Published in final edited form as:
Cogn Psychol. 2011 August ; 63(1): 1–33. doi:10.1016/j.cogpsych.2011.05.001.

Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production

Nazbanou Nozari[a], **Gary S. Dell**[a], and **Myrna F. Schwartz**[b]

[a] Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Matthews Ave., Urbana, IL 61801, USA

[b] Moss Rehabilitation Research Institute, 50 Township Line Road Elkins Park, PA 19027, USA

Abstract

Despite the existence of speech errors, verbal communication is successful because speakers can detect (and correct) their errors. The standard theory of speech-error detection, the perceptual-loop account, posits that the comprehension system monitors production output for errors. Such a comprehension-based monitor, however, cannot explain the double dissociation between comprehension and error-detection ability observed in the aphasic patients. We propose a new theory of speech-error detection which is instead based on the production process itself. The theory borrows from studies of forced-choice-response tasks the notion that error detection is accomplished by monitoring response conflict via a frontal brain structure, such as the anterior cingulate cortex. We adapt this idea to the two-step model of word production, and test the model-derived predictions on a sample of aphasic patients. Our results show a strong correlation between patients' error-detection ability and the model's characterization of their production skills, and no significant correlation between error detection and comprehension measures, thus supporting a production-based monitor, generally, and the implemented conflict-based monitor in particular. The successful application of the conflict-based theory to error-detection in linguistic, as well as non-linguistic domains points to a domain-general monitoring system.

Keywords

Speech monitoring; Speech errors; Error detection; Aphasia; Computational models

Introduction

The fact that we survive and even thrive in spite of the error-prone nature of our cognitive systems makes the study of error processing crucial to the understanding of human cognition. Most probably, the reason that we function well despite erring often is that we have the ability to detect our own errors and counteract their effects, either by correcting them or by catching them before they cause trouble. It is the error-detection process that we target in this paper.

Correspondence address: Nazbanou Nozari, nnozari2@illinois.edu, 217-244-1294 (phone), 217-333-2922 (fax).

There are two types of errors (Reason, 1990) with distinct properties. The first type, classically labeled as "mistakes", comprises errors that result from the lack of information necessary to arrive at the correct response (e.g. you might call a yellow-jacket a bee if you do not know the difference between the two). In contrast, some errors result from a hasty response or a momentary lapse of attention, rather than lack of knowledge (e.g. when you confuse left and right when giving someone directions to your home). This second type of errors, referred to as "slips", is particularly important, because slips are common, correctable, and in principle preventable.

In the current paper, we address the question of error detection in speech, by adults who are using their native language. These errors are thus most likely to be "slips" in Reason's terminology, although we often use the term "error" to label them. Our goal is two-fold: for one, we review the problems with the existing account of error detection in speech production and propose an alternative account, which does not suffer similar criticisms. To develop our model we use the similarity between speech errors and slips in other tasks, which brings us to the second goal of the paper: to provide support for a central, generic error detection system (e.g. Miltner, Braun, & Coles, 1997) by showing that a central mechanism that had previously been shown to explain error detection in forced-choice response tasks is equally plausible and effective when applied to a natural task such as speech production.

We start by discussing the "perceptual loop" theory (Levelt, 1983, 1989), the most widely-accepted account of speech monitoring. We review the evidence against this theory and introduce a new account, in which conflict (Botvinick, Braver, Carter, Barch, & Cohen, 2001; Yeung, Botvinick, & Cohen, 2004) is used as a signal for error detection. We then implement our theory in the interactive two-step model of word production (Dell & O'Seaghdha, 1991) and establish its predictions with two computational simulations. Next, we assess the explanatory power of our account against the perceptual loop theory by analyzing the error-detection performance of a group of aphasic patients. We conclude by proposing the conflict-based model as the default mechanism of error-detection in language production and discuss the similarities between this account and the monitoring of motor movements, thus pointing to the possibility of a domain-general monitoring mechanism.

The perceptual loop account of monitoring

The perceptual loop theory (Levelt, 1983, 1989) is an elegant account of speech monitoring because it assumes no specialized device or mechanism for error detection. According to this theory, speakers detect errors in their speech by listening to themselves. Error detection then boils down to comprehending that the produced utterance is different from the intended one. Since humans need the comprehension system to process the speech of others anyway, using this system for self-monitoring sounds plausible and is parsimonious. Similar to detecting errors in other people's speech by listening to them, speakers can detect errors in their own speech through an "external channel" (implying that the spoken utterance is processed by the auditory system).

There is some support for the perceptual loop theory's claim that error detection in self and others' speech is similar. Although normal speakers (as well as Broca's aphasia patients) detect more errors in others speech than in their own (Oomen, Postma & Kolk, 2001), the *relative* detectability of different error types is similar in monitoring self and others' speech. An equal proportion of semantically-related errors (e.g. "dog" for "cat"), as well as form-related errors (e.g. "mat" for "cat"), is detected by people when they monitor their own speech (when auditory feedback is not blocked by noise; Postma & Noordanus, 1996) and when they monitor the speech of others, at least under conditions where the intended meaning is known to both the speaker and the listener (Oomen & Postma, 2002; but see Tent

& Clark, 1980). Such similarity in the pattern of error detection can be taken as evidence that a similar mechanism underlies the detection of both self and others' speech errors.

However, the external channel is not sufficient to explain all the empirical findings concerning error detection. An important finding in this regard is the timeline of detection. We show this using the classic example of the corrected error "v-horizontal" (Levelt, 1989), where the production of the word "vertical" is halted as soon as the first phoneme is produced. The latency between the initiation of the erroneous utterance (the onset of "v" in "vertical") and stopping articulation (halting the production of "v-") is reported to be shorter than 150 ms in about 15%, and shorter than 100 ms in about 5% of overt errors (Blackmer & Mitton, 1991). Given that at least 200 ms is needed for word recognition (Marslen-Wilson & Tyler, 1980) and on average 150- 200 ms is required to halt ongoing behavior (Hartsuiker & Kolk, 2001; Logan & Cowan, 1984), such short latencies between the error onset and cut-off (error-to-cutoff time) are incompatible with error detection through the external channel. Therefore a second channel of processing has been proposed within the perceptual loop theory, in which *inner speech* undergoes monitoring. Like the external channel, this "internal channel" is monitored by the comprehension system, with the only difference being the level at which the speech representation is monitored. The representation is thought to be more abstract in the case of inner speech (phonological representations; Wheeldon & Levelt, 1995; Oppenheim & Dell, 2008).

The combination of the internal and external channels provided a simple and plausible account of monitoring, which succeeded in explaining the nature and time course of error detection behavior in normal speakers (but see Oomen & Postma, 2001). Thus, since its introduction, the perceptual loop theory has maintained its status as the leading theory of speech monitoring in spite of a number of alternative accounts (e.g., Laver, 1973; MacKay, 1987, 1992; Schlenck, Huber, & Willmes, 1987). The theory, however, has not gone without criticism. Two of its major assumptions have been called into questions: (1) Do speakers routinely monitor their inner speech even while producing overt speech? (2) Is comprehension really the basis of error detection? In addition, even if the two assumptions hold, there is doubt that the perceptual loop account can explain the pattern of detection and repair in speech rates faster than normal (Oomen & Postma, 2001). We will review these criticisms below and reevaluate the viability of the perceptual loop account of error detection in speech production.

Problems with the perceptual loop account

Do speakers routinely monitor their inner speech while producing overt speech?—There is little doubt that humans are capable of monitoring their inner speech when they are not speaking aloud. The empirical evidence for this claim comes from a variety of tasks, ranging from phoneme monitoring (e.g. judging whether there is /l/ in the name of the pictured object, when subjects see a picture of a nose; Ganushchak & Schiller, 2006, 2008a, 2009; Wheeldon & Levelt, 1995; Wheeldon & Morgan, 2002) to reporting errors when silently reciting tongue-twisters (Oppenheim & Dell, 2008). This however, does not necessarily imply that people often monitor inner speech (or even that they can do so) while overtly articulating. Vigliocco and Hartsuiker (2002) bring up a theoretical problem with simultaneous monitoring of inner and overt speech. Since the inner speech precedes articulation of overt speech by one or a few words (buffering of inner speech), listening to both would be similar to constantly listening to an echo of your voice, which would make comprehension difficult. Moreover, note that monitoring these echoed signals must be performed while attending to the main task, which is the production of speech.

Recently, Huettig and Hartsuiker (2010) directly tested the hypothesis that inner speech is monitored when there is overt production. They registered the eye movements of their

participants while they named a picture (e.g. a heart) in the presence of a phonological competitor (e.g. the word "harp") and two unrelated words. Regardless of whether the subjects monitor their inner speech or their overt utterance during picture naming, the phonological similarity of the competitor to the target is expected to cause temporary fixation on the competitor. The critical point is the timing of this fixation. If inner speech is perceived during overt production, one would expect the competitor picture (harp) to attract more looks than unrelated pictures at an early time window (−50 ms to +55 ms, based on different estimations of the head-start of inner-speech perception on overt word onset; Indefrey & Levelt, 2004; Hartsuiker & Kolk, 2001; Levelt, 1989). But if only overt speech is monitored, the earliest time point at which the participants fixate the phonological competitor would be around 300 ms post-onset of speech (the same as if they heard the word "heart" spoken to them). The data were compatible with the second possibility. Subjects began fixating the phonological competitor 350–500 ms after they started naming the target. There was thus no indication that they monitored their inner speech during overt production.

Is comprehension really the basis of error detection?—For now, let us assume that monitoring inner speech during the course of overt production is in fact possible and speakers routinely do it. The perceptual loop theory's second main assumption is that monitoring is carried out by the comprehension system. A corollary to this assumption is that there should be a correlation between the ability to comprehend and the ability to detect errors. Doubts about the existence of such correlation were first raised by Schlenck et al. (1987), and these doubts were supported later by Nickels and Howard (1995; but see Özdemir & Roelofs, 2007), who failed to find a correlation between error-detection performance by aphasic patients and three measures of their input processing abilities: lexical decision tasks, nonword-minimal-pair judgment, and synonym judgment. The lack of correlation held up even when the authors included only those patients who showed intact inner speech processing ability, as defined by good performance on rhyme and homophone judgment tasks.

More support for the dissociation between comprehension and error detection came from studies of individual aphasic patients, some of whom showed poor error detection in spite of good comprehension (Butterworth & Howard, 1987; J. Marshall, Robson, Pring, & Chiat, 1998; Liss, 1998). McNamara, Obler, Au, Durso and Albert (1992) reported a similar finding in Parkinson patients, who missed 75% of their errors (a rate comparable to that of Alzheimer's patients with poor comprehension), in spite of having good comprehension. However, whether a monitoring deficit in spite of good comprehension is problematic for the perceptual loop or not, is not entirely clear. Hartsuiker and Kolk (2001) argue that comprehension might be necessary but not sufficient for error detection. So, patients who have detection problems in spite of good comprehension might have problems with the comparison process required for detection, either in storing the intended and comprehended representations, or in determining if and how they differ.

Interestingly, many patients with good comprehension and poor self-correction detect errors in other people's speech perfectly (e.g. Kinsbourne & Warrington, 1963; Maher, Rothi, & Heilman, 1994; J. Marshall et al., 1998). One account of the dissociation between detecting errors in self and others' speech is that these patients are in denial; they avoid acknowledging their disorder by not reacting to their errors. This view does not seem very convincing for two reasons: first, some patients show cross-modal differences in error detection, meaning that they can detect errors in their writing (e.g., patient RMM; J. Marshall et al., 1998) while they fail to detect errors in their oral speech. There is no compelling reason why patients should be in denial about the problem in only one mode of production. Second, even within the same modality, error detection differs across tasks. Patient CM (J. Marshall et al., 1998) fails to detect most of his errors in picture naming, but

does so perfectly in auditory word repetition. Again, a denial account is inconsistent with this finding.

An alternative explanation for poor self-monitoring in spite of good comprehension is an extension of the capacity limitation account for why normal speakers detect errors more frequently in others' speech compared to their own. It has been proposed that aphasic patients might suffer from greater capacity limitations, so the advantage of monitoring others' speech is even greater for this population. There is evidence to doubt this assumption as well. J. Marshall et al. (1998) asked patient CM (the one who could detect his auditory-repetition, but not his picture-naming errors) to listen to a spoken word, choose the correct picture from a semantic competitor, repeat the name and then judge the accuracy of his response. Contrary to the prediction of the limited-capacity account, he did quite well on this task. It therefore seems that, at least in some patients, the denial and capacity limitation accounts of poor detection can be refuted and, hence, we are left with the conclusion that, although good comprehension is associated with good detection of errors in others' speech (e.g. Kinsbourne & Warrington, 1963), it is less associated with detection of one's own errors.

An even stronger case against the assumption that comprehension is the basis for error detection could be made if good error detection was to be shown in spite of poor comprehension. R. Marshall, Rappaport and Garcia-Bunuel, (1985) report a patient with auditory agnosia, who fails to understand spoken speech, but has better-than-expected ability to detect her speech errors. Hartsuiker and Kolk (2001) rightly point out that the patient had good reading comprehension, so although the external loop was absent, comprehension through the internal loop could have been responsible for her successful monitoring performance. But recall that monitoring inner speech during overt production is suspect (Vigliocco & Hartsuiker, 2002; Huettig & Hartsuiker, 2010).

Moreover, R. Marshall et al.'s (1985) patient shows differential monitoring ability for different error types. She makes a lot of semantic errors and fails to detect them, while her fewer phonological errors are almost always repaired. This pattern of error detection is the opposite of what would be expected by an internal-channel-only monitor, because studies of error detection under noise-masked condition have shown that the contribution of the external channel to the detection of semantic errors is minimal (Postma & Noordanus, 1996; also see Hartsuiker, Kolk, & Martensen, 2005, for a model of the division of labor between the internal and external monitoring channels). A similar pattern of differential error detection based on the error type was reported in a transcortical sensory aphasic (with poor comprehension and fluent grammatical speech) who corrected all her phonological errors, but failed to acknowledge her semantic errors, even her semantic jargons, which were her dominant type of error (Stark, 1988). Stark proposed that the patient's trouble with detection may in fact arise from trouble with production, rather than trouble with comprehension.

But before these cases can be taken as evidence against a comprehension-based monitor, an alternative explanation must be refuted. It is possible that semantic errors are in fact detected, but the patients' awareness of the difficulty of the process of repair prevents them from bothering with error acknowledgment (Oomen, Postma, & Kolk, 2005). Countering this argument, Stark (1988) could not find any behavior in the patient who did not correct her semantic errors indicating that she had noticed them, and later a more objective approach confirmed that other patients who do not acknowledge their errors truly do not notice them. J. Marshall et al. (1998) compared the percentage of pauses in two jargon aphasics to that of normal speakers. If, in fact, the lack of overt indication of error detection was due to a conscious decision to ignore the error because of the estimated difficulty of the subsequent repair, one would expect to see more hesitations in the speech of patients. The results were

the opposite. The two patients were marginally more fluent than normal speakers and showed no sign of covert monitoring behavior. Considering all of the data from aphasic speakers together, it appears that there is evidence for a dissociation of comprehension from monitoring of self-speech, thus opposing a major assumption of the perceptual loop theory.

To summarize, we reviewed the evidence that questions the two core assumptions of the perceptual loop account, that inner speech can be monitored during overt production, and that error detection abilities are predicted by comprehension abilities. Nevertheless, there is no doubt that monitoring overt speech through comprehension occurs and is useful for a variety of purposes. For one thing, monitoring errors in other people's speech is certainly carried out, at least in part, by means of comprehension. Furthermore, the increased rate of detection of phonological errors in one's own speech when auditory feedback is present (Lackner & Tuller, 1979), also points to a contribution to error detection from comprehension. In the same vein, the role of other monitoring processes, such as that of proprioceptive receptors in calibrating articulation (Postma, 2000), is undeniable. Given these points, it is *not* the goal of this paper to refute the contribution of the perceptual loop to speech monitoring altogether, but to question its role as the primary mechanism for error detection in self-speech.

