Journal of Memory and Language 111 (2020) 104069

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Repairing speech errors: Competition as a source of repairs

Sieb G. Nooteboom*, Hugo Quené

Utrecht University, the Netherlands

ARTICLE INFO

Keywords:
Speech errors
Self-monitoring
Internal speech
Overt speech
Repairing

ABSTRACT

This paper focuses on the source of self-repairs of segmental speech errors during self-monitoring. A potential source of repairs are candidate forms competing with the form under production. In the time interval between self-monitoring internal and overt speech, activation of competitors probably decreases. From this theory of repairing we derived four main predictions specific for classical SLIP experiments: (1) Error-to-cutoff times are shorter after single elicited errors than after other errors. (2) Single elicited errors are relatively more often detected than other errors, but more so after internal than after external error detection. (3) The correct form is the most frequent form used as repair, but more so for single elicited than for other errors. (4) Cutoff-to-repair times are shorter for single elicited than for other errors. A re-analysis of data formerly obtained in two SLIP experiments mainly supports the theory of repairing for multiple but not for single non-elicited errors.

Introduction

This paper is about repairing errors of speech. Errors of speech, speech errors or slips of the tongue are involuntary deviations from the intended form of utterances, often in the form of exchanges or other misplacements of speech sounds, morphemes or words. Some examples, taken from Fromkin (1973), are *a Tanadian from Toronto* instead of a *Canadian from Toronto*, *heft lemisphere* instead of *left hemisphere*, *a language needer learns* instead of *a language learner needs*, *take him to the lab first* instead of *take him to the lab last*. Speech errors have been studied at least since the late nineteenth century by phoneticians, linguists, psycholinguists, neurologists, neuropsychologists and psychoanalysts mainly because they provide a window on the mechanisms underlying speech. Given how complex these mechanisms are, human speech is amazingly fluent and errors of speech are relatively rare. Garnham, Shillcock, Brown, Mill, and Cutler (1982), counting speech errors in a corpus of spontaneous speech of 175.000 words, found one speech error in every 900 words. Rossi and Peter-Defare (1998), counting speech errors in a number of conversations lasting 45 min each, found that the frequency of speech errors varied over conversations from one per 680 words to one per 1700 words, with an average of one per 900 words.

Speakers not only make a speech error every now and then, they also regularly detect and repair their own speech error, as in *chew the fla…the fat with Slim* (also taken from Fromkin, 1973). Roughly 60% of segmental speech errors in spontaneous speech are detected and repaired by the speaker (Levelt, 1983; Nooteboom, 1980, 2005). The main question we attempt to answer in this paper is where such repairs come from. The answer, of course, depends on one's theory of "self-monitoring for speech errors", the assumed processes involved in the detection and repair of speech errors employed by speakers. For a survey of models of self-monitoring the reader is referred to Postma (2000) and to Nozari, Dell, and Schwartz (2011). A main distinction is between production-based and perception-based models of self-monitoring. Examples of production-based models of self-monitoring are provided by Laver (1980; see also Schlenck, Huber, & Wilmes, 1987; Van Wijk & Kempen, 1987), who assumed special purpose editors within the speech generation system, and MacKay (1987), who proposed that, because a speech error in some sense is a relatively new structure, it will cause prolonged activation of some node in the neural network generating speech; this prolonged activation will increase awareness and thereby lead to error detection. A different production-based mechanism for error detection is proposed by Nozari et al. (2011), to be discussed below.

For quite some time the most influential theory of self-monitoring was a perception-based theory, the so-called "dual perceptual loop theory" proposed by Levelt (1983; 1989) and Levelt, Roelofs, and Meyer (1999). This theory is part of a more wide-ranging theory of speech production. Levelt and colleagues distinguish between a module called CONCEPTUALIZER generating preverbal messages, a module called FORMULATOR containing both a submodule responsible for *grammatical encoding*, ordering lemmas delivered by a LEXICON, and a submodule responsible for *phonological encoding*, providing each lemma with a pronounceable word-like form. The output of the FORMULATOR is a "phonetic plan" (equivalent to "internal speech"). The "phonetic

* Corresponding author at: Cor Ruyslaan 20, 3584 GD Utrecht, the Netherlands.
E-mail address: s.g.nooteboom@uu.nl (S.G. Nooteboom).

https://doi.org/10.1016/j.jml.2019.104069
Received 9 February 2019; Received in revised form 28 October 2019; Accepted 9 November 2019
Available online 27 November 2019
0749-596X/

plan" goes two different ways (that is where the "dual" comes from): First, it forms the input to the same SPEECH COMPREHENSION SYSTEM that is also used for listening to other-produced speech, transforming this input into "parsed speech" that in its turn forms the input for a *monitor* that is part of or embedded in the earlier mentioned CONCEPTUALIZER. Second, the "phonetic plan" also forms the input for the ARTICULATOR generating overt speech that forms the input for a module called AUDITION, generating a "phonetic string" that, similarly to the "phonetic plan", is fed into the SPEECH COMPREHENSION SYSTEM, again leading to "parsed speech" fed into the *monitor*. When the *monitor* detects an error, the word form is re-compiled by going through the processes of speech preparation again.

