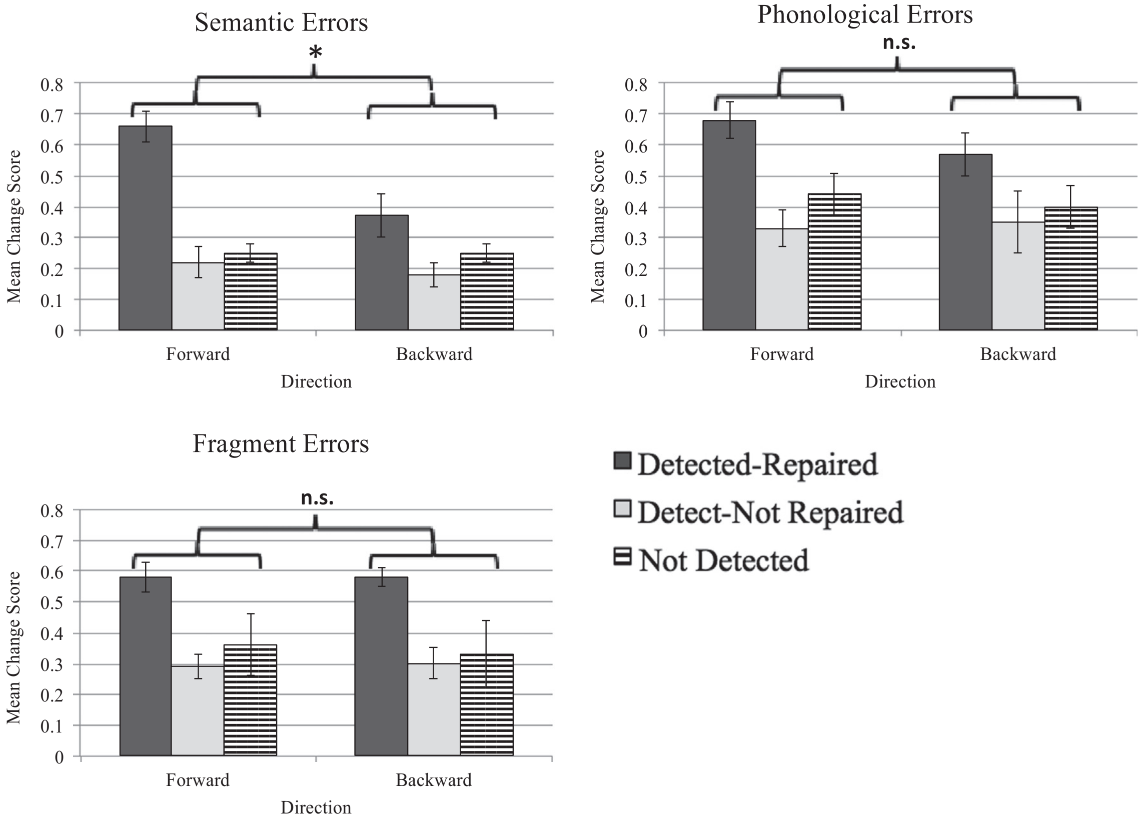

Fig. 2. Proportion mean change score as a function of detection/repair event and direction for each error type.

system might re-initiate the spread of activation from the target's semantic features and allow selection at word and phoneme levels to go forward in the usual (automatic) manner (see Fig. 1). Obviously, this is more likely to return the target if the target has strong connection weights. The implication, then, is that stronger connection weights are both more likely to be successfully recomputed (repaired) after error and more likely to be named correctly on another occasion. In short, stronger connection weights mediate the relationship between repair and higher change scores.

4.2.2. *Repair as a learning event*

The notion that repair involves response recomputation also may explain the learning effect of repair found for semantic errors. In the computational model of incremental lexical learning developed by Oppenheim et al. (2010), s-weights were selectively strengthened and weakened in accordance with an error-based, supervised learning algorithm. Programmed with knowledge of the desired outcome on each trial, the learning algorithm computed the mismatch between the actual and desired response and adjusted the relevant s-weights away from the error and towards the desired response. An intriguing possibility is that in the human production system, the recomputed response instantiates the desired outcome and thus drives the incremental weight change. That would explain why we found learning only in connection with successful repair.

At a more molar level of explanation, the role of spontaneous error correction in lexical learning could function in the way that extrinsic feedback does in other types of verbal learning. It is well known that people learn words, facts, and verbal associations better when, as part of the learning experience, they attempt to retrieve the information and receive feedback when they make an error. This effect seems to require correct-answer feedback; simply informing participants that their answer was right or wrong does not facilitate learning and retrieval compared to a no-feedback condition (Metcalfe and Finn, 2011; Pashler et al., 2005). We found a parallel effect here: lexical learning was not facilitated by error detection manifested as response disavowal, but it was associated with spontaneous retrieval/production of the correct response. This parallel may hint at a common mechanism of action. Specifically, extrinsic and self-generated error correction might provide comparable information to a mechanism that learns by correcting deviations from an ideal or desired state (Oppenheim et al., 2010; Pashler et al., 2005).