An alternative class of monitors: the production-based monitors

Note that the main source of failure of the perceptual loop account is its contingency on comprehension. The argument against monitoring through inner speech does not question the existence of inner speech but rather that it can be profitably processed through the comprehension system during overt production. Thus, for a monitoring theory to avoid these problems, it should not rely on the comprehension system for error detection. Examples of such monitors have been proposed (De Smedt & Kempen, 1987; Laver, 1973, 1980; Schlenk et al., 1987; Van Wijk & Kempen, 1987). The main feature of this class, which we refer to as "production-based" monitors [1], is that they view error detection to be mostly independent of comprehension, and instead to rely on the information generated by the production system itself.

Of these, Laver's (1980) model is the most detailed. It assumes multiple independent monitors between the layers of the production system and a final sensory loop, which is the equivalent of the external channel in the perceptual loop account. According to this account, the production process is *held up* while monitoring is taking place (unlike the perceptual loop account, in which the production process and monitoring are carried out in parallel). This theory has been criticized, because the hold-up nature of multiple monitors is bound to interfere with the fluency of speech (Postma, 2000). Moreover, since the production process is put on hold at the time of monitoring, the only errors ever to become overt are the ones missed by all the production-based monitors. Such errors can only be detected by the sensory loop, which takes a few hundred milliseconds to detect them and plan a repair. So the theory fails to explain very short error-to-cutoff intervals in overt errors such as "v-horizontal".

An important objection to multiple specialized monitors -and more generally to any specialized monitor that compares the actual output against the "correct" output- is reduplication of knowledge: "If an editor knows the correct output all along, one wonders why the correct output wasn't generated in the first place." (MacKay, 1987, p. 167). This problem motivated a new idea of monitoring, in which error detection does not consist of a

[1]Levelt has used the term "production-based" in a different sense. He refers to cases where the speaker has access to the intermediate components of the production system. We use the term to refer to cases where the monitor uses the information internal to the production system.

comparison between the actual and the correct response, but is based on the patterns of information flow in the production system. One such proposal is MacKay's Node structure Theory (NST; MacKay, 1987, 1992). A crucial mechanism of error detection in the NST is the detection of new patterns, not previously experienced by the speaker. Given the evidence supporting the speakers' high sensitivity to the statistical patterns of the information received (Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Warker & Dell, 2006), it is quite conceivable that an unfamiliar pattern will raise a red flag in the system. This red flag, in NST's term, is triggering the speaker's awareness, which is proposed as the mechanism for error detection. Although the NST should be credited for pioneering the idea that detection of a speech error is possible without the system already having the correct response at its disposal, it assumes that nodes are shared between the production and the comprehension systems, and thus suffers the criticism of the dissociability of comprehension and error detection abilities, brought up earlier against the perceptual-loop account.

Another mechanism for a monitor that uses the patterns of information flow in production is a comparison between the amount of activation a node sends to its connected nodes and the amount of feedback it receives (Postma & Kolk, 1993). Obviously, this proposal hinges on the interactive nature of the production system. In an interactive network, such as Dell's (1986) model, when the word node "cat" becomes activated, it sends activation to phonemes /k/, /æ/ and /t/. By virtue of feedback, these phonemes send activation back to the word node "cat". Now, if by some mistake, instead of the onset /k/ another onset (e.g. /b/) becomes activated, the amount of feedback that the word node "cat" receives will be lower than the amount it would have received if the correct phoneme had been activated. In addition, a different node (e.g. "bat") receives feedback without having had much activation previously. It is possible that a monitor would use such discrepancies as a signal that an error must have occurred. Finally, it has been proposed that error detection might be achieved by a competition detector, which measures the total activation of the nodes in a pool and generates a signal if the total activation exceeds a preset threshold (Mattson & Baars, 1992; Schade & Laubenstein, 1993). Some have likened this theory to conflict theories of error detection in action monitoring (Huettig & Hartsuiker, 2010).

Although production-based monitors have promise, they have not been able to supplant the perceptual loop theory as the most widely-accepted account of monitoring, because they either have not been laid out -and tested- in sufficient detail (e.g. Postma & Kolk's proposed model), or have empirical findings contradicting their assumptions (e.g. MacKay's NST or Laver's monitor). Below, we propose a new monitoring mechanism, production-based in nature, that avoids the problems with the comprehension-based monitors such as the perceptual loop, but is detailed enough in its assumptions and predictions to be testable.

The conflict-based account of error detection

Conflict can, in principle, arise at any point during the performance and aftermath of an information processing task. We target a specific type of conflict in this paper which emerges from the competition of several alternatives at the time of response selection, and propose that this conflict can be a basis for speech error detection. The notion of conflict monitoring in service of error detection is not new. It was originally proposed that conflict could be used as a signal to recruit and regulate cognitive control (Botvinick et al., 2001). If control is defined as an interaction between an executive system and a subordinate system (Logan & Cowan, 1984), conflict within the subordinate system (e.g. motor system) can be used as a signal for the executive system to increase the amount of control and subsequently resolve conflict. The evidence leading to this proposal was provided by neuroimaging data on a medial frontal brain structure, the Anterior Cingulate Cortex (ACC). Although the ACC shows activity during a wide range of tasks, varying from visual and motor processing to language and memory, it has been argued that the majority of tasks in which the ACC's

activity has been documented can be categorized in three groups: overriding a prepotent response (e.g. naming the ink color of the word RED printed in green), generating a response from a host of equally possible responses (e.g. completing the stem ST- with any word that comes to mind) or during commission of errors (See Botvinick et al., 2001, for a review). The common element in these categories is argued to be high amount of conflict at the response level. Why there is response conflict in the first two categories is obvious. In Stroop-like tasks, the natural tendency is to generate the prepotent response, but the correct response is in fact the subordinate one, leading to conflict. In stem-completion tasks, no one response is salient enough, so in order to make a single-word response, the subject must suppress a number of equally strong alternatives and resolve the conflict. But the most interesting of all is the ACC's activity during error commission. Does error commission also indicate response conflict?

Botvinick, Nystrom, Fissell, Carter, and Cohen (1999) tested specifically for the correlation between error commission and conflict by measuring the ACC's activity using a variation of the Eriksen flanker task. In this task, targets were arrows pointing left or right with the direction defining the binary response to be made. Flankers consisted of surrounding arrows that pointed in either the same or opposite direction of the central target. The functional magnetic resonance imaging (fMRI) findings showed that this brain region was activated on both error trials and correct trials with high conflict (e.g. ≪ > ≪; see also Carter, Braver, Barch, Botvinick, Noll, & Cohen, 1998), thus hinting at a possible relationship between high conflict and error commission.

Another source of valuable information about error processing has been electrophysiological studies (ERP). Below, we will review the findings concerning the ERP component, called the Error Related Negativity (ERN), with the aim of illuminating certain properties necessary for a model of monitoring. ERN is a negative deflection in the ERP, the onset of which coincides with (Holroyd & Coles, 2002) or precedes (Gehring, Goss, Coles, Meyer, & Donchin, 1993) the onset of the EMG activity giving rise to the erroneous response in speeded tasks, and the peak of which is observed about 80–100 ms after the response (e.g. Dehaene, Posner, & Tucker, 1994; Falkenstein, Hohnsbein, Hoormann, & Blanke, 1990; Gehring et al., 1993). The ERN's origin has been traced to the frontocentral regions, particularly the ACC and SMA (Gehring et al., 1993) or more precisely to the inferior ACC (Dehaene et al., 1994), and has been corroborated by MEG findings (Miltner, Lemke, Weiss, Holroyd, Schevers, & Coles, 2003).

There is strong evidence to suggest that the ERN reflects the operation of a generic error-detection system. First, it has been shown to be independent of the modality of error commission. Errors committed not only with hands, but also with feet (Holroyd, Dien, & Coles, 1998), eyes (Nieuwenhuis, Ridderinkhof, Blow, Band, & Kok, 2001; Van't Ent & Apkarian, 1999), and voice (Masaki, Tanaka, Takasawa, & Yamazaki, 2001) all elicit ERNs. Moreover, this negativity has been traced back to the same brain region regardless of whether the errors were committed by hand or by foot (Holroyd et al., 1998). In addition, ERP studies have provided evidence for a central detection-correction loop, involving the frontal structures. Some of these studies have shown the correlation between the ERN amplitude and a variety of post-error adjusting behaviors (e.g., Debener, Ullsperger, Siegel, Fiehler, von Cramon, & Engel, 2005; Dehaene et al., 1994; Gehring et al., 1993) implying that the ERN-generating center is involved in a cycle of error detection and cognitive regulation to avoid further errors (but see Gehring & Fenscik, 2001; Núñez-Castellar, Kühn, Fias, & Notebaert, 2010; van Meel, Heslenfeld, Oosterlaan, & Sergeant 2007). Other studies have focused on the pathologies of frontal structures, showing that patients with damage to the ACC have error rates comparable to normal subjects, but generate no ERNs (Stemmer et al., 2004; Swick & Turken, 2002). Abnormal ERNs are also well-documented in obsessive-

compulsive adults (Endrass, Schuermann, Kaufmann, Spielberg, Kniesche, & Kathmann, 2010; Gehring, Himle, & Nisenson, 2000; Hajcak & Simons, 2002; Ruchsow, Gron, Reuter, Spitzer, Hermle, & Kiefer, 2005) and children (Santesso, Segalowitz, & Schmidt, 2006) as well as adults with the Gilles de la Tourette syndrome (Johannes, Wieringa, Müller-Vahl, Dengler, & Münte, 2002). These abnormalities have been ascribed to the malfunction of a central error-processing system involving the ACC and the basal ganglia (Devinsky, Morrell, & Vogt, 1995; Holroyd, Nieuwenhuis, Mars, & Coles, 2004). Together, these pieces of evidence point to a central loop which is involved in the detection and correction of errors, some part of which generates the ERN.

In addition to suggesting a central monitoring mechanism, research findings on the ERN have defined specific properties for the monitor. Of particular importance among these findings is the independence of the ERN from awareness. The very early onset of this negativity with regard to the onset of the motor response suggests that it is unlikely to result from conscious processing of the response. Also, direct empirical evidence shows that the magnitude of the ERN is unaffected by whether participants are aware of the error (Endrass, Franke, & Kathmann, 2005; O'Connell, Dockree, Bellgrove, Kelly, Hester, Garavan et al., 2007; Nieuwenhuis et al., 2001; but see Steinhauser & Yeung, 2010). In fact, many errors which were declared "unperceived" by the participants were followed by corrections (Nieuwenhuis et al. 2001; Postma, 2000; Ullsperger & von Cramon, 2006), sometimes of significantly shorter latencies compared to those corrections made to "perceived" errors (Nieuwenhuis et al. 2001). The seemingly automatic nature of error detection (and at least some corrections), together with the neuroimaging findings showing the ACC's activity during both error commission and high-conflict situations, naturally led to the proposal that monitoring for conflict might be the basis of error detection (Yeung et al., 2004). The theory was further corroborated by the presence of a pre-response ERN-analogue (N2) on correct but high-conflict trials (Nieuwenhuis, Yeung, van den Wildenberg, & Ridderinkhof, 2003; Pritchard, Shappell, & Brandt, 1991; Smith, Smith, Provost, & Heathcote, 2010; Van Veen & Carter, 2002; Yeung et al., 2004; Yeung & Nieuwenhuis, 2009; but see Holroyd & Coles for a different view on the ERN).

Although there are relatively few studies of the ERN in language, the findings clearly show that the negativity is present during uncertain or errorful language processing. Sebastián-Gallés, Rodríguez-Fornells, De Diego-Balaquer, and Díaz (2006) showed that Catalan-dominant (but not Spanish-dominant) Spanish-Catalan bilinguals generated ERNs when they made a mistake in a lexical decision task involving Catalan word/nonwords. On the production side, Ganushchak and Schiller (2006, 2008a, 2009) used a button-push go-nogo version of the phoneme monitoring task to investigate the ERN in verbal self-monitoring. The task requires the participants to monitor for a certain phoneme (e.g. /l/) in the name of an object the picture of which (e.g. a nose) is presented on the screen. Participants are to push a button only if the picture name contains the specific phoneme. With this task, Ganushchak and Schiller (2006) demonstrated ERNs on error trials whose amplitude decreased under severe time pressure. This result agrees with that of Gehring et al. (1993) who found smaller ERNs when participants were asked to sacrifice accuracy for speed in the Flanker task. Ganushchak and Schiller (2008a) tested phoneme monitoring in the presence of semantic (e.g., "ear" for the picture of a nose) or neutral (semantically irrelevant) distractors. They found the largest-amplitude ERNs following errors that occurred after semantic distractors. The authors interpreted these result in terms of a monitor that is not only sensitive to conflict at the lower-level motor representations of the response, but also to the more abstract steps of lexical access, such as semantic processing. Finally, Ganushchak and Schiller (2009) demonstrated that German-Dutch bilinguals performing the phoneme monitoring task in Dutch showed an ERN on error trials, just like the Dutch speakers did.

However, they showed higher-amplitude ERNs under time pressure compared to the control condition, the opposite of the finding in the Dutch speakers.

The ERN has also been found following verbal responses. Masaki et al (2001) were the first to show this negativity following spoken responses in a Stroop task. Later, Möller, Jansma, Rodríguez-Fornells, and Münte (2007) showed negativities with a frontocentral distribution on speech errors arising from the SLIP task (Baars, Motley, & MacKay, 1975). Ganushchak and Schiller (2008b) collected ERN-like components in a semantic-blocking paradigm, in which participants had to name pictures in semantically-related or unrelated blocks. In addition, they replicated with spoken responses the button-push finding that the amplitude of the ERN increases when monetary reward is offered for higher accuracy (Gehring et al., 1993; Hajcak, Moser, Yeung, & Simons, 2005; Pailing & Segalowitz, 2004), thus adding to the growing body of evidence that the ERN reflects a error monitoring system that applies regardless of the specific domain that generated the error. Furthermore, they showed larger ERNs in the semantically-related compared to the semantically-unrelated blocks, which they attributed to the higher amount of response conflict in the semantically-related block.

Most recently, Riès, Janssen, Dufau, Alario, and Burle (2011) asked participants to determine the grammatical gender of the name of a pictured object, and in a second experiment, to name the object itself. Both experiments found a negativity, with the time-course and scalp distribution previously described for the ERN, not only on the incorrect, but also on the correct trials. However, the amplitude of the negativity was larger for the incorrect trials. The authors argued that the presence of the negative potential on both correct and incorrect trials in speech production, in agreement with the findings in the visuomotor literature, points towards an online monitor that is shared, at least in part, between the speech-production system and the other cognitive systems.

To summarize, converging evidence from linguistic and non-linguistic studies, using neuroimaging and electrophysiological measures, supports the idea that a generic monitoring system is in place, and that this system consists of a frontal brain region (most likely the ACC) which might use response conflict as a signal for error detection. An example of such a monitoring model has been detailed and shown to explain the data in speeded forced-choice tasks (Yeung et al., 2004). We propose a speech monitoring model, in which, similar to Yeung et al.'s model, measuring conflict between response options and relaying that conflict to an executive center is the basis for error detection. Although the model builds on this domain-general approach to monitoring, it is not simply an application of Yeung et al.'s model to speech. Our model's properties have been tailored to the task of unprompted error-detection in natural speech production, which is quite different from the forced-choice laboratory tasks. Moreover, the model is based on a previously existing model of lexical access in production and this further constrains how conflict is defined and used. In the simulation section below, we outline the principles of our proposed monitor and its predictions. Using an implemented production model not only helps us explain the details of the theory in a precise fashion, but also allows us to generate concrete and testable predictions, which will guide our patient-data analysis.

Model simulations

We used the interactive two-step model of word production (Figure 1) in which semantic features in a semantic layer are connected to *lemmas* (Kempen & Huijbers, 1983) in a word layer and those lemmas, in turn, are connected to their relevant phonemes. The strength of the connections between the semantic and the word layer is defined by the *s* (semantic) weight and the strength of the connections between the word and the phoneme layer by the *p* (phonological) weight. The values of these parameters determine how strongly the

information is transferred from one layer to another, and therefore, determine how well the system functions. In order to simulate a damaged system (such as an aphasic production system), the value of *s* or *p* or both weights is decreased (see Foygel & Dell, 2000, for the details of this lesioning).