A major property of this theory is that there are two stages where a speech error can be detected, viz. internal speech analyzed by the SPEECH COMPREHENSION SYSTEM before articulation starts and overt speech analyzed by AUDITION plus the SPEECH COMPREHENSION SYSTEM after articulation has started. That there are two stages of error detection was confirmed by Hartsuiker and Kolk (2001). They made a reasoned computational implementation of the dual perceptual loop theory, and demonstrated that this implementation could successfully simulate experimentally obtained distributions of error-to-cutoff times and cutoff-to-repair times, but only if both internal and external error detection were incorporated. Nooteboom and Quené (2017) demonstrated that, as one would predict from the dual perceptual loop theory, the distribution of error-to-cutoff times obtained in two experiments eliciting segmental speech errors under strictly temporally controlled conditions, is bimodal. The two peaks were found to be temporally separated by about 500 ms, suggesting that this is the average delay between error detection in internal and in overt speech. This finding implies that, at least statistically, repaired speech errors can be classified as detected in internal or in overt speech on the basis of the durations of error-to-cutoff times. This makes it possible to investigate differences between these two classes of repaired speech errors. In the investigation to be reported below we will capitalize on this possibility.

A problem with the dual perceptual loop theory is how a monitor incorporated in the CONCEPTUALIZER can detect a segmental error. One would expect that an error can be detected by comparing a candidate form containing the error with a correct version of the same form. But at the level of the CONCEPTUALIZER there are no word-like forms. Levelt and colleagues were forced into this somewhat awkward position by their assumption that in speech preparation there is no cascading of information from one module to the next. This means that each module generates an output containing, apart from very few exceptions, only a single unit for each slot in the string being generated. Thus, assuming that segmental errors are generated during phonological encoding, in the internal speech generated by phonological encoding there is no competition for example between an error form and the corresponding correct form, struggling for the same slot in the string of pronounceable word-like forms. Meanwhile, however, there is much evidence that in reality such competition between simultaneously activated forms struggling to fill the same slot, is frequent. This evidence comes from quite a few articulatory studies investigating articulatory gestures during experiments eliciting segmental speech errors (Frisch & Wright, 2002; Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; McMillan & Corley, 2010; Mowrey & MacKay, 1990; Slis & Van Lieshout, 2016). In such experiments articulatory blending of the error sound and the correct sound occurs frequently, in fact much more frequently than one can spot by listening to the speech produced: there are many whole or partial intrusions or deletions of articulatory gestures, but often these have no easily identifiable audible consequences (Pouplier & Goldstein, 2005). Although such articulatory blendings often cannot be heard as speech errors, it has been found that such deviations from error-free speech lead to longer reaction times in a phoneme identification task (Nooteboom & Quené, 2013). These results together show that cascading of information from the FORMULATOR to the ARTICULATOR is quite normal. This also implies that internal speech often contains multiple candidate forms competing for the same slot.

Nozari et al. (2011) proposed a production-based model of self-monitoring internal speech. Their model assumes that during speech preparation there can be conflict on a representational level between simultaneously active and competing options for a particular slot in the string being generated. In case no error is made, only one item has a high level of activation and no conflict signal is generated. When an error has been made, there may be multiple competing items with high levels of activation. In such cases conflict information is passed on to a domain general executive center. This idea was implemented in a two-step model of word production as earlier proposed by Dell, Schwartz, Martin, Saffran, and Gagnon (1997) and Foygel and Dell (2000). Simulations with the Nozari et al. (2011) extension showed that conflict detection is layer specific, i.e. separate for a semantic and a phonological representation, and that amount of conflict correlates well with error detection in human speakers. The model by Nozari et al. (2011) presupposes that both at the level of semantic features and at the level of phonological features multiple items generated by the speech production system may be active simultaneously, correct and incorrect competing for the same slot. This was confirmed by Nozari, Freund, Breining, Rapp, and Gordon (2016). They examined "selection control" by manipulating the overlap in either semantic or segmental features in a naming experiment, measuring reaction times. They also examined "post-monitoring control" in a reversal task both after semantic and after segmental overlap. Results support a model in which selection control operates separately at lexical and segmental selection stages, but post-monitoring control operates on the segmentally-encoded outcome. For our purposes it is relevant that these results confirm that there is potential competition, showing up in increased reaction times, between simultaneously active multiple forms, both during speech preparation and after speech is initiated.