4.3. *What is special about semantic errors?*

The evidence we obtained for learning through monitoring is the first of its kind. We had no *a priori* expectation that the effect would occur only in connection with repair and, specifically, repair of semantic errors. Assuming this can be replicated, it will be important to explore whether the restriction to semantic errors has to do with type of representation (semantic vs. phonological), memory system (declarative vs. procedural; e.g., Gupta and Dell, 1999); or monitoring system (comprehension vs. production based). We are currently conducting an analysis of monitoring latencies that may shed light on this issue.

4.4. *Clinical implications*

The verbal learning literature teaches that recall accuracy often underestimates item strength, and item strength is key to long-term learning and re-learning (Kornell et al., 2011). Clinicians implicitly acknowledge the strength-accuracy distinction when they use cueing hierarchies to coax retrieval of words that vary in accessibility. We obtained strong evidence that misnamed items that are self-repaired are stronger than those that are misnamed without detection and repair. Such information may have a useful role to play in aphasia treatment research. For example, experimental investigations of naming treatments often create individualized sets of treatment- (and matched control) items to insure a desired level of difficulty. For this purpose, a large set of naming items is administered two or more times, and errorful items are selected for treatment based on some criterion (e.g., items never named successfully, or those failed at least once). Based on the present results, self-repaired items are more likely to spontaneously switch to correct over the course of the treatment experiment, in which case their inclusion in a treatment set has the potential downside of reducing experimental sensitivity (e.g., where one is comparing two treatment approaches). On the other hand, the greater strength of these self-monitored items may render them more amenable to treatment and more likely to maintain the benefits of treatment over time. That is, such items might be more likely to be within what learning theorists call the range of "desirable difficulty" (Bjork, 1994). Evidence to this effect would bolster confidence in the present findings, while at the same time adding support for the relevance of general learning principles to the treatment of language in aphasia (Middleton et al., 2015; Schwartz et al., 2015).

A treatment manipulation like the one just described could also be implemented in an experimental research design. Not only would this enable analysis at the level of individual items, as the present study did, it would, in theory, at least, allow for matching of items assigned to undetected, detected, and repaired categories. An experiment along these lines might be especially useful for exploring the functional consequences of detection and repair in PWA with more clinically significant monitoring problems than the present participants, such as those with jargon aphasia, for example.

Acknowledgements

This work was supported by National Institutes of Health research Grants (RO1 DC000191 and RO3 DC012426). A portion of this work was presented at the 2014 meeting of the Academy of Aphasia, Miami FL, October 2014. Dr. N. Bonnie Nozari participated in early discussions of this work and conducted a preliminary analysis of the present dataset. Drs. Gary Dell and Daniel Mirman provided helpful input at various stages of the study.

References

Alajouanine, T., Lhermitte, F., 1973. The phonemic and semantic component of jargon aphasia. In: Goodglass, H., Blumstein, S. (Eds.), Psycholinguistic Aspects of Aphasia. Johns Hopkins University Press, Baltimore.

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59 (4), 390–412.

Bjork, R.A., 1994. Memory and metamemory considerations in the training of human beings. In: Metcalfe, J., Shimamura, A. (Eds.), Metacognition: Knowing About Knowing. MIT Press, Cambridge, MA.

Brysbaert, M., New, B., 2009. Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav. Res. Methods Instrum. Comput. 41, 977–990.

Damian, M.F., Als, L.C., 2005. Long-lasting semantic context effects in the spoken production of object names. J. Exp. Psychol.: Learn. Mem. Cogn. 31, 1372–1384.

Fillingham, J.K., Sage, K., Lambon Ralph, M.A., 2006. The treatment of anomia using errorless learning. Neuropsychol. Rehabil. 16 (2), 129–154.

Foygel, D., Dell, G.S., 2000. Models of impaired lexical access in speech production. J. Mem. Lang. 43, 182–216.

Gupta, P., Dell, G.S., 1999. The emergence of language from serial order and procedural memory. In: MacWhinney, B. (ed.) The emergence of language, 28th Carnegie Mellon Symposium on Cognition. Lawrence Erlbaum, Mahwah, NJ.

Hartsuiker, R.J., Kolk, H.H.J., 2001. Error monitoring in speech production: a computational test of the perceptual loop theory. Cognit. Psychol. 42, 113–157.

Humphreys, K.R., Menzies, H., Lake, J.K., 2010. Repeated speech errors: evidence for learning. Cognition 117, 151–165.

Jaeger, T.F., 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. J. Mem. Lang. 59 (4), 434–446.

Jefferies, E., Lambon Ralph, M.A., 2006. Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. Brain 129, 2132–2147.

Kittredge, A.K., Dell, G.S., Verkuilen, J., Schwartz, M.F., 2008. Where is the effect of frequency in word production? Insights from aphasic picture naming errors. Cognitive Neuropsychology 25, 463–492.

Kornell, N., Bjork, R.A., Garcia, M.A., 2011. Why tests appear to prevent forgetting: a distribution-based bifurcation model. J. Mem. Lang. 65, 85–97.