This model names an object in two steps. Imagine the target word is “cat”. In the first step, the semantic features pertaining to cat become activated. Activation spreads in the network, activating the lemma “cat”, but also its competitors in the word layers, such as “dog”. Each node’s activation is determined linearly from the sum of activations received from the nodes connected to it and is subject to decay and random noise. Cascading in the model allows for further spread of activation down to the phoneme layer, and the interactive nature of the model causes nodes in the higher layers to receive feedback from the nodes in the lower layers. After 8 time-steps the most active node in the word layer is selected. When describing the simulations, we will call this selected node the “word-layer response” to describe the outcome of the first step of the process of lexical retrieval, although no overt response is made at this stage.

The second step starts by giving a jolt of activation (100 units) to the word-layer response (e.g., “cat”), creating a non-linearity in the process of mapping semantics to phonology. Activation spreads for another 8 time-steps, at the end of which the most active node in each phoneme cluster (e.g., /k/ in onset, /æ/ in vowel and /t/ in coda) is selected and the model’s “final response” is generated by combining these phonemes.

The “word-layer response” can be correct (e.g. “cat” for “cat”), or a semantic (e.g. “dog” for “cat”), formal (e.g. “cap” for “cat”) or unrelated-word (e.g. “fog” for “cat”) error. In a normal speaker, nearly all responses at this layer are correct. If an error is made, it is almost exclusively of semantic type, because the error receives activation from the semantic features that it shares with the target. Most semantic errors in the model are in fact first-step errors. The final response of the model can belong to all the above categories but also to a unique category of nonword errors (e.g. “lat” for “cat”). The reason that nonwords are limited to the second step of lexical retrieval is that whichever node is selected in the first step will be, by definition, a word. Therefore, it is only through an error in retrieving the phonemes (i.e. the second step) that a nonword can be created.

If detecting conflict is useful for error detection in a language production system, that system should exhibit three principles:

1. *Detection Sensitivity*. The amount of conflict must be predictive of the probability of error occurrence. In other words, the distribution of correct and incorrect responses must be distinguishable by a quantified measure of conflict.
2. *Layer Specificity.* Conflict at each layer of the system should specifically predict the error type arising from that step. Recall that semantic errors are common first-step errors, while nonword errors are second-step errors. Therefore, high conflict at the word layer should be predictive of the occurrence of a semantic (but not a nonword) error. Likewise, high conflict at the phoneme layer must signal a nonword (but not a semantic) error.
3. *Integrity Contingency.* Finally, the conflict-based monitor is a production-based monitor, meaning that it relies on the information generated in the process of production to detect errors. Thus, the reliability of conflict between the model’s nodes as an error signal should be directly related to the strength of the production weights. When these weights are strong, correct trials will generally be associated with low conflict. So when conflict is high (due to noise factors), the system can use this information as a signal that something must have gone wrong. On the other

hand, when the weights are weak, transmission of information between layers suffers, with noise playing a large role and thus creating conflict on all trials, regardless of whether they end in the correct response or not. In this case, conflict should no longer be a useful signal to discern error trials. Thus, the more degraded the system, the less reliable the error signal.

In the simulations that follow, we show that the model exhibits detection sensitivity and layer specificity in a simulation of a normal speaker. To investigate integrity contingency in the model, we simulate 5 aphasic patients, and demonstrate how gradual decrease of the *s* and *p* weights affects the reliability of the error signal derived from the amount of conflict between the competing nodes.

Simulation I

In the two-step model of word production the *s* and *p* parameters were both set to 0.04 to simulate picture naming performance of a normal individual. Using these parameters, when the model is run through 10000 trials, the final response is correct in about 98% of the times. The other 2% comprises mostly semantic errors, with few nonword and formal errors (These results match data from control subjects doing a picture naming task with high name-agreement pictures; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997). For each of the 10000 trials, the following information was registered: word-layer response, conflict at the word layer (see below for how conflict was measured), final response and conflict at the phoneme layer. For the latter, we measured conflict among the model's six possible onset phonemes.

Conflict was quantified using two measures: The first measure was simply the difference between the two nodes with the highest levels of activation in a layer, which we call the difference between the maximums or *diff(max).* For example, if activations in the word layer looked like cat = 0.099, dog = 0.018, hat = 0.008, mat = 0.006 and fog = 0.008, *diff(max)* = 0.099–0.018= 0.081. The second measure was the standard deviation *(sd)* of the activation of all nodes in the layer at which conflict was to be determined. In the above example, $sd=\sqrt{\sum_i^n \frac{(x_i-\bar{x})^2}{n-1}}$, where x_i is the activation of each node in the layer, $\bar{x}$, is the mean and is the total number of nodes in that layer, thus *sd* = 0.04. Note that both measures determine the difficulty with which a "winner" is selected. The first measure deals with only the strongest competitor, while the second measure takes into account competition from *all* the competitors. In both cases, the measures were converted into *−ln(diff(max)) and −ln(sd),* because the transformed measures are easier to interpret (higher values represent more conflict) and they generate distributions with less skew. The pattern of the results obtained by using the *−ln*(*diff(max)*) vs. the *−ln*(*sd)* measure was similar, but the *−ln*(*sd)* was a noisier measure of conflict. Ultimately, though, the choice of the conflict measure is an empirical question (see Botvinick et al., 2001 for a full discussion of this issue).

10000 values (for the 10000 trials) were gathered for each measure of conflict at each layer. These values were then categorized based on whether they belonged to a correct or an incorrect trial (determined by the model's final response) and for each measure, separate distributions were built for correct and incorrect responses. If conflict is a useful measure for error detection, the distributions of correct and incorrect responses that are built using the measures of conflict should have little overlap. If not, these distributions should not be easily distinguishable (Figure 2). To determine the overlap of the two distributions we calculated the Cohen's $d=\frac{m_e-m_c}{S_p}$, where m_e is the mean of the distribution of the conflict measure for error responses, m_c is the mean of the distribution of conflict measure for

correct responses and S_p is the pooled standard deviation of the two distributions calculated as $S_p = \sqrt{\frac{(n_c - 1)s_c^2 + (n_e - 1)s_e^2}{n_c + n_e}}$, where the s_c and s_e are standard deviations of the correct and error distributions respectively, and n_c and n_e are the number of trials in those distributions (Hartung, Knapp, & Sinha, 2008). A larger Cohen's d means less overlap, which means the distribution of the two response types are well-differentiated by the conflict-measure (detection sensitivity; see Figure 2, left panels). If the distribution of correct and error responses are very similar with regard to the amount of conflict, the Cohen's d will be close to zero (Figure 2, right panels). When discussing the Cohen's d's, we report the values based on the *−ln(diff(max))* measure of conflict (which proved to be the more sensitive measure) and put the *−ln(sd)*-based values in parentheses. To check the reliability of our estimates of Cohen's d, we calculated the 95% CIs. The large number of trials led to confidence intervals so narrow that in most cases the upper and lower bounds were the same as the reported value when rounded up to two decimals. For this reason, we do not report the confidence intervals for each Cohen's d.

Of the 10000 trials, 209 resulted in incorrect responses and 9791 in correct responses at the phoneme layer (final response). When conflict was measured at the word layer, Cohen's d was 3.11 (1.15 based on the *−ln(sd)*), which indicates good discriminibility, particularly for the *−ln(diff(max))* measure. However, when conflict was measured at the phoneme layer Cohen's d was only 0.24 (0.19 based on the *−ln(sd)*), which indicates low discriminability. Although at first glance this finding looks problematic for the theory, this is in fact exactly what is predicted by the model's second principle: Conflict at each layer must specifically predict the type of error arising from that layer. In the normal speaker simulated here, the majority of errors are semantic (199 semantic errors, 2 formal and 8 nonword errors as the final response pattern). The error distribution therefore, contains mainly first-step errors, which according to layer specificity must be detectable by conflict at the word, and not at the phoneme layer; precisely what the Cohen's d's show.

To test whether this interpretation was correct, we built separate distributions for semantic and nonword errors. Of the 10000 word-layer responses, 199 of them were semantic errors. This number was unchanged in the model's final response, confirming our assumption that semantic errors are dominantly (and in this case, exclusively) first-step errors. Distributions of the two measures of conflict were then generated for correct vs. semantic errors at the word layer. Cohen's d was 3.26 (1.18 based on the *−ln(sd)*) for the detection of semantic errors. The 8 nonword errors in the model, as discussed earlier *must* be second-step errors. When distributions of the conflict measures were built for nonword errors vs. correct responses at the phoneme layer, Cohen's d was 2.83 (1.52 for *−ln(sd))*, much higher than the Cohen's d measured at the phoneme layer, when all errors were lumped together.

As a control, we also did the reverse pairing. The distributions of correct and error responses for the semantic errors were paired with conflict measured at the phoneme layer, and distributions of correct and error responses for nonword errors were paired with conflict measured at the word layer. Recall that this is incorrect pairing of error type and the level at which conflict is measured, and detection sensitivity is expected to be low. Cohen's d was 0.09 (0.11) for the semantic and 0.57 (0.50) for the nonword errors. The low values of the Cohen's d's for the reverse pairing, together with large values when the correct pairing of error type and conflict-layer was made, confirm the principle of layer specificity, which entails a more specific definition of detection sensitivity.

Simulation II

One of the objections raised against the comprehension-based monitor was that some patients with poor comprehension did not have much trouble detecting their own errors. The logic was that if the monitoring device receives its information from the comprehension system, then it should not operate well if the source of information is malfunctioning. The same rationale applies to a production-based monitor. We claim that error detection uses the information (in the case of our model, the amount of conflict) generated by the production system. If the production system does not function properly, the information being used for error detection should be unreliable. Under such circumstances, the speaker either continues to monitor using the unreliable information or stops monitoring altogether. In the former case, the monitor will likely generate many false alarms, while, in the latter, there will necessarily be many misses. In our modeling terms, both of these scenarios would manifest as lower Cohen's d's.

In simulation II we lesioned the model systematically to create patients with different degrees of damage to their *s* and *p* weights. From previous data-fitting studies (Dell et al., 1997; Dell, Martin & Schwartz, 2007; Schwartz, Dell, Martin, Gahl, & Sobel, 2006) we know that a weight value of 0.04 indicates normal production, and values lower than 0.01 indicates severe damage. 0.02 is a middle value, for mild to moderate damage. Six speakers (one normal speaker, as in simulation I and five simulated aphasic patients) were created using permutations of these three values of *s* and *p*. Table 1 shows that of the simulated aphasic patients (second to sixth entries), the first two have one normal weight and one moderately-damaged (but still functioning) weight. The second two have one normal weight and one weight which is severely damaged. The last patient's production system is degraded to the degree that its production is close to random (weights are too low for systematic mapping from one layer to another).

Table 1 also shows the Cohen's d's for each simulated speaker, calculated at the word and phoneme layers, using the amount of conflict measured for semantic and nonword error trials respectively. The numbers show that when layer specificity is taken into account, detection sensitivity is high when the relevant weights are strong. As the weights become weaker, the sensitivity decreases (e.g. a decrease in the value of the *s* weight from 0.04 to 0.02 causes a drop in detection sensitivity from 3.26 to 1.41), until the weight value becomes so low that there is little connectivity between the nodes in the production network. In this case, detection sensitivity becomes so low that the error signal is almost meaningless (e.g., 0.37 when the *s* weight drops to 0.008).

One might wonder why the Cohen's d's for the detection of semantic errors are so different for the normal speaker, the third and the fifth virtual patients in Table 1, despite all their *s* weights being 0.04. The reason is that the relationship between semantic errors generated at the word-layer and semantic errors that emerge as the final response changes as a function of *p* weights. When *p* weights are strong, the numbers of word-layer and final-response semantic errors in the model are very similar because the incorrectly chosen word at the word layer is correctly pronounced, creating a semantic error as the final responses. Simulation of the normal speaker shows 199 semantic errors at the word layer, all of which turned into semantic errors at the phoneme layer. Moreover, no new semantic errors were created during phonological encoding, so there is a one to one relationship between semantic errors at the word layer and in the final response profile of this speaker.

The third patient makes 292 semantic errors at the word level (on 292 trials "dog" is chosen at that level). But the final response profile of the patient contains 275 semantic errors. Of these, only 274 come from the "dog" node, and one is a new error, created by the misselection of the phonemes for the chosen node "cat" at the word layer. So, in this patient,

18 of the word-layer semantic errors are converted to other words during phonological encoding, and one correct word-layer response is turned into a semantic error because of lesioned *p* weights. Thus the relationship between the word-layer and final semantic responses is not perfect like the normal speaker. For the fifth patients, the *p* weights are very small, and the association between the word level and semantic errors is further undermined. The simulations with this patient show 265 word-level, but only 164 final-response, semantic errors, of which only 124 come from the word-layer "dog" node. Therefore, 141 of the word-level semantic errors were converted to other response types, and 20 new semantic errors were created from responses other than "dog" at the word layer.

In summary, our simulations exhibited the three principles deemed necessary for the conflict-based monitor to be a plausible mechanism for error detection: Conflict is a good predictor for the occurrence of an error (detection sensitivity), when measured at the layer from which the error originates (layer specificity). Since this monitoring mechanism relies on the information produced by the speech production system, it functions best when the production system is healthy, and becomes increasingly less accurate as the production system becomes more degraded (integrity contingency).

Throughout the modeling section we have reported Cohen's d as a measure of the usefulness of conflict in signaling that an error is probable. Certainly the high level of discriminability shown by the normal model's values of Cohen's d supports an effective detection mechanism, but ultimately the model must simulate natural error detection in human speakers. Estimates from connected speech corpora place the natural detection rate by a neurologically-healthy adult speaker at around 50% (Nooteboom, 2005; slightly higher in Nooteboom, 1980; Note that there are no comparable data for single-word utterances). By "natural" detection rate, we mean the rate with which errors are followed by corrections or other acknowledgment of error, in the absence of any explicit instruction to detect or report errors.

Why are the real-life hit rates only moderately good, and does the model simulate this? To address these questions we augmented the model with a criterion that governs whether or not a certain degree of conflict indicates an error. Specifically, we used a model of incremental criterion placement in signal detection proposed by Kac (1962), which appears to account for detection data about as well as other models (Thomas, 1973). People learn to set their criteria by making detection responses and getting feedback on their accuracy. In our case, speakers determine whether or not they made a speech error, and then determine whether that assessment was correct or not (presumably by using the external channel for feedback). In Kac's model, the criterion is initially placed in some neutral location. Whenever a signal is not detected, the criterion is lowered by a small amount, and whenever a false alarm occurs, it is raised by a similar amount. No adjustments are made on hits or correct rejections. After many trials, the criterion will hover around a location that reflects a probability matching strategy. The reason that Kac's model yields probability matching is that, at equilibrium, the chance of a criterion increase due to a false alarm must equal the chance of a decrease due to a miss. From this, it follows that the overall probability of a "detect" response (in our case, detect that an error has occurred) will match the overall probability of a signal (in our case, the occurrence of an actual error).

Figure 3 shows the distributions of the $-ln(diff(max))$ measure of conflict for correct responses and semantic errors of the simulated normal speaker (simulation I). We applied Kac's model to these distributions and found the criterion to fall at the conflict level equal to 5.20. When this criterion is applied to the error distribution, the hit rate is 47%. Thus, the combination of the conflict model and the incremental criterion-setting model mimics the fact that normal speakers detect only half of their errors. The reason for the medium hit rate,

in spite of the strong difference in the distributions of error and correct conflict values, is the probability matching feature of the criterion placement algorithm. When errors are uncommon, criterion placement will be conservative, that is, the model will miss detecting many errors, but rarely generate false alarms. It must be noted though, that Kac's model only estimates how a criterion is derived when there are no specific demands influencing criterion placement. Speakers may – and most probably do- change their criteria for self-monitoring under different circumstances. The criterion is most probably lowered (to increase hit rates) under conditions when detection is important.