An interesting feature of the model by Nozari et al. (2011) is that, when the overall conflict in the system increases, distinguishing correct and error trials becomes more difficult, and therefore error detection suffers. As it happens, the two experiments reported below together provide data to test this particular prediction. This is so because in Experiment 2, the overall conflict was higher than in Experiment 1. We will discuss this further in the results section of Experiment 2 and in the general discussion.

In the research described in this paper, we are only concerned with *segmental* speech errors, and not with lexical, semantic, syntactic or appropriateness errors. It has been pointed out to us that it is not self-evident that we describe the detection and repair of segmental errors in terms of processes involving pronounceable word or nonword forms, both correct and incorrect. One could imagine that the units involved in competition are segments. This is what we used to believe (cf. Nooteboom & Quené, 2008). In the current paper we are agnostic as to the role of segments versus pronounceable word-like forms in the initial generation of segmental speech errors (although it may be relevant that all segmental speech errors can be interpreted as blends of competing word forms). However, we found out that the detection of segmental errors in internal and overt speech is best described as the result of time-consuming scanning of word-like forms generated by the production system, from early to late (Nooteboom & Quené, 2019). As to repairing, we hardly ever observe single segments being used as repairs. Nearly always, repairs are recognizable word-like or morpheme-like forms and sometimes syllables. To us it seems most natural to describe the processes involved in detection and repair of segmental errors in terms of competing word-like forms. This we have first spelled out in Nooteboom and Quené (2017) and later more elaborately in Nooteboom and Quené (2019). In those two publications we inadvertently employed the term "lexical forms". Because non-lexical error forms, as generated by the speech preparation system, are involved, the term "lexical" was unfortunate. It should be made clear that our

candidate word forms are generated by what Levelt (1989) calls "phonological encoding".

Nooteboom and Quené (2017; 2019) proposed that after error detection and rejection in internal speech, no recompilation of a repair is necessary: generally, the competing candidate form with the next highest activation is immediately available as repair. Most often this most activated competitor candidate will be the correct form, because activation of the correct candidate will be sustained from the lexical level, in contrast to the activation of other, erroneous candidates. Nooteboom and Quené (2017; 2019) also proposed that during the time delay of about 500 ms between error detection in internal and in overt speech, the competing candidate form remains active, but with a gradual decrease in its activation. Provided it is possible to classify repaired speech errors as detected in internal or in overt speech, we may derive from this proposal some testable predictions. In order to test these predictions, we need a set of repaired segmental speech errors elicited under strictly temporally controlled conditions. Two experiments reported in Nooteboom and Quené (2017) provide such a set of repaired speech errors. In Nooteboom and Quené (2017) the repairs were not analyzed. This will be done in the current paper. In that 2017 paper, only single interactional speech errors, mostly in word initial position, were elicited and analyzed. In addition to those errors elicited by the SLIP task (see below), however, many more errors were also made in that study but were not analyzed at the time. Here we will also analyze interactional single segmental errors in other than the targeted positions, i.e. non-elicited errors, and multiple errors that of course contain at least one non-elicited error. The main opposition we are creating in this way is between elicited and non-elicited errors.

The reason for doing this is that from our proposed mechanism for repairing, different predictions can be derived for these different categories of errors. It has been pointed out to us that in doing so, possibly we conflate the often observed special role of word initial segments in segmental speech errors with our experimental opposition between elicited speech errors and other speech errors. This problem will be taken care of in the second experiment to be described here, in which we elicited speech errors not only against segments in initial position, but also against vowel segments which were always in second position. It has also been pointed out to us that preferably we should compare single elicited errors with single non-elicited errors, and not with multiple errors. In the experiments to be described below, we keep single and multiple non-elicited errors separate in the analyses, although this could result in statistical power problems. It has been suggested to us that we should focus on the single other errors, and skip the multiple errors. The reason that we still include the multiple errors, is that they provide an interesting test case for our theory: Naïvely, one would expect multiple errors to be detected faster than single errors, because they exhibit stronger deviations from the correct target. But our theory predicts that they will be detected slower than single errors, because more candidate forms are involved in the generation of multiple errors than of single errors.