Lackner, J.R., Tuller, B.H., 1979. Role of efference monitoring in the detection of self-produced speech errors. In: Cooper, W.E., Walker, E.C.T. (Eds.), Sentence Processing. Erlbaum, Hillsdale, NJ, pp. 281–294.

Lecours, A., Lhermite, F., 1969. Phonemic paraphasias: linguistic structures and tentative hypothesis. Cortex 5, 193–228.

Levelt, W.J.M., 1983. Monitoring and self-repair in speech. Cognition 14, 41–104.

Marshall, J., Robson, J., Pring, T., Chiat, S., 1998. Why does monitoring fail in jargon aphasia? Comprehension, judgment, and therapy evidence. Brain Lang. 63 (1), 79–107.

Marshall, R.C., Neuburger, S.I., Phillips, D.S., 1994. Verbal self-correction and improvement in treated aphasic clients. Aphasiology 8 (6), 535–547.

Marshall, R.C., Rappaport, B.Z., Garcia-Bunuel, L., 1985. Self-monitoring behavior in a case of severe auditory agnosia with aphasia. Brain Lang. 24, 297–313.

Marshall, R.C., Tompkins, C.A., 1982. Verbal self-correction behaviors of fluent and nonfluent aphasic subjects. Brain Lang. 15, 292–306.

Metcalfe, J., Finn, B., 2011. People's Hypercorrection of High confidence errors: did they know it all along? J. Exp. Psychol.: Learn. Mem. Cogn. 37 (2), 437–448.

Middleton, E.L., Chen, Q., Verkuilen, J., 2015. Friends and foes in the lexicon: homophone naming in aphasia. J. Exp. Psychol.: Learn. Mem. Cogn. 41 (1), 77–94.

Middleton, E.L., Schwartz, M.F., 2013. Learning to fail in aphasia: an investigation of error learning in naming. J. Speech Lang. Hear. Res. 56, 1287–1297.

Middleton, E.L., Schwartz, M.F., Rawson, K.A., andGarvey, K., 2015. Test-enhanced learning versus errorless learning in aphasia rehabilitation: Testing competing psychological principles. Journal of Experimental Psychology: Learning, Memory, and Cognition. 41 (4), 1253–1261.

Nickels, L., Howard, D., 1995. Phonological errors in aphasic naming: comprehension, monitoring and lexicality. Cortex 31 (2), 209–237.

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

Nozari, N., Kittredge, A.K., Dell, G.S., Schwartz, M.F., 2010. Naming and repetition in aphasia: steps, routes, and frequency effects. J. Mem. Lang. 63, 541–559.

Oomen, C.C.E., Postma, A., Kolk, H.H.J., 2001. Prearticulatory and postarticulatory self-monitoring in Broca's aphasia. Cortex 37, 627–641.

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.

Pashler, H., Cepeda, N.J., Wixted, J.T., Rohrer, D., 2005. When does feedback facilitate learning of words? J. Exp. Psychol.: Learn. Mem. Cogn. 31 (1), 3–8.

Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behav. Brain Sci. 27, 169–226.

Postma, A., 2000. Detection of errors during speech production: a review of speech monitoring models. Cognition 77, 97–131.

Postma, A., Noordanus, C., 1996. Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. Lang. Speech 39 (4), 375–392.

Quene, H., Van den Bergh, H., 2008. Examples of mixed-effects modeling with crossed random effects and with binomial data. J. Mem. Lang. 59 (4), 413–425.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, http://www.r-project.org/.

Roach, A., Schwartz, M.F., Martin, N., Grewal, R.S., Brecher, A., 1996. The Philadelphia Naming Test: Scoring and rationale. Clin. Aphasiol. 24, 121–133.

Schlenck, K.-J., Huber, W., Willmes, K., 1987. "Prepairs" and repairs: different monitoring functions in aphasic language production. Brain Lang. 30, 226–244.

Schwartz, M.F., Dell, G.S., Martin, N., Gahl, S., Sobel, P., 2006. A case-series test of the interactive two-step model of lexical access: evidence from picture naming. J. Mem. Lang. 54, 228–264.

Schwartz, M.F., Middleton, E.L., Hamilton, R., 2015. Word retrieval impairment in adult aphasia. In: Bahr, R.H., Silliman, E.R. (Eds.), Routledge Handbook of Communication Disorders. Routledge, New York.

Stark, J., 1988. Aspects of automatic versus controlled processing, minitoring, metalinguistic tasks, and related phenomena in aphasia. In: Dressler, W., Stark, J. (Eds.), Linguistic Analysis of Aphasic Langauge. Springer-Verlag, New York.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Bates, E., 2004. A new on-line resource for psycholinguistic studies. J. Mem. Lang. 51 (2), 247–250.

Wepman, J.M., 1958. The relationship between self-correction and recovery from aphasia. J. Speech Hear. Disord. 23 (3), 302–305.

Yeung, N., Botvinick, M.M., Cohen, J.D., 2004. The neural basis of error detection: conflict monitoring and the error-related negativity. Psychol. Rev. 111 (4), 931–959.