Just as hit rates were derived by applying the criterion to the error distributions, false alarms can be calculated by applying the criterion to the distribution of correct responses. This renders a 1% false alarm rate for a normal speaker. Although 1% seems like a reasonably low value, given that the number of correct trials is much larger than the error trials, suspicion may raise as to whether it is in fact realistic. Healthy adult speakers rarely have the impression of having erred when they have not. False alarms, however, may manifest as disfluencies. Hesitations or other breaks in the flow of speech are common, with an estimated rate of 6 (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001) to 26 (Fox Tree, 1995) per 100 spoken words. These include overt repairs, disfluencies which clearly reflect error detection and correction, and covert repairs, disfluencies that may reflect monitoring and repair, but without any audible reparandum (Postma & Kolk, 1992). Levelt (1983) categorizes repeats (e.g. "I … I went…") and filled pauses (e.g. "I…uh…went") as covert repairs and brings up the possibility that some of these are in fact false alarms generated by the monitoring system. Bortfeld et al. (2001) report an average rate of 1.47 for repeats and 2.56 for filled pauses per 100 words (with slightly higher overall rates in the older population), which together comprise about 4% of the spoken words. If, a sizable fraction of these are the result of the monitoring system falsely detecting an errors, our simulated 1% false alarm rate is not far off the mark. In the General Discussion we will return to the question of the effectiveness of error detection by arguing that although we propose the conflict-based monitoring as the default natural monitoring mechanism, error detection is in fact boosted by a number of ancillary mechanisms, each of which add to the probability of the success of detection.

We also examined the effects of Kac's (1962) criterion-setting algorithm on the effectiveness of error monitoring through conflict when connection weights are weaker (as in aphasia). These effects can be appreciated by looking at semantic error detection as a function of *s* weights. With *p* weights kept constant at 0.04 (normal), we determined the correct and error distributions of conflict scores and the resulting criteria as the *s* weights dropped from 0.04 (normal) to 0.03, 0.02 and 0.015. The model predicts a decrease in the hit rates from 0.47 (normal) to 0.39, 0.35 and 0.34, accompanied by an increased rate of false alarms from 0.01 (normal) to 0.02, 0.07 and 0.11 for the respective weights. In the section that follows, we investigate the predicted association between hit rates for error detection and production-derived connection weights in a sample of aphasic patients. False alarms, as mentioned above, have many different manifestations in natural speech, which makes quantifying them very difficult, especially in aphasic speech. So, although we do not test specific quantitative predictions about false alarms in error detection in aphasia, we discuss relevant qualitative findings in our sample data.

Patient data

The plausibility of the conflict-based theory was established by simulations showing a certain degree of detection sensitivity, as required under the first principle. The principles of layer specificity and integrity contingency, however, are stronger and lead to predictions that can be empirically tested. According to integrity contingency, the ability to detect errors

must correlate with the indices of the functionality of the production system (parameters *s* and *p* in our model), as opposed to indices of comprehension. Layer specificity makes an even more specific prediction; that the ability to detect semantically-related errors must correlate with the *s* parameter (which determines conflict at the word layer) and the ability to detect phonologically-related (mostly nonword) errors must correlate with the *p* parameter (which determines conflict at the phoneme layer). To this end, we determined the percentage of detected errors in a group of aphasic patients who completed a picture naming task. We then assessed the correlation between the detection of different error types and measures of comprehension and production.

The first step in the patient study was to devise a reliable method for coding natural error detection in patients. For this step, data were obtained from 63 aphasic patients, who participated in the 175-item Philadelphia Naming Test (The PNT; Roach, Schwartz, Martin, Grewal, & Brecher, 1996). In this test, a single picture (a black and white line drawing) appears and remains on the screen until the patient responds or for 30 seconds in case of no response. There are no specific instructions for self-correction in this task, but patients do have time to detect and correct their errors. Therefore, detecting errors is self-initiated, and somewhat reflexive of the natural process of error detection, rather than a task demand. In addition, the single-word-response nature of this task, which eliminates contextual cues, makes it suitable for comparison with our model, which is a model of single-word (and not sentence) production.

Patients' responses (word by word, including disfluencies, pauses, tangents and incomplete responses) were transcribed, once on-line (by a trained expert at the Moss Research Institute during the testing session), and once off-line using the session's recordings. In the next step, the responses were "coded" into correct and a number of error categories. For the purpose of this study, all of the original transcriptions of trials were then recoded by the first author with regard to the nature of the errors made and the patients' detection of those errors. A second coder, a graduate student trained on the coding scheme, also recoded the transcription independently of the first author. The codings were then compared, resulting in an initial agreement of 78%. The coding criteria were consulted in resolving the disagreements until full agreement was achieved.

Coding error categories

For each picture, the response was coded as either correct, incorrect or omission. We assigned the "omission" code to trials in which no response was produced, as well as trials in which the patient gave a relevant description of the item or a multi-word response irrelevant to the description of the picture (e.g., "I can't remember."). A trial was not coded as an omission, though, when such responses were followed by a single-word response in the noun category. For example, if in response to the picture of a cat, the patient said, "I have one of those (pause) dog", we accepted the response after the pause, coding it as a semantic error.

If a response was coded as an error, a category had to be specified. A detailed description of all error categories with examples is available upon request. The two important error categories are semantic and phonological errors. Semantic errors were defined as any nouns related in meaning to the target word (e.g. "dog" for the target "cat"). In agreement with previous coding schemes using the PNT, verbs (e.g., "ride" for the target "bike") did not count as semantic errors, but were registered as omissions. Phonological errors were coded as any responses, word or nonword, which showed a clear phonological similarity to the target, as defined in Schwartz et al. (2006). It is noteworthy that most of the targets in the PNT do not have a rich phonological neighborhood (e.g., tractor, helicopter, pumpkin, etc.), so most of the phonological errors end up being nonwords. Dialectical variations in

pronunciation and phonological distortions due to mild speech apraxia, as determined by the Apraxia Battery for Adults (Dabul, 2000), were not coded as errors.

Fragments that showed a clear phonological similarity to the target were also coded as phonological errors. For example, for the target “thermometer”, if the patient responded “thero-” we counted that as a phonological error. However, one phoneme fragments (e.g. “th-”) were counted as fragments and not as errors. Also, if the fragment had the correct sequence of phonemes but was simply incomplete (e.g. “thermo-”) it did not count as an error. As a policy, in keeping with previous applications of PNT data, we coded fragments in a conservative manner, meaning that where the two coders did not agree that a fragment showed clear similarity to the target, it was not coded as an error.

Coding error detection

An error was coded as “detected” if the patient gave any indication of response rejection. This included a repair attempt, regardless of whether the repair was successful or not (e.g. “dog… cat” or “dog…cow” in response to the target “cat”) or simply rejection of the response (e.g., “dog… no…”). We did not code repetition of the same response as detection, due to the uncertain nature of repetitions (Fraundorf, & Watson, 2008; Levelt, 1983; MacKay, 1976; Postma & Kolk, 1993).

Other measures

In addition to error detection, a number of other measures were registered for each patient. These consisted of two production and three comprehension measures. For production, the strength of the *s* and *p* weights for each patient was determined by fitting the interactive two-step model to the patient’s naming data. Briefly, this entails a search of the *s* and *p* values so that the model maximizes its fit (minimizing Chi-square) to the patient naming response proportions in the correct category, as well as five error categories: semantic, phonologically-related word (formal), mixed semantic and formal, unrelated word, and nonword errors (See Dell, Lawler, Harris, & Gordon, 2004 for the details of the relevant coding and fitting process). After the model is fitted to the data, the deviation of its predicted proportions from the actual response proportions for each patient is calculated as the uncorrected root mean squared deviation (RMSD). For the current sample, the average RMSD was 0.021, which is a good fit, slightly better than the .024 found in Schwartz et al.’s (2006) study of picture naming by 94 patients.

Comprehension was assessed at three levels: (1) semantic comprehension was measured by the 52-item Pyramids and Palm Trees test (Howard & Patterson, 1992), in which a pictured item must be matched to the closest associate among a set of two pictured choices on each trial (e.g. a picture of a pyramid must be matched to a picture of a palm tree or a picture of a pine tree). Abstract semantic knowledge, without accessing the correct lexical item, is enough to successfully complete this task. (2) Lexical comprehension was measured by the 30-item Synonym Judgment test (Saffran, Schwartz, Linebarger, Martin, & Bochetto, 1988). On each trial, the subject views three written words that are spoken aloud by the examiner and must decide which two are most similar in meaning (e.g., violin, fiddle, clarinet). This test requires semantic comprehension not only at the abstract level, but also at the word level, because meaning must be accessed through word-forms, without the aid of pictures. This test has two variants, noun and verb. Because all the PNT targets are nouns, we only used the noun variant of the test in this study. Since lexical comprehension is a key component of the perceptual-loop monitor, we used a second test to ensure that the results obtained by using the Synonym Judgment test are corroborated. We used the Peabody Picture Vocabulary Test, Third Edition-form A (Dunn & Dunn, 1997), in which the patient must match a heard word to one of the four pictures that best represent the meaning of that

word. (3) Finally, phonological comprehension was measured using the 40-item Phonological Discrimination test (N. Martin, 1996). In this task, the patient hears two items in immediate succession (20 lexical trials, and 20 nonlexical trials) and has to judge whether the items in each trial were the same or different. Non-identical pairs differ by a single onset or final phoneme. This task could be accomplished without access to meaning or even lexical knowledge.

Refining the sample for a study of error detection

The goal of the analyses was to find which measures reliably predict success or failure of error detection. To this purpose, we created a subsample of the 63 patients whose data were most likely to be accurately coded and to be revealing of error detection behavior. Specifically, patients were excluded from the subsample if they met either of the following criteria:

1. Moderate to severe speech apraxia (as measured by the Apraxia Battery for Adults; Dabul, 2000): the reason for this exclusion is that in such patients a portion of the phonological errors may not be due to low *p* weights, but instead to the malfunction of the articulatory system (e.g., Romani, Olson, Semenza & Grana, 2002). In this case, detection of such errors is not expected to correlate with the strength of the patient's *p* weight.
2. High rate of omissions: This was defined by calculating two scores for each patient. The raw score was the proportion of correct responses overall, and the normalized score was the proportion of correct responses when omission trials were thrown out. If the difference between these two scores was greater than 20%, the patient was excluded. One reason for excluding these high-omission patients is that when there are too many omissions, the model's estimate of the *s* and *p* weights becomes unreliable (Dell et al., 2004). Another is that many silent omissions could be due to covert error detection, which directly speaks to the ability of the patient in detecting errors; but we had no way of discriminating between omissions that did or did not reflect covert error detection.

After these criteria were enforced, 29 patients were selected. Their error detection behavior was assessed and related to the measures of comprehension and production ability.

Results

Correlations

Recall from simulation I that conflict at each layer was predictive of the error type originating mainly from that layer. We showed that the probability of the model making a semantic error was best predicted by the amount of conflict at the word layer while the probability of making a nonword error was best predicted by the amount of conflict in the phoneme layer. Moreover, simulation II showed that the reliability of the error signal varied as a function of the weights in the production system. If weights were strong, detection sensitivity was high. As weights decreased, so did the sensitivity with which errors were detected. Taken together, these two principles suggest that the strength of the *s* weights (which predicts conflict at the word layer) must be correlated with the detection of the semantic errors (which originate mainly from that layer). The same relationship should hold between the strength of the *p* weights and the detection of the phonological errors.

Table 2 shows the patients' scores on the four comprehension tests (expressed in percentage correct), their *s* and *p* parameters estimated from their naming performance, along with the total number of their semantic and phonological errors and the proportion of those errors detected. A total of 384 semantic and 461 phonological errors were coded for the 29

patients, out of which 251 (65%) and 262 (57%) were detected respectively. Figure 4 shows the correlations between the production weights and each error type. Semantic-error detection was correlated with the strength of the *s* weights (r= .59, p =.001), while phonological-error detection showed a correlation with the strength of the *p* weights (r= .43, p = .021). As expected by the principle of layer specificity, these positive correlations were not observed when the level-specific detection rates were paired with the "wrong" model weights. In fact, the corrections were negative (−.34 for detection of phonological errors and the strength of *s* weights; −.55 for semantic error detection with the strength of the *p* weights). These negative correlations may have arisen in part because of a negative correlation between the values of the *s* and *p* weights in our sample (r = −.32). In any case, it is clear that higher weights at a particular level are associated with better detection rate for errors only at that level.

Table 3 presents the correlations between the comprehension measures and percentages of detected semantic and phonological errors, none of which reaches significance. This lack of significance contrasts with the significant correlations between error detection and the appropriate layer-specific production measures, suggesting that the sample size had adequate power. Moreover, note that none of the comprehension correlations was larger than .24, and that the mean of the eight such correlations was −.01, which we consider to be a convincing null result. Finally, recall that the lack of a correlation between comprehension and error detection is consistent with the study of Nickels and Howard (1995), who used a different set of comprehension tests (but see Roelofs, 2005).

As Table 2 shows, some of the patients make very few errors of a certain type. This is especially true of semantic errors; 9 patients make 5 or fewer semantic errors. When turned into proportions, small numbers can cause problems, for example, imagine a patient who makes only a single semantic error. If he detects that error, it will be registered as 100% detection, but if he misses, suddenly the detection rate drops to zero. This bounciness makes for noisy data. For this reason, we checked our results using a simplified hierarchical logistic mixed model, which uses the information on each trial, nested under each patient, to assess the effect of the independent variables on error detection.

The two-step nature of the production model, which gives rise to the layer specificity of conflict-based error detection, made it reasonable to build two logistic regression models one concerned with semantic-error detection and one with phonological-error detection. Each model contained random effects for subjects and items, and detection of the error was the binary dependent variable. Critically, the independent variables for each model were chosen to explicitly pit the relevant production predictor against the relevant comprehension predictors, because the key question is whether detection depends on good production or good comprehension. For the regression examining detection of semantic errors, the predictors included *s* weight (for production) and measures of comprehension processes that would be expected to be needed to detect a semantic error such as "dog" for "cat". The relevant comprehension processes involve the phonological, lexical, and semantic levels. It is thus reasonable to include all four of our comprehension measures in the logistic regression model with the detection of semantic errors as its dependent variable.

For the regression examining detection of phonological errors (e.g. "cag" for "cat"), the relevant predictors are *p* weight (for production) and a comprehension measure that indexes the ability to recognize and compare strings of speech sounds, which is exactly what is assessed by the Phonological Discrimination test. This was, therefore, the only comprehension measure entered in the model with the phonological error detection as the dependent variable. The semantic model was thus built with five fixed effects: the strength of subjects' *s* weights, as well as their scores on the Pyramids & Palm Trees, Synonym

Judgment (Noun), PPVT-III, and Phonological Discrimination tests. The phonological model had only two fixed effects, the strength of the *p* weights and the Phonological Discrimination scores.

The findings of the correlation analysis were confirmed: In the semantic error analysis, the *s* weights were predictive of semantic error detection (coefficient = 52.17; p = .042. A positive coefficient means that the larger the *s* weights, the greater the proportion of semantic errors that were detected). But Pyramids & Palm Trees, Synonym Judgment, PPVT-III or Phonological Discrimination scores had no predictive power (coefficients = −.001, .001, −.009, and −0.014; p = .98, p = .93, p = .64, and p= .44 respectively). In the Phonological error analysis, the *p* weights were predictive of the detection of phonological errors (coefficient = 100.44; p = .008), but Phonological Discrimination scores were not (coefficient = −.02; p = .19).

In summary, we found no correlation between error detection and comprehension at semantic, lexical or phonological level. This result is compatible with the cases of aphasic patients reported to have shown a discrepancy between the ability to comprehend and the ability to monitor their own speech for errors. These cases and our results cannot be explained by a comprehension-based monitor such as the perceptual loop. On the other hand, we did observe a correlation between the detection of specific error types and specific production parameters, in the fashion predicted by our simulations of the conflict-based model. In the remainder of this section, we discuss some cases where the difference between the predictions of the comprehension and production-based monitors can be observed at the level of individual patients. In addition, we report two phenomena exhibited by some patients, which, we believe, are more easily explained by the conflict-based account. In the summaries below, the reported brain imaging findings were obtained from research CT or MRI studies performed near to the time of behavioral testing.

Patient 24

This patient was a 59-year old male, 16 months post-onset of acute stroke due to Left Middle Cerebral Artery infarct. His CT scan showed that the lesion was mostly frontal (BA 44) to parietal (BA 39). His comprehension profile reflected low comprehension ability at all three levels, with 46% correct on the Pyramids & Palm Trees test (mean of the sample of the 29 patients = 87.59, 95% CI = 83.45–91.73), 13% correct on the Synonym Judgment (Noun) test (sample mean = 83.45, 95% CI = 75.32–91.58), 61% correct on PPVT-III (sample mean = 80.93, 95% CI = 76.09–85.77), and 55% correct on the Phonological discrimination test (sample mean = 88.72, 95% CI = 84.08–93.37). His production profile, showed marked weakness of the semantic weights (*s*= 0.003; sample mean *s* weight = 0.027, 95% CI = 0.023–0.031) but relatively preserved phonological weights (p= 0.032; sample mean *p* weight = 0.024, 95% CI = 0.021–0.028).