The various predictions we make are related to how the SLIP task works. In this task as implemented in Nooteboom and Quené (2017), word pairs, in this case CVC monosyllables, are presented on a screen. Each stimulus word pair, i.e. a word pair that is to be spoken aloud, is preceded by five precursor word pairs, the last three of which prime a reversal between the two targeted segments, as in *bouw jool, lijf deed, koet pop, kuur poet, kas piet*, precursors of the stimulus word pair *paf kiep*. After the stimulus word pair, a sequence of *??????* is shown on the screen as a cue that the last word pair seen has to be spoken as soon as possible. The cue to speak is followed by the Dutch word for *repair?* plus a question mark, in order to elicit repairs. The time interval between all successive word pairs, between the last word pair and the cue to speak, and between the *??????* cue to speak and the *repair?* cue is in these experiments always 1000 ms. The precursors boost the activation of those competitor word candidates that have segments exchanged with the target words (e.g. of *kaf* and *piep*), and thus they lower the relative activation of all other potential competitor candidates. Of course, the precursors also provide extra activation to the correct target forms because of the segmental overlap between precursors and the correct forms. We therefore assume that when elicited errors occur, the main competition has been between the correct form and the error form elicited by the SLIP task. Nevertheless, in a SLIP task non-elicited errors also occur frequently, and this suggests that in those cases more than two activated candidates are in competition. It is reasonable to expect that if there is major competition between only two highly activated candidates, then this competition is resolved in favor of one of the two candidates more rapidly than if there is major competition between more than two less activated candidates. Following Seyfeddinipur, Kita, and Indefrey (2008) and Tydgat, Stevens, Hartsuiker, and Pickering (2012), we also assume that interruption (a.k.a. "cutoff") is often postponed until after a repair has come available, i.e., until after the competition between error form and candidate repair has been resolved. This leads to our first prediction:

(1) Error-to-cutoff times are longer after single non-elicited and after multiple errors than after single elicited errors.

Note that this prediction is not self-evident from the general viewpoint of detecting differences. From that viewpoint one would expect that the greater the difference between two items, for example word-like forms, the faster detection is. If our first prediction is borne out by the data, this would be strongly in favor of our proposed theory.

If indeed competition is resolved more rapidly for single elicited errors than for other errors, this implies that for other errors more often than for single elicited errors, the time available for self-monitoring internal speech expires before the competition has been resolved. In those cases the chances are that error detection and repairing are shifted to overt speech. This leads to our second prediction:

(2) Single other errors and multiple errors are detected in internal speech relatively less often than single elicited errors are.

One may note that this second prediction is an immediate consequence of the first prediction plus the assumption of two consecutive, temporally separated, stages of self-monitoring.

A major, and as far as we know new, assumption of our theory of repairing speech errors is that activation of competing candidates decreases during the time delay between detection in internal and in overt speech. We also have assumed that in most cases in which an error has been made during segmental encoding, the competitor with highest activation is the intended correct pronounceable word-like form. This is so because of our assumption that activation of the correct form is sustained from the lexical level, whereas the activation of other error forms is not. If these assumptions are correct, we assume that both after internal and after external error detection, the most frequent repair is the correct form, but, because of the decreasing activation, less so after external than after internal error detection. Also, one would expect that this predicted effect is particularly strong for single elicited errors, because there the main competition is between error form and correct form. For other errors, the advantage of correct candidate forms is much less evident, because in those cases the main competition, given the error, must have been between the error form and another error form. This gives our prediction 3:

(3) (a) Incorrect forms are used less often as repairs than correct forms are, and (b) this effect is less strong after external detection than after internal detection.

We have assumed that after other errors there are often more competitors, which are also less activated, than after single elicited errors. Thus, one would expect that in selecting a repair competition generally is resolved more rapidly after single elicited errors than after

other errors. In those cases in which interruption is postponed until a repair is available and cutoff-to-repair times have a duration of 0 ms, this difference between single elicited and other errors will, of course, not affect the cutoff-to-repair times. But we know that cutoff-to-repair times of 0 ms, so-called immediate repairs, are relatively rare. For the cases in which the cutoff-to-repair times have a measurable duration, we would expect shorter cutoff-to-repair times after single elicited than after other errors. But because activation of candidates for repair decreases during the time interval between detection in internal and in overt speech, the differences between competing activated repair candidates also decrease between the two stages of error detection. Therefore the predicted difference between repairs of single elicited and other errors would decrease in the time interval between detection in internal and in overt speech, leading to our prediction 4.

(4) (a) Cutoff-to-repair times are shorter for single elicited than for single other errors and multiple errors, and (b) this effect is weaker after error detection in overt speech than after error detection in internal speech.

These four (composite) predictions will be tested against speech errors made in two experiments as described in Nooteboom and Quené (2017). The reader might argue that our predictions were made after we had done the experiments, and therefore they may not be really predictions