There are two characteristics of this patient which make him an excellent case for testing the predictions of the conflict-based monitor: (1) The discrepancy between the values of the two production parameters predicts differential ability to detect semantic and phonological errors, if the monitor is in fact production-based. The prediction would be that since the patient's *s* weight is lower than the sample's average, his semantic-error detection should also be lower than average semantic-error detection in this sample. On the other hand, his better-than-average *p* weights, would predict superior phonological-error detection compared to the sample's average detection of such errors. (2) At the phonological level alone, there is a discrepancy between production and comprehension phonological processing abilities. The patient's *p* parameter is close to normal and above the sample's average *p* weight (see above), while his score on the Phonological Discrimination task is

markedly lower than the sample's average score. So, the conflict-based monitor predicts good phonological detection while the perceptual-loop account predicts the opposite.

As expected from the patient's pattern of deficit, he made many semantic errors (n = 30) but fewer phonological errors (n = 12). He detected 50% of his semantic errors, which is lower than the sample's average (mean percentage of semantic-error detection in the sample = 65.28; 95% CI = 55.96–74.59) but his 100% detection of his 9 phonological errors, was well above the sample's average (mean percentage of phonological-error detection in the sample = 56.83; 95% CI = 46.54–67.12). Thus, the difference in the production parameters is directly reflected in this patient's differential ability in the detection of semantic vs. phonological errors. Moreover, perfect detection of phonological errors is unexpected from a patient with such poor phonological input processing if the monitor is comprehension-based, but is well expected from the perspective of a production-based monitor, since the patient has near-normal phonological weights.

Patients 05, 17 and 20

Patient 05 was a 26-year old female, 7 months post-onset of stroke due to left parietal-occipital intraparenchymal hemorrhage. Her MRI showed the lesion to be mostly confined to inferior parietal areas, BA 7, 40, 39. She had good semantic and lexical comprehension with 87% correct on Pyramids & Palm Trees, 80% correct on Synonym Judgment (Noun), 86% correct on PPVT-III and almost perfect phonological comprehension (95% correct on the Phonological Discrimination test). Her production profile showed higher than average semantic weights ($s = 0.046$) but poorer than average phonological weights ($p = 0.018$), the opposite of the pattern observed in patient 24. A production-based monitor, thus, would predict better than average semantic error detection, but lower-than average phonological error detection. Also, when production and comprehension abilities are compared only at the phonological level, again the patient shows a discrepancy in the opposite direction of patient 24. Her phonological comprehension is intact, but her phonological production is degraded.

On the PNT, she made a single semantic error which she detected, and 11 phonological errors, only 2 of which (18%) she detected. Although her single semantic error does not allow for any conclusions to be drawn about semantic error detection, recall from the above that 18% is well below the average phonological-error detection in the sample. This is what the production-based monitor would predict from the weak *p* weights. A comprehension-based monitor, though, would have a hard time explaining why a patient with very good phonological comprehension would be so poor in detecting her phonological errors. However, as we pointed out in the introduction, it has been suggested that in addition to perfect comprehension, other abilities are required for successful error-detection through comprehension. Speakers must be able to successfully hold representations of both the target and the uttered word in their working memory and perform some kind of phonological comparison between the two.

To explore whether problems other than comprehension were interfering with the patient's ability to detect her phonological errors, we looked at a number of other measures in this patient. To assess the general ability of the patient to hold items in the memory, we evaluated her short-term memory span with Immediate Serial Recall Span for Words (R. Martin, Shelton, & Yaffee, 1994). The patient scored 3.2, meaning that she correctly recalled more than 50% of the 3-word lists, as well as about 20% of the 4-word lists. For a comprehension-based monitor to successfully compare the response to the target, holding only two words in working memory is sufficient.

To more specifically assess the patient's ability to compare phonological strings in short-term memory, we used the Rhyme Probe task (based on Freedman & R. Martin, 2001). The

subject listens to a string of two words, quickly followed by a third, and then must determine if the final word rhymed with either of the preceding words by saying or pointing to "Yes/No". The string of words gradually increases and the test is terminated when a subject drops to 75% accuracy on any list. Performance yields the subject's maximum phonological short term memory span. Patient 05 scored 3.16, meaning that she could successfully hold 3 words in her working memory and compare them to a fourth one to make a decision about the similarity of their phonological properties (again, all that is needed for monitoring is successful comparison of two words). We conclude that this patient seems to have the cognitive resources required by a comprehension-based monitor for successful phonological error-detection. Yet she rarely detects these errors. A similar case has been reported by Oomen et al. (2005).

It is noteworthy that two other patients in our sample demonstrated a strikingly similar profile to patient 05. Patient 17 was a 48-year-old male, 148 months post-onset of acute stroke due to Left Middle Cerebral Artery infarct. According to his CT scan the lesion was mostly frontal, including extensive damage to BA 44. He had good comprehension scores (96 on Pyramids & Palm Trees test, 100 on the Synonym Judgment, 83 on PPVT-III, and 88 on Phonological Discrimination), strong semantic weights ($s = 0.045$), but poor phonological weights ($p = 0.017$). His short-term memory span score was 3.4, and his Rhyme Probe score, 3.66. He made 2 semantic errors, both of which he detected, but only 28% of his 18 phonological errors were detected. This 28% is well below the sample's mean detection rate for phonological errors, while the patient's phonological comprehension score is no lower than average. Together with his sufficient memory span and phonological comparison skills, this finding is problematic for a comprehension-based monitor.

Patient 20 was a 46-year-old male, 81 months post-onset of stroke due to left intracerebral subarachnoid hemorrhage with mostly posterior frontal lesion, affecting BA 6 and adjacent prefrontal regions on CT scan. He also had very good comprehension scores (96 on Pyramids & Palm Trees test, 100 on the Synonym Judgment, 92 on PPVT-III, and 93 on Phonological Discrimination), strong semantic weights ($s = 0.05$), but poor phonological weights ($p = 0.017$). His short-term memory span score was 4, and his Rhyme Probe score, 2.38. Similar to patient 17, he detected both of his semantic errors, but only 25% of his 12 phonological errors. The latter is too low when compared to his above-average phonological comprehension and his sufficient working memory and phonological short-term memory capacity.

To summarize, we reported four aphasic patients whose error-detection ability was dissociated from their comprehension ability. Three of these patients had good comprehension but poor self-monitoring, and one showed the reverse pattern. Although problematic for the perceptual loop theory, this pattern is well-predicted by the conflict-based account: detecting a certain error type for all four patients was predicted by the strength of the production weight that was responsible for the generation of that error type.

Doubts and false rejections

In normal speakers, false alarms are thought to manifest mainly as disfluencies. Our data suggest that in aphasic patients, false alarms can surface in other ways, as doubts and false rejections.

By *doubts* we refer to cases where the patient responds correctly, but immediately questions the accuracy of his/her response (e.g. in response to the target "towel" the patient says "towel, is it towel?"). *False rejections* are cases where the patient overtly rejects his/her correct response and may or may not replace it with an incorrect one (e.g. in response to the target "pineapple", the patients responds "pineapple, no." or "pineapple, no, watermelon").

Although not very common, 7 patients in our sample showed more than one instance of doubts/false rejections, with one showing a remarkably high number (15 doubts and 3 false rejections). Interestingly, this patient had high scores on all four comprehension measures (87 on the Pyramids & Palm Trees test, 100 on the Synonym Judgment test, 84 on the PPVT-III, and 98 on the Phonological Discrimination test), along with good working memory span (3.6) and Rhyme Probe (2) scores. If the response is monitored by the comprehension system, there is no compelling reason for this patient to show so many false alarms.

The conflict-based account, though, has an explanation. As weights become weaker, there is overall more conflict on correct trials, compared to when weights are strong. In other words, the amount of conflict on correct trials approaches that on the error trials, which means lower detection sensitivity. Lower sensitivity leads to fewer hits (more misses) and more false alarms. Missing more errors as the weights decrease is common in our sample (as evidenced by our correlation analysis). However, it appears that many patients are continuing to detect and some have reasonable hit rates despite damage. The cost of achieving these hits will then be more false alarms. We thus propose that the doubts and false rejections seen in our data are aphasic manifestations of false alarms of the error detection system.

General Discussion

A comprehensive study of some particular cognitive function consists of studying that function when it is carried out perfectly, when it encounters problems and ends in errors and, finally, how those errors are processed and repaired. The first two of these -the analysis of correct performance and error generation- require a detailed study of the cognitive domain in question. For example, knowing a lot about word-production response times does not give you much information about how visual scenes are processed. Similarly, visual illusions are not particularly informative about the nature of speech errors, because the errors in each domain reflect the specific processes within that domain. However, detection of errors might not be quite as domain-specific. A domain-general error detection system may be responsible for the detection of all slips independent of the cognitive process from which those slips arise (Rabbitt, 1966a, b).

In recent years, the domain-general error detection system has found support in an electrophysiological potential, the ERN, observed in speeded forced-choice tasks. Even though the exact mechanism by which the ERN is generated is still debated, there is no doubt about its relation to error detection, and most importantly its significance as a signal for central processing of errors committed by different effectors and in different tasks. In light of this, a number of domain-general theories of error processing have been proposed and supported (e.g. Yeung et al., 2004). We believe that speech production provides an excellent opportunity for the application of the conflict-based approach to error processing, because (1) all speakers slip and all neurologically unimpaired speakers can detect their slips, (2) the pressure to keep the speech flow in spite of problems like slips makes speaking a “speeded” task, and (3) the growing body of ERP evidence attests to the similarities between monitoring linguistic and non-linguistic processes. However, unlike the forced-choice tasks, speech production is a natural task, in which the response alternatives are constrained by the speaker’s intention and functional properties of the production system. Therefore, there are enough similarities between speech production and forced-choice tasks to presume that a common error-detection mechanism might be shared between the two, but at the same time, there are clear differences, which require that the theory be tailored to language production before a generalization can be made. The possibility of a domain-general error detection mechanism for language production has been recently proposed (Riès

et al., 2011), but no implemented or otherwise detailed and testable model has as of yet been developed.

For this paper, we had two goals: (1) To develop a new model of error detection in language production. (2) In so doing, to provide additional support for the notion of a central generic monitoring mechanism for detecting errors that extends beyond the scope of laboratory tasks with limited response choices. To this end, we implemented a simple version of a conflict-based monitor in the interactive two-step model of word production (Dell et al., 1997), laid out and tested its assumptions in two simulations, and tested explicit predictions about patient data using the model.

Our first goal was motivated by the evidence pertaining to the insufficiency of the current theory of speech monitoring, the perceptual loop account, in explaining the empirical data. We reviewed the assumptions of this account, along with the various pieces of evidence questioning those assumptions. Instead, we proposed a conflict-based model of error-detection which is not subject to the two main objections against the perceptual loop theory: it does not assume processing of inner speech during overt speech production and, more fundamentally, it does not rely on comprehension for error detection. Its central premise is that an error is signaled by the presence of high conflict between various options at the time of selection among activated words, and later, among activated phonemes. Conflict can be measured in different ways. In our implementation we measured it in two ways, once by taking into account the competition from all alternatives, and once by considering only the strongest competitor. In either case, the crucial measure was the difference between the activation of the selected (response) node and the activation of the competitor(s). The greater this difference (the more distinct the response node), the less the conflict between the selected node and other potential responses and the lower the chance of making an error.

We identified three principles that should hold for conflict-based error-detection in a speech production model. According to the first principle (detection sensitivity), the amount of conflict should correlate directly with the probability of error commission. This principle is modulated by the second principle (layer specificity), which constrains the relationship between the amount of conflict at specific layers of the production system and the probability of certain error types. To demonstrate these principles, we used the interactive two-step model of word production. In the model, the nonlinearity of the process of mapping meaning to sound is enforced by selection at two points in the process of lexical access: if “cat” is to be produced, selection 1 happens at the word layer, when a lemma (e.g. “cat”) is chosen; and selection 2 happens at the phoneme layer, when the sounds of the selected lemma (e.g. /k/,/æ/,/t) are chosen. Although there is only one overt response, the two-step process of lexical retrieval makes it desirable to measure conflict at two layers, where a node is selected among other potential nodes.

Also, the two-step nature of the model, which enforces separate mapping of meaning to word units and word to sound units, creates differential error probabilities during each step. If selection 1 suffers, a semantic error may be generated and, if selection 2 suffers, a phonological error (often a nonword error) is created. Thus, conflict measured at the word layer (selection 1) must be correlated with the probability of semantic error commission and conflict at the phoneme layer (selection 2) with the probability of nonword error commission. This relationship between the conflict arising at a specific point in the model and a certain error type is summarized by the principles of detection sensitivity and layer specificity and our simulation of a normal speaker (simulation I) confirms it.

The third principle of the model (integrity contingency) establishes the relationship between the informativeness of the error signal and the quality of the production system, as

determined by the *s* and *p* weights. In a damaged production system with weak weights, noise has more effect, and conflict is more likely to be omnipresent, regardless of whether the trial ends in a correct or an incorrect response. In this case, using conflict as a signal will not reliably discriminate between errors and correct responses. Thus, the third principle predicts a direct relationship between the first two principles and the strength of the production weights. When weights are strong, conflict at each layer should accurately predict the probability of the relevant type of error. As weights become weaker, this precision should decrease, so that when production is greatly disrupted (as in a case of severe aphasia), the conflict signal no longer carries useful information. Our second simulation confirms the third principle.

It is this third principle that is most relevant to the distinction between the perceptual loop model and the conflict-based model. While the former predicts a correlation between comprehension and error detection and little correlation between the quality of the production system and error detection, the latter makes the opposite prediction. The patient-data section of the paper tested the predictions of these two theories on a sample of aphasic patients. We replicated the previous finding of no correlation between comprehension and error detection (Nickels & Howard, 1995), and in the same sample showed correlations between detection ability and production weights, as predicted by the conflict-based model.

Finally, we applied a criterion-setting model (Kac, 1962) to the data generated using the conflict-based model and showed that the model is capable of simulating the detection rate observed in normal speakers and its drop in the aphasic patients. Based on our findings, we propose the conflict-based account of monitoring for detecting speech errors. It is noteworthy that a conflict-based model may be capable of explaining other error-related findings. One such example is the fidelity of speech errors to the phonotactic constraints of the language (e.g. Fromkin, 1971). For example, maybe producing the phonotactically illegal /kd/ onset cluster is much less probable than producing the legal /kl/ onset cluster, because at the phonological level, an activated /k/onset and /d/onset is viewed as more conflicted than an activated /k/onset and /l/onset and hence the illegal cluster is inhibited or covertly repaired. Similarly, an overall pattern of phonological activation that is, or is similar to, a word, might be viewed as less conflicted than one corresponding to a nonword, a perspective that is consistent with claims that speech errors that create nonwords are more detectable than ones that form words (Baars et al.,1975; Hartsuiker, Corley, & Martensen, 2005; Nooteboom & Quené, 2008). In this way, conflict is more like the inverse of the "goodness" of an activation pattern, such as the Hopfield Energy measure of conflict used by Botvinick et al. (2001) and Yeung et al. (2004). Unconflicted or "good" activation patterns require a clear distinction between activated and unactivated units (as in our measures) *and* that the set of activated units cohere with background knowledge (e.g. phonotactic or lexical constraints).

Our claim that the primary mechanism for speech monitoring is detection of conflict does not imply that we deny the role of perceptual processes in speech monitoring altogether. Empirical studies have shown that although error detection is possible in noise-masked conditions, the percentage of detected errors (at least for phonological errors) decreases (Lackner & Tuller, 1979), suggesting some role of an external monitoring loop that processes the auditory signal. Electrophysiological data second this observation: on some error trials a slow positive wave with centroparietal distribution (Pe) appears, the timing of which, unlike the ERN, is compatible with processing of peripheral information (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Falkenstein, Koshlykova, Kiroj, Hoormann, & Hohnsbein, 1995). On the basis of these and similar arguments, the possibility of a hybrid (production/comprehension) model of monitoring has been discussed in recent years (Nickels & Howard, 1995; Postma, 2000; Schlenk et al., 1987; Slevc, 2006; Vigliocco

& Hartsuiker, 2002). In a similar vein, Logan and Crump (2010) presented evidence for a dual monitoring system in type-writing, which seems to take advantage of both production-based and perceptual signals for detecting errors. It is possible that in such a hybrid model, the conflict signal, since it is available before production, contributes different information to the repair process than does post-response perceptual feedback. Perhaps the former readies the system for repair while the latter provides information about the details that need to be fixed.

Cues for error detection could be provided in a number of other ways as well. A good example is the processing of social signals. Language production is, for the most part, an interaction between a speaker and a listener (e.g. Branigan, Pickering, McLean, & Cleland, 2007; Clark, 1996; Clark & Wilkes-Gibbs, 1986) and there is growing interest in how social interaction affects the cognitive processes underlying speech generation (e.g. Horton & Gerrig, 2005). Speech monitoring might also take advantage of the interactive nature of conversation. It is plausible that the listener's confusion would act as a cue for the speaker that revision is required. Such an adaptive error-detection strategy has been proposed for jargon aphasics whose detection ability improves over time (J. Marshall et al., 1998).

Also, we have only modeled error detection at the level of individual words. Most accounts of language production/acquisition propose that speakers are sensitive to the transitional probabilities of words in their context, meaning that certain categories (noun, verbs, etc.) appear before or after other categories with probabilities that are learned by the speakers (e.g. Chang, Dell, & Bock, 2006; Wonnacott, Newport, & Tanenhaus, 2008). It is thus conceivable that when there is a context involved, violations of such contingencies will provide additional cues for the detection of certain error types. This proposal shares a perspective with MacKay's NST, in which occurrence of a unit in a novel context is proposed as a signal for error detection.

In short, we propose the conflict-based account as the core mechanism for speech monitoring, while acknowledging the possibility of complementary mechanisms which, in the case of pathology of the core mechanism, might gain a more substantial role in monitoring speech for errors.

Towards a domain-general account of error detection

Throughout the paper, we argued for a domain-general error-detection mechanism, and claimed that the model proposed here implements that view for language production. In this section, we elaborate on the similarities and differences between our model and other conflict-based theories (e.g. Botvinick et al., 2001; Yeung et al., 2004). The model proposed in this paper is not simply an implementation of prior conflict-based theories in the domain of language production. The differences between our model and models of Botvinick et al. (2001) and Yeung et al. (2004) stem from the differences in the target tasks. For us it is error detection during natural speech production, for them it is error detection under laboratory conditions for forced-choice tasks. This leads to two ways in which our models differ:

The point of conflict measurement—We measured conflict only at the time of response selection (once at the word level, and once at the phonological level). This differs from conflict-based models where the point of measurement is post-response. In those models, activation of a hasty response conflicts with the subsequent activation of the correct answer which emerges from further processing after the response is made. We focus on our two selection points because, first, they are motivated by the production theory, and second, because they successfully simulate the empirical findings. On the theoretical side, conflict-resolution for selection is an inevitable step in word production. The greater the conflict, the more difficult the selection. In fact, the amount of conflict at the time of selection is most

relevant for estimating the accuracy of the selection. Thus it seems natural that the conflict signal is relayed for the (secondary) task of error detection at the time when the system must face conflict resolution in order to accomplish the (primary) task of production. A consequence of measuring conflict at the time of selection is that our model is blind to the correct response. It picks the most activated node and at the same time assesses its confidence in its pick by comparing that node's activation to others. Post-response measures of conflict, on the other hand, are sensitive to the conflict between the executed and the correct response (which is derived from continuing process of the stimulus in forced-choice tasks).

On the empirical side, measuring conflict at the point of selection leads to a model which explains the data well. Conflict at the time of word selection is strongly predictive of word-level errors, and conflict at the time of phonological selection is strongly predictive of phonological errors. When a criterion is derived using the simulated distributions, the model correctly predicts that normal speakers detect around half of their errors. Moreover, our decision to assess conflict at the point with which semantic errors occur (the word level) is consistent with findings in the literature. Ganushchak and Schiller (2008a) found that ERNs have larger amplitudes on errors that follow semantic distractors. Similarly, Ganushchak and Schiller (2008b) found larger ERNs in semantically-related (compared to unrelated) blocks. If the amplitude of ERN is assumed to be directly proportional to the amount of conflict, our model predicts this finding. Yeung et al.'s (2004) model on the other hand, always predicts larger ERN for low-conflict stimuli.

Choice of the conflict measure—The conflict based models that were applied to binary tasks use a measure of conflict derived from Hopfield energy (e.g. Botvinick et al. 2001, Yeung et al., 2004). This measure assesses the extent to which the activation pattern in a network is compatible with its connection weights. Specifically, response alternatives are assumed to have mutual inhibitory connections and thus their simultaneous activation indicates conflict. Our model does not have such connections, and thus we chose simpler and, we believe, more intuitive measures of conflict, such as the difference in the activation between the selected node and its principal competitor(s). Furthermore, Hopfield energy, as used in Yeung et al. (2004), is most informative if calculated over a period of time (as opposed to a single point in time), which makes it unsuitable for a model like ours which associates conflict at the exact point of response selection with the error signal.

Given these differences, what does our model share with the other conflict-based models? Before answering this question we briefly digress by considering the role of monitoring in complex motor movements. In motor movement -just as in language production- initial theories of monitoring emphasized monitoring through perception (e.g. correctness of arm movements determined by visual input after the motion). However, behavioral studies of arm movements showed that the trajectory of the movement can be amended at latencies much shorter (30–45 ms) than the time needed for the sensory feedback to be processed (Cooke & Diggles, 1984; van Sonderen, Gielen, & Denier van der Gon, 1989). To explain such fast and efficient corrections, a feedback loop was proposed, in which a copy of the efferent signal (the efference copy) is generated and monitored for its compatibility with the action, so that a predicted output is compared to the actual output. This so-called *forward model* (Desmurget & Grafton, 2000; Jordan & Rumelhart, 1992; Kawato, 1999; Miall & Wolpert, 1996), similar to the conflict-based model, uses the information generated by the production system (in this case, the motor system) for error detection rather than relying on the peripheral information processed by the perceptual system. We consider the relation between the conflict-based model and forward models by focusing on this question: Is monitoring a comparative process, and if so, what exactly is the nature of this comparison?

Whether to view error detection as a comparative process or not, depends on what the error is being compared to. In some theories, detection is achieved by comparison to the correct response (e.g. the perceptual loop account). Earlier, we mentioned that not all agree that it is plausible for the system to have the correct response at its disposal, while making an error. However, there are other potential comparative processes which are not subject to the same criticism. One such example is the forward model discussed above. Note that the comparison is between the actual and the *predicted* outcome, which the system derives from the information available from the action and the knowledge it has about the state of the environment. The predicted outcome might or might not be the desired outcome (i.e. the correct response), but the process is nevertheless a comparative process. A similar proposal has been formulated in the *response monitoring theory* (e.g., Steinhauser, Maier, & Hübner, 2008), in which error detection consists of a comparative process between the actual response and a representation of a response that is derived from further processing of the input information. The idea is that slips result from premature response generation, before the system has had time to use all the information and settle on the final (and presumably correct) response.

Conflict-based error detection can also be viewed as a comparative process. In its simplest form, such as the model used in this paper, errors are detected based on a comparison between the activation of two (or more) alternatives at the time of responding. The conflict model of Yeung et al. (2004) uses a more sophisticated comparison. Its monitor detects the conflict between the executed response and the response derived from continued processing of the stimulus. Note that in both cases the comparison is between an actual response and an internally-generated criterion of comparison.

Against this background, we can now consider the sense in which the models are similar. For one thing, all view error detection as a loop involving a lower system (e.g. motor or language production system) that generates signals of conflict and some other domain-general system (frontal, perhaps the ACC) that interprets these so as to mitigate or ultimately, to learn from, the error. Furthermore, all agree that the information necessary for error detection is provided by a non-perceptual comparison between various options generated within the system being monitored. Whether this comparison includes the desired response or not should matter little for error detection. In cases where the ultimate purpose of error detection is learning to avoid producing the error in the future, a forward model or a conflict-based process could be situated in a larger loop that compares the output of the model to the desired outcome. This loop could be a learning-reinforcement algorithm (Holroyd et al., 2002) or alternatively, a supervised learning algorithm (Jordan & Rumelhart, 1992).

In conclusion, this research assessed whether conflict-based error-detection can apply to speech production, and more generally, whether mechanisms of this sort that have been proposed for simple button-push tasks can apply to a more natural and complex task, specifically, speaking. We believe that the work presented here supports such an application.

Acknowledgments

This research was funded by a grant from the National Institutes of Health's National Institute for Deafness and Other Communication Disorders: DC000191. We would like to thank Michael Coles, Daniel Simons, Kathryn Bock, Susan Garnsey, Kara Federmeier, Aaron Benjamin, Robert Slevc, Robert Hartsuiker and an anonymous reviewer for their help.

References

Baars BJ, Motley MT, MacKay DG. Output editing for lexical status in artificially elicited slips of the tongue. Journal of Verbal Learning and Verbal Behavior. 1975; 14:382–391.

Blackmer ER, Mitton JL. Theories of monitoring and the timing of repairs in spontaneous speech. Cognition. 1991; 39(3):173–194. [PubMed: 1841032]

Bortfeld H, Leon S, Bloom J, Schober M, Brennan S. Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. Language and Speech. 2001; 44:123–147. [PubMed: 11575901]

Botvinick MM, Braver TS, Carter CS, Barch DM, Cohen JD. Evaluating the demand for control: Anterior cingulate cortex and crosstalk monitoring. Psychological Review. 2001; 108:624–652. [PubMed: 11488380]

Botvinick MM, Nystrom LE, Fissell K, Carter CS, Cohen JD. Conflict monitoring versus selection-for-action in anterior cingulated cortex. Nature. 1999; 402:179–181. [PubMed: 10647008]

Branigan HP, Pickering MJ, McLean JF, Cleland AA. Participant role and syntactic alignment in dialogue. Cognition. 2007; 104:163–197. [PubMed: 16876778]

Butterworth B, Howard D. Paragrammatism. Paragrammatism. 1987; 26:1–37.

Carter CS, Braver TS, Barch D, Botvinick MM, Noll D, Cohen JD. Anterior cingulate cortex, error detection, and the online monitoring of performance. Science. 1998; 280:747–749. [PubMed: 9563953]

Chang F, Dell GS, Bock K. Becoming syntactic. Psychological Review. 2006; 113:234–272. [PubMed: 16637761]

Clark, HH. Using language. Cambridge: Cambridge University Press; 1996.

Clark HH, Wilkes-Gibbs D. Referring as a collaborative process. Cognition. 1986; 22:1–39. [PubMed: 3709088]

Cooke JD, Diggles VA. Rapid error correction during human arm movements: Evidence for central monitoring. Journal of Motor Behavior. 1984; 16:348–363. [PubMed: 15151894]

Dabul, B. praxia Battery for Adults. 2. Austin, Texas: Pro-Ed; 2000.

Debener S, Ullsperger M, Siegel M, Fiehler K, von Cramon DY, Engel AK. Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. Journal of Neuroscience. 2005; 25(50):11730–11737. [PubMed: 16354931]

Dehaene S, Posner MI, Tucker DM. Localization of a neural system for error detection and compensation. Psychological Science. 1994; 5:303–305.

Dell GS. A spreading-activation theory of retrieval in sentence production. Psychological Review. 1986; 93:283–321. [PubMed: 3749399]

Dell GS, Lawler EN, Harris HD, Gordon JK. Models of errors of omission in aphasic naming. Cognitive Neuropsychology. 2004; 21(2):125–145. [PubMed: 21038196]

Dell GS, Martin N, Schwartz MF. A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. Journal of Memory and Language. 2007; 56:490–520. [PubMed: 21085621]

Dell GS, O'Seaghdha PG. Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). Psychological Review. 1991; 98:604–614. [PubMed: 1961775]

Dell GS, Schwartz MF, Martin N, Saffran EM, Gagnon DA. Lexical access in aphasic and nonaphasic speakers. Psychological Review. 1997; 104:801–838. [PubMed: 9337631]

De Smedt, K.; Kempen, G. Incremental sentence production, self-correction, and coordination. In: Kempen, G., editor. Natural language generation: recent advances in artificial intelligence, psychology, and linguistics. Dordrecht: Martinus Nijhoff Publishers; 1987. p. 365-376.

Desmurget M, Grafton S. Forward modeling allows feedback control for fast reaching movements. Trends in Cognitive Sciences. 2000; 4:423–431. [PubMed: 11058820]

Devinsky O, Morrell MJ, Vogt BA. Contributions of anterior cingulate cortex to behaviour. Brain. 1995; 118:279–306. [PubMed: 7895011]

Dunn, LM.; Dunn, LM. Examiner's Manual for the PPVT-III: Peabody Picture Vocabulary Test. 3. Circle Pines, MN: American Guidance Service; 1997.

Endrass T, Franke C, Kathmann N. Error awareness in a saccade countermanding task. Journal of Psychophysiology. 2005; 19(4):275–280.

Endrass T, Schuermann B, Kaufmann C, Spielberg R, Kniesche R, Kathmann N. Performance monitoring and error significance in patients with obsessive-compulsive disorder. Biological Psychology. 2010; 84(2):257–263. [PubMed: 20152879]

Falkenstein, M.; Hohnsbein, J.; Hoormann, J.; Blanke, L. Effects of errors in choice reaction tasks on the ERP under focused and divided attention. In: Brunia, C.; Gaillard, A.; Kok, A., editors. Psychophysiological brain research. Tilburg, the Netherlands: Tilburg University Press; 1990. p. 192-195.

Falkenstein M, Hohnsbein J, Hoormann J, Blanke L. Effects of crossmodal divided attention on late ERP components: II. Error processing in choice reaction tasks. Electroencephalography and Clinical Neurophysiology. 1991; 78:447–455. [PubMed: 1712280]

Falkenstein M, Koshlykova NA, Kiroj VN, Hoormann J, Hohnsbein J. Late ERP components in visual and auditory Go/Nogo tasks. Electroencephalography and Neurophysiology. 1995; 96:36–43.

Fox Tree JE. Effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. Journal of Memory and Language. 1995; 34:709–738.

Foygel D, Dell GS. Models of impaired lexical access in speech production. Journal of Memory and Language. 2000; 43(2):182–216.

Fraundorf, SH.; Watson, DG. Dimensions of variation in disfluency production in discourse. In: Ginzburg, J.; Healey, P.; Sato, Y., editors. Proceedings of LONDIAL 2008, the 12th Workshop on the Semantics and Pragmatics of Dialogue. London: King's College London; 2008. p. 131-138.

Freedman ML, Martin RC. Dissociable components of short-term memory and their relation to long-term learning. Cognitive Neuropsychology. 2001; 18(3):193–226. [PubMed: 20945211]

Fromkin VA. The non-anomalous nature of anomalous utterances. Language. 1971; 47:27–52.

Ganushchak LY, Schiller NO. Effects of time pressure on verbal self-monitoring. Brain Research. 2006; 1125:104–115. [PubMed: 17113572]

Ganushchak LY, Schiller NO. Brain error-monitoring activity is affected by semantic relatedness: an event-related brain potentials study. Journal of Cognitive Neuroscience. 2008a; 20(5):927–940. [PubMed: 18201131]

Ganushchak LY, Schiller NO. Motivation and semantic context affect brain error-monitoring activity: An event-related brain potentials study. Neuroimage. 2008b; 39:395–405. [PubMed: 17920932]

Ganushchak LY, Schiller NO. Speaking one's second language under time pressure: an ERP study on verbal self-monitoring in German-Dutch bilinguals. Psychophysiology. 2009; 46:410–419. [PubMed: 19207202]

Gehring WJ, Fencsik D. Functions of the medial frontal cortex in the processing of conflict and errors. Journal of Neuroscience. 2001; 21:9430–9437. [PubMed: 11717376]

Gehring WJ, Goss B, Coles MGH, Meyer DE, Donchin E. A neural system for error detection and compensation. Psychological Science. 1993; 4:385–390.

Gehring WJ, Himle J, Nisenson LG. Action-monitoring dysfunction in obsessive-compulsive disorder. Psychological Science. 2000; 11:1–6. [PubMed: 11228836]

Hajcak G, Moser JS, Yeung N, Simons RF. On the ERN and the significance of errors. Psychophysiology. 2005; 42:151–160. [PubMed: 15787852]

Hajcak G, Simons RF. Error-related brain activity in obsessive-compulsive undergraduates. Psychiatry Research. 2002; 110:63–72. [PubMed: 12007594]

Hartsuiker RJ, Corley M, Martensen H. The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars, Motley, and MacKay (1975). Journal of Memory and Language. 2005; 52:58–70.

Hartsuiker RJ, Kolk HHJ. Error monitoring in speech production: A computational test of the perceptual loop theory. Cognitive Psychology. 2001; 42(2):113–157. [PubMed: 11259106]

Hartsuiker, RJ.; Kolk, HHJ.; Martensen, H. The division of labor between internal and external speech monitoring. In: Hartsuiker, RJ.; Bastiaanse, R.; Postma, A.; Wijnen, F., editors. Phonological

encoding and monitoring in normal and pathological speech. Hove: Psychology Press; 2005. p. 187-205.

Hartung, J.; Knapp, G.; Sinha, BK. Statistical Meta-Analysis with Application. Hoboken, New Jersey: Wiley; 2008.

Holroyd CB, Coles MGH. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. Psychological Review. 2002; 109:679–709. [PubMed: 12374324]

Holroyd CB, Dien J, Coles MGH. Error-related scalp potentials elicited by hand and foot movements: Evidence for an output independent error-processing system in humans. Neuroscience Letters. 1998; 242:65–68. [PubMed: 9533395]

Holroyd, CB.; Nieuwenhuis, S.; Mars, RB.; Coles, MGH. Anterior cingulate cortex, selection for action, and error processing. In: MI, editor. Cognitive neuroscience of attention. New York: Guilford Press; 2004.

Horton WS, Gerrig RJ. The impact of memory demands on audience design during language production. Cognition. 2005; 96:127–142. [PubMed: 15925573]

Howard, D.; Patterson, K. Pyramids and Palm Trees: a test of semantic access from pictures and words. Bury St. Edmunds, Suffolk: Thames Valley Test Company; 1992.

Huettig F, Hartsuiker RJ. Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. Language and Cognitive Processes. 2010; 25(3):347–374.

Indefrey P, Levelt WJM. The spatial and temporal signatures of word production components. Cognition. 2004; 92(1–2):101–144. [PubMed: 15037128]

Johannes S, Wieringa BM, Müller-Vahl KR, Dengler R, Münte TF. Excessive action monitoring in Tourette syndrom. Journal of Neurology. 2002; 249:961–966. [PubMed: 12195438]

Jordan ML, Rumelhart DE. Forward models: Supervised learning with a distal teacher. Cognitive Science. 1992; 16:307–354.

Kac M. A note on learning signal detection. IRE Transactions on Information Theory. 1962; IT-8:126–128.

Kawato M. Internal models for motor control and trajectory planning. Current Opinion in Neurobiology. 1999; 9:718–727. [PubMed: 10607637]

Kempen G, Huijbers P. The lexicalisation process in sentence production and in naming: Indirect election of words. Cognition. 1983; 14:185–209.

Kinsbourne M, Warrington EK. Jargon aphasia. Neuropsychologia. 1963; 1:27–37.

Lackner, JR.; Tuller, BH. Role of efference monitoring in the detection of selfproduced speech errors. In: Cooper, WE.; Walker, ECT., editors. Sentence processing: studies dedicated to Merrill Garrett. Hillsdale, NJ: Earlbaum; 1979. p. 281-294.

Laver, JDM. The detection and correction of slips of tongue. In: Fromkin, VA., editor. Speech errors as linguistic evidence. The Hague: Mouton; 1973. p. 132-143.

Laver, JDM. Monitoring systems in the neurolinguistic control of speech production. In: Fromkin, VA., editor. Errors in linguistic performance: slips of the tongue, ear, pen, and hand. New York: Academic Press; 1980. p. 287-305.

Levelt WJM. Monitoring and self-repair in speech. Cognition. 1983; 14:41–104. [PubMed: 6685011]

Levelt, WJM. Speaking: From intention to articulation. Cambridge, MA: MIT Press; 1989.

Liss JM. Error-revision in the spontaneous speech of apraxic speakers. Brain and Language. 1998; 62:342–360. [PubMed: 9593614]

Logan GD, Cowan WB. On the ability to inhibit thought and action: A theory of an act of control. Psychological Review. 1984; 91:295–327.

Logan GD, Crump MJC. Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. Science. 2010; 330:683–686. [PubMed: 21030660]

MacKay DG. On the retrieval and lexical structure of verbs. Journal of Verbal Learning and Verbal Behavior. 1976; 15:169–182.

MacKay, DG. The organization of perception and action: a theory for language and other cognitive skills. New York: Springer-Verlag; 1987.

MacKay DG. Awareness and error detection: New theories and research paradigms. Consciousness & Cognition: An International Journal. 1992; 1(3):199–225.

Maher LM, Rothi LJG, Heilman KM. Lack of error awareness in an aphasic patient with relatively preserved auditory comprehension. Brain & Language. 1994; 46:402–418. [PubMed: 7514943]

Marshall J, Robson J, Pring T, Chiat S. Why does monitoring fail in jargon aphasia? Comprehension, judgment, and therapy evidence. Brain & Language. 1998; 63(1):79–107. [PubMed: 9642022]

Marshall RC, Rappaport BZ, Garcia-Bunuel L. Self-monitoring behavior in a case of severe auditory agnosia with aphasia. Brain & Language. 1985; 24:297–313. [PubMed: 3978408]

Marslen-Wilson WD, Tyler LK. The temporal structure of spoken language understanding. Cognition. 1980; 8(1):1–71. [PubMed: 7363578]

Martin N. Auditory Discrimination of Word and Non-word Pairs. Unpublished test. 1996

Martin RC, Shelton JR, Yaffee LS. Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. Journal of Memory and Language. 1994; 33:83–111.

Masaki H, Tanaka H, Takasawa N, Yamazaki K. Error-related brain potentials elicited by vocal errors. NeuroReport. 2001; 12:1851–1855. [PubMed: 11435911]

Mattson, ME.; Baars, BJ. Error-minimizing mechanisms: boosting or editing?. In: Baars, BJ., editor. Experimental slips and human error: exploring the architecture of volition. New York: Plenum Press; 1992.

McNamara P, Obler LK, Au R, Durso R, Albert ML. Speech monitoring skills in Alzheimer's disease, Parkinson's disease, and normal aging. Brain and Language. 1992; 42:38–51. [PubMed: 1547468]

Miall RC, Wolpert DM. Forward models for physiological motor control. Neural Networks. 1996; 9(8):1265–1279. [PubMed: 12662535]

Miltner WHR, Braun CH, Coles MGH. Event-related potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. Journal of Cognitive Neuroscience. 1997; 9:788–798.

Miltner WHR, Lemke U, Weiss T, Holroyd C, Schevers MK, Coles MGH. Implementation of error-processing in the human anterior cingulate cortex: A source analysis of the magnetic equivalent of the error-related negativity. Biological Psychology. 2003; 64:157–166. [PubMed: 14602360]

Möller J, Jansma BM, Rodríguez-Fornells A, Münte TF. What the brain does before the tongue slips. Cerebral Cortex. 2007; 17:1173–1178. [PubMed: 16831855]

Nickels L, Howard D. Phonological errors in aphasic naming: comprehension, monitoring and lexicality. Cortex. 1995; 31:209–237. [PubMed: 7555004]

Nieuwenhuis S, Ridderinkhof KR, Blow J, Band GPH, Kok A. Error-related brain potentials are differentially related to awareness of response errors: evidence from an antisaccade task. Psychophysiology. 2001; 38(5):752–760. [PubMed: 11577898]

Nieuwenhuis S, Yeung N, van den Wildenberg W, Ridderinkhof KR. Electrophysiological correlates of anterior cingulate function in a Go/NoGo task: Effects of response conflict and trial-type frequency. Cognitive, Affective, and Behavioral Neuroscience. 2003; 3:17–26.

Nooteboom, SG. Speaking and unspeaking: detection and correction of phonological and lexical errors of speech. In: Fromkin, VA., editor. Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen, and Hand. New York: Academic Press; 1980. p. 87-96.

Nooteboom, SG. Listening to oneself: Monitoring speech production. In: Hartsuiker, RJ.; Bastiaanse, R.; Postma, A.; Wijnen, F., editors. Phonological encoding and monitoring in normal and pathological speech. Hove: Psychology Press; 2005. p. 167-186.

Nooteboom SG, Quené H. Self-monitoring and feedback: a new attempt to find the main cause of lexical bias in phonological speech errors. Journal of Memory and Language. 2008; 58:837–861.

Núñez-Castellar E, Kühn S, Fias W, Notebaert W. Outcome expectancy and not accuracy determines posterror slowing: ERP support. Cognitive, Affective, & Behavioral Neuroscience. 2010; 10:270–278.

O'Connell RG, Dockree PM, Bellgrove MA, Kelly SP, Hester R, Garavan H, et al. The role of cingulate cortex in the detection of errors with and without awareness: a high-density electrical mapping study. The European Journal of Neuroscience. 2007; 25(8):2571–2579. [PubMed: 17445253]

Oomen CCE, Postma A. Effects of time pressure on mechanisms of speech production and self-monitoring. Journal of Psycholinguistic Research. 2001; 30:163–184. [PubMed: 11385824]

Oomen CCE, Postma A. Limitations in processing resources and speech monitoring. Language & Cognitive Processes. 2002; 17(2):163–184.

Oomen CCE, Postma A, Kolk HHJ. Prearticulatory and postarticulatory self-monitoring in Broca's aphasia. Cortex. 2001; 37:627–641. [PubMed: 11804213]

Oomen, CE.; Postma, A.; Kolk, HHJ. Speech monitoring in aphasia: Error detection and repair behaviour in a patient with Broca's aphasia. In: Hartsuiker, RJ.; Bastiaanse, R.; Postma, A.; Wijnen, F., editors. Phonological encoding and monitoring in normal and pathological speech. Hove: Psychology Press; 2005.

Oppenheim GM, Dell GS. Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. Cognition. 2008; 106:528–537. [PubMed: 17407776]

Özdemir R, Roelofs A, Levelt WJM. Perceptual uniqueness point effects in monitoring internal speech. Cognition. 2007; 105:457–465. [PubMed: 17156770]

Pailing PE, Segalowitz SJ. The error-related negativity as a state and trait measure: motivation, personality, and the ERPs in response to errors. Psychophysiology. 2004; 41:84–95. [PubMed: 14693003]

Postma A. Detection of errors during speech production: A review of speech monitoring models. Cognition. 2000; 77(2):97–131. [PubMed: 10986364]

Postma A, Kolk HHJ. The effects of noise masking and required accuracy on speech errors, disfluencies, and self repairs. Journal of Speech and Hearing Research. 1992; 35:537–544. [PubMed: 1608244]

Postma A, Kolk HHJ. The covert repair hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. Journal of Speech & Hearing Research. 1993; 36(3):472–487. [PubMed: 8331905]

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

Pritchard, WS.; Shappell, SA.; Brandt, ME. Psychophysiology of N200/N400: A review and classification scheme. In: Jennings, JR.; Ackles, PK.; Coles, MGH., editors. Advances in psychophysiology. Vol. 4. London: Jessica Kingsley; 1991. p. 43-106.

Rabbitt PMA. Error correction time without external error signals. Nature. 1966a; 212:438. [PubMed: 5970176]

Rabbitt PMA. Errors and error corrections in choice-response tasks. Journal of Experimental Psychology. 1966b; 71:264–272. [PubMed: 5948188]

Reason, J. Human error. Cambridge, MA: Cambridge University Press; 1990.

Riès S, Janssen N, Dufau S, Alario F, Burle B. General-Purpose Monitoring during Speech Production. Journal of Cognitive Neuroscience. 2011; 23(6):1419–1436. [PubMed: 20350170]

Roach A, Schwartz MF, Martin N, Grewal RS, Brecher A. The Philadelphia Naming Test: Scoring and Rationale. Clinical Aphasiology. 1996; 24:121–133.

Roelofs, A. Spoken word planning, comprehending, and self-monitoring: Evaluation of WEAVER++. In: Hartsuiker, RJ.; Bastiaanse, R.; Postma, A.; Wijnen, F., editors. Phonological encoding and monitoring in normal and pathological speech. Hove: Psychology Press; 2005.

Romani C, Olson A, Semenza C, Grana A. Patterns of phonological errors as a function of a phonological versus an articulatory locus of impairment. Cortex. 2002; 38(4):541–567. [PubMed: 12465668]

Ruchsow M, Gron G, Reuter K, Spitzer M, Hermle L, Kiefer M. Error related brain activity in patients with obsessive compulsive disorder and in healthy controls. Journal of Psychophysiology. 2005; 19(4):298–304.

Saffran JR, Newport EL, Aslin RN, Tunick RA, Barrueco S. Incidental language learning: Listening (and learning) out of the corner of your ear. Psychological Science. 1997; 8:101–105.

Saffran EM, Schwartz MF, Linebarger M, Martin N, Bochetto P. The Philadelphia Comprehension Battery. Unpublished test. 1988

Santesso DL, Segalowitz SJ, Schmidt LA. Error-related electrocortical responses are enhanced in children with obsessive compulsive behaviors. Developmental Neuropsychology. 2006; 29:431–445. [PubMed: 16671860]

Schade, U.; Laubenstein, U. Repairs in a connectionist language-production model. In: Kohler, R.; Rieger, BB., editors. Contributions to quantitative linguistics. Dordrecht: Kluwer; 1993.

Schlenck KJ, Huber W, Willmes K. "Prepairs" and repairs: Different monitoring functions in aphasic language production. Brain & Language. 1987; 30(2):226–244. [PubMed: 2436704]

Schwartz MF, Dell GS, Martin N, Gahl S, Sobel P. A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. Journal of Memory and language. 2006; 54(2): 228–264.

Sebastián-Gallés N, Rodríguez-Fornells A, de Diego-Balaguer R, Díaz B. First- and second-language phonological representations in the mental lexicon. Journal of Cognitive Neuroscience. 2006; 18:1277–1291. [PubMed: 16859414]

Slevc LR. Mechanisms of self-monitoring. Unpublished manuscript. 2006

Smith JL, Smith EA, Provost AL, Heathcote A. Sequence effects support the conflict theory of N2 and P3 in the Go/NoGo task. International Journal of Psychophysiology. 2010; 75(3):217–226. [PubMed: 19951723]

Stark, J. Aspects of automatic versus controlled processing, monitoring, metalinguistic tasks, and related phenomena in aphasia. In: Dressler, W.; Stark, J., editors. Linguistic Analyses of Aphasic Language. New York: Springer-Verlag; 1988.

Steinhauser M, Maier M, Hübner R. Modeling behavioral measures of error detection in choice tasks: Response monitoring versus conflict monitoring. Journal of Experimental Psychology: Human, Perception and Performance. 2008; 34:158–176. [PubMed: 18248146]

Steinhauser M, Yeung N. Decision processes in performance monitoring. Journal of Neuroscience. 2010; 30:15643–15653. [PubMed: 21084620]

Stemmer B, Segalowitz SJ, Witzke W, Schonle PW. Error detection in patients with lesions to the medial prefrontal cortex: An ERP study. Neuropsychologia. 2004; 42:118–130. [PubMed: 14615082]

Swick D, Turken AU. Dissociation between conflict detection and error monitoring in the human anterior cingulate cortex. Proceedings of the National Academy of Sciences, USA. 2002; 99:16354–16359.

Tent J, Clark JE. An experimental investigation into the perception of slips of the tongue. Journal of Phonetics. 1980; 8(3):317–325.

Thomas EAC. On a class of additive learning models: Error-correcting and probability matching. Journal of Mathematical Psychology. 1973; 10:241–264.

Ullsperger M, von Cramon DY. How does error correction differ from error signaling? An event-related potential study. Brain Research. 2006; 1105:102–109. [PubMed: 16483557]

van Meel CS, Heslenfeld DJ, Oosterlaan J, Sergeant JA. Adaptive control deficits in attention-deficit/hyperactivity disorder (ADHD): The role of error processing. Psychiatry Research. 2007; 151:211–220. [PubMed: 17328962]

van Sonderen JF, Gielen CCAM, Denier van der Gon JJ. Motor programs for goal-directed movements are continuously adjusted according to changes in target location. Experimental Brain Research. 1989; 78:139–146.

van Veen V, Carter CS. The timing of action monitoring in rostral and caudal anterior cingulate cortex. Journal of Cognitive Neuroscience. 2002; 14:593–602. [PubMed: 12126500]

van Wijk C, Kempen G. A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. Cognitive Psychology. 1987; 19(4):403–440.

van't Ent D, Apkarian P. Motoric response inhibition in finger movement and saccadic eye movement: A comparative study. Clinical Neurophysiology. 1999; 110:1058–1072. [PubMed: 10402093]

Vigliocco G, Hartsuiker RJ. The interplay of meaning, sound, and syntax in sentence production. Psychological Bulletin. 2002; 128(3):442–472. [PubMed: 12002697]

Warker JA, Dell GS. Speech errors reflect newly learned phonotactic constraints. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2006; 21:387–398.

Wheeldon LR, Levelt WJM. Monitoring the time course of phonological encoding. Journal of Memory & Language. 1995; 34(3):311–334.

Wheeldon LR, Morgan JL. Phoneme monitoring in internal and external speech. Language & Cognitive Processes. 2002; 17(5):503–535.

Wonnacott E, Newport EL, Tanenhaus MK. Acquiring and processing verb argument structure: Distributional learning in a miniature language. Cognitive Psychology. 2008; 56:165–209. [PubMed: 17662707]

Yeung N, Botvinick MM, Cohen JD. The neural basis of error detection: Conflict monitoring and the error-related negativity. Psychological Review. 2004; 111(4):931–959. [PubMed: 15482068]

Yeung N, Nieuwenhuis S. Dissociating response conflict and error likelihood in anterior cingulate cortex. Journal of Neuroscience. 2009; 29(46):14506–14510. [PubMed: 19923284]

- A new model of monitoring in speech production is proposed.
- Conflict in the production system is used as a signal for error detection.
- Predictions derived from computational simulations are tested on aphasic patients.
- Conflict-based model is supported over perceptual loop account of error detection.

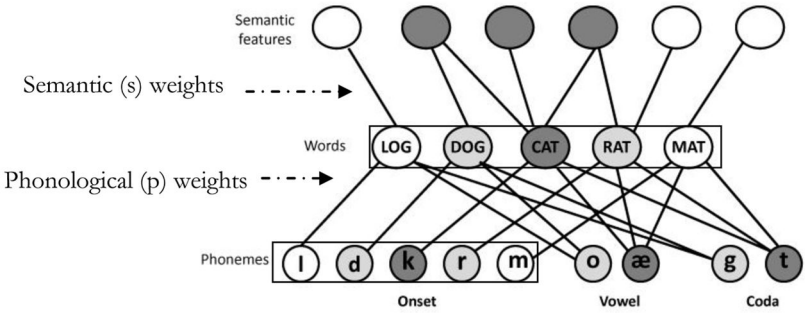

Figure 1.

The interactive two-step model of word production. Boxes indicate the places where conflict was measured, once at the word layer at the time of lemma selection (the upper box), and once at the phoneme layer at the time of onset selection (the lower box).

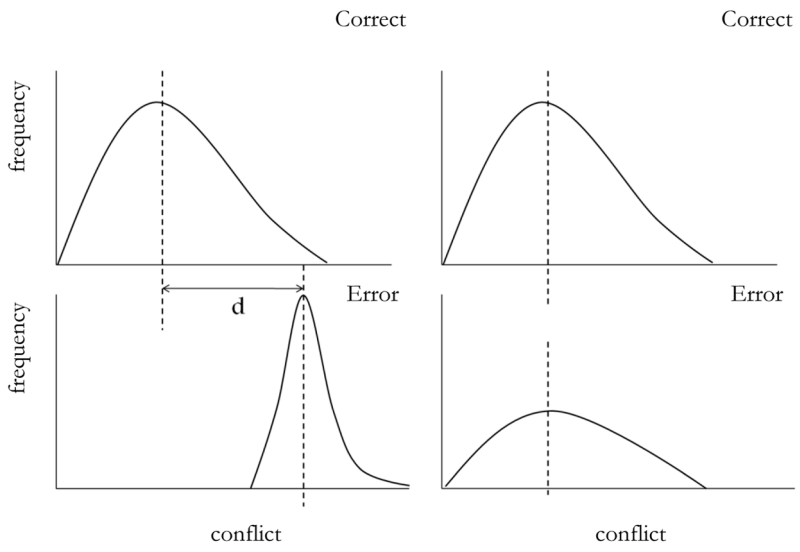

Figure 2.

Demonstration of the principle of detection sensitivity using hypothetical distributions. The graphs plot the number of trials showing a certain amount of conflict, and are categorized based on whether the trial ended in the correct or incorrect response. The two panels on the left show the case where detection sensitivity is high. If conflict correctly signals error occurrence, errors should on average show higher conflict compared to correct trials. On the other hand, if conflict is not a useful signal for error detection, the distribution of correct and incorrect responses should not differ with regard to measures of conflict (right panels). d = Cohen's d (see text).

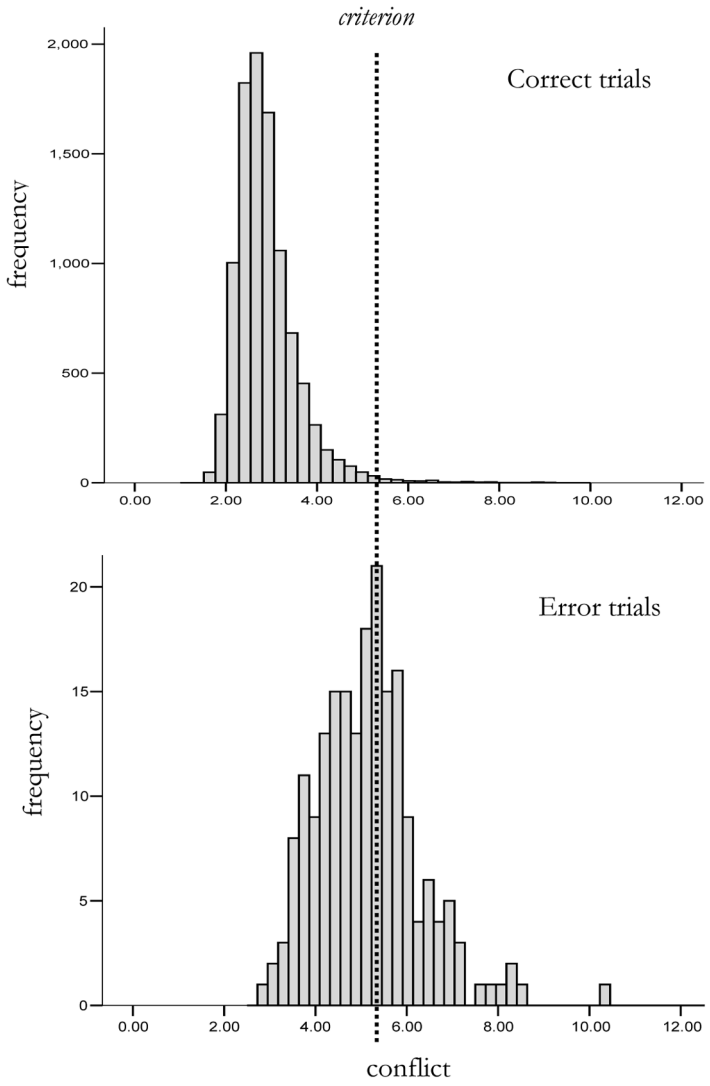

Figure 3.

Distribution of the *−ln(diff(max))* measure of conflict for semantic errors and correct trials in a simulated normal speaker. The dashed line indicates the criterion derived from Kac's (1962) model. The area to the right of the criterion consists of conflict values translated into error signals. This area indicates false alarms (≈ 1%) in the correct distribution (upper panel), and detected errors or hits (≈47%) in the error distribution (lower panel).

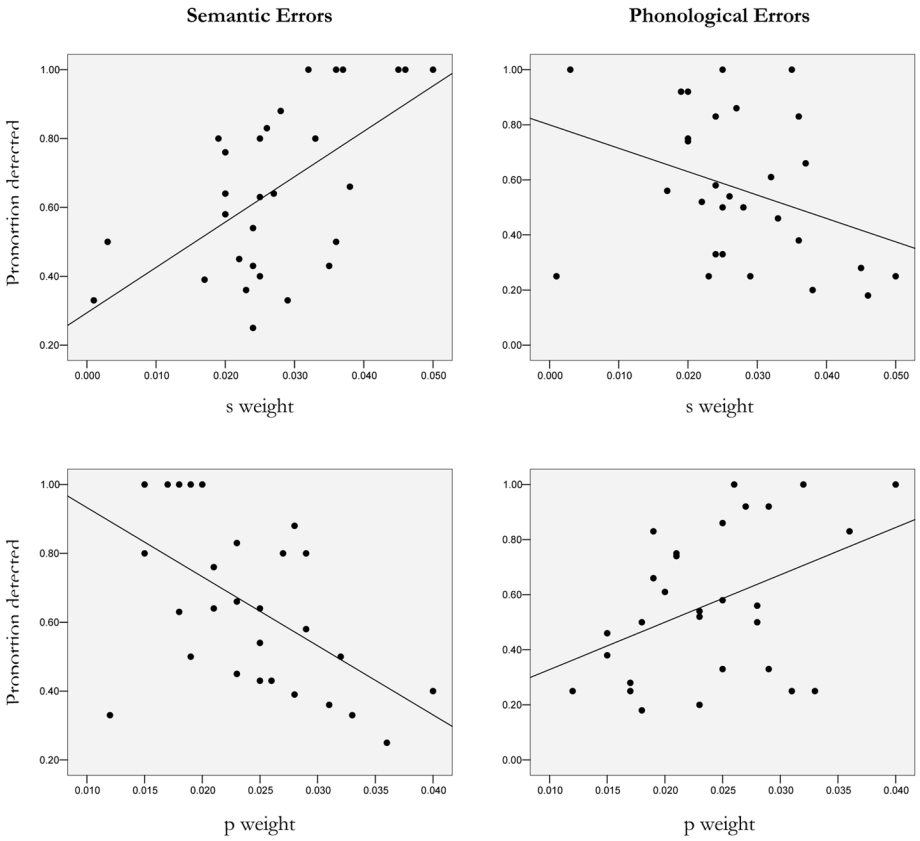

Figure 4.

Correlations between the production parameters and error detection in the 29 patients. The two left-hand panels show semantic error detection, and the two right-hand panels, phonological error detection. Stronger *s* weights are associated with better semantic error detection, and stronger *p* weights with better phonological error detection. The correlation between the *s* weights and phonological error detection is in the opposite direction, as is the correlation between the *p* weights and semantic error detection.

Table 1

Detection sensitivity and layer specificity in the 6 simulated speakers with normal to severely damaged production systems. The last two columns show the Cohen's d's for semantic and nonword errors at word and phoneme layers calculated using the $-ln(diff(max))$ measure of conflict, with the $-ln(sd)$-based values reported in the parentheses. The Bold numbers identify cases where there is some detection sensitivity. Note that as the weights decrease, detection sensitivity decreases as well. When the production system is completely distorted (s=p = 0.008) detection sensitivity is very low.

Weights		**level of conflict measurement**	**Cohen's d for Error type**	
s	***p***		**Semantic**	**Nonword**
0.04	0.04	Word layer	**3.26(1.18)**	0.57 (0.50)
		Phoneme layer	0.09 (0.11)	**2.83 (1.52)**
0.02	0.04	Word layer	**1.41 (0.76)**	0.55 (0.33)
		Phoneme layer	0 (0)	**3.05 (1.95)**
0.04	0.02	Word layer	**2.81 (1.18)**	0.05 (0)
		Phoneme layer	0.04 (0.04)	**1.42 (0.95)**
0.008	0.04	Word layer	0.37 (0.34)	0.03 (0.37)
		Phoneme layer	0.09 (0.20)	**3.58 (2.18)**
0.04	0.008	Word layer	**2.03 (0.82)**	0.29 (0.29)
		Phoneme layer	0.09 (0)	0.55 (0.43)
0.008	0.008	Word layer	0.25 (0.19)	0.25 (0.18)
		Phoneme layer	0.12 (0.04)	0.31 (0.26)

Table 2

Data of the individual patients. The *s* and *p* column show the weights of the production system. The next four columns are percentages of correct responses on the comprehension tests and the last two columns report the proportion of detected semantic and phonological errors with the total number of errors in parentheses.

Patients	*s*	*p*	Pyramids & Palm Trees	Synonym judgment	PPVT-III	Phonological discrimination	Proportion of semantic errors detected (Total)	Proportion of phonological errors detected (Total)
01	0.017	0.028	81	73	55	73	0.39 (18)	0.56 (16)
02	0.024	0.025	96	93	90	100	0.43 (21)	0.33 (12)
03	0.037	0.019	98	100	86	85	1 (1)	0.66 (25)
04	0.032	0.02	94	93	75	98	1 (7)	0.61 (23)
05	0.046	0.018	87	80	86	95	1 (1)	0.18 (11)
06	0.019	0.027	90	93	82	95	0.8 (41)	0.92 (13)
07	0.027	0.025	98	67	98	98	0.64 (11)	0.86 (21)
08	0.035	0.026	87	100	84	98	0.43 (7)	1 (6)
09	0.029	0.033	100	93	89	98	0.33 (9)	0.25 (4)
10	0.023	0.031	96	100	101	90	0.36 (11)	0.25 (8)
11	0.028	0.028	73	87	81	95	0.88 (8)	0.5 (8)
12	0.02	0.029	92	100	90	93	0.58 (19)	0.92 (12)
13	0.022	0.023	85	93	93	98	0.45 (18)	0.52 (27)
14	0.036	0.015	90	63	70	98	1 (1)	0.38 (45)
15	0.02	0.021	88	100	88	85	0.64 (28)	0.75 (20)
16	0.02	0.021	87	80	73	95	0.76 (29)	0.74 (23)
17	0.045	0.017	96	100	83	88	1 (2)	0.28 (18)
18	0.026	0.023	94	100	88	70	0.83 (6)	0.54 (13)
19	0.038	0.023	88	93	79	98	0.66 (3)	0.2 (5)
20	0.050	0.017	96	100	92	93	1 (2)	0.25 (12)
21	0.024	0.025	71	73	72	90	0.54 (13)	0.58 (7)
22	0.025	0.029	90	80	77	88	0.8 (20)	0.33 (6)
23	0.001	0.012	75	33	66	85	0.33 (3)	0.25 (16)
24	0.003	0.032	46	13	61	55	0.5 (30)	1 (12)
25	0.025	0.018	94	87	98	95	0.63 (8)	0.5 (20)

Patients	*s*	*p*	Pyramids & Palm Trees	Synonym judgment	PPVT-III	Phonological discrimination	Proportion of semantic errors detected (Total)	Proportion of phonological errors detected (Total)
26	0.024	0.036	85	53	79	93	0.25 (12)	0.83 (6)
27	0.025	0.040	92	100	87	93	0.4 (10)	1 (1)
28	0.036	0.019	90	100	78	78	0.5 (4)	0.83 (30)
29	0.033	0.015	81	73	46	53	0.8 (5)	0.46 (41)
Average	0.027	0.024	87.59	83.45	80.93	88.72	0.65 (12.00)	0.57 (15.90)

Table 3

Correlation between comprehension measures and detection of semantic and phonological errors for the 29 patients. Pearson's *r* is reported along with the relevant p-value in the parentheses. PPVT= Peabody Picture Vocabulary Test.

	Semantic error detection	Phonological error detection
Pyramids & Palm Trees	0.20 (p = 0.29)	−0.23 (p = 0.22)
Synonym Judgment	0.24 (p = 0.22)	−0.08 (p = 0.67)
PPVT-III	−0.03 (p = 0.88)	−0.04 (p= 0.83)
Phoneme Discrimination	0.02 (p = 0.91)	−0.15 (p = 0.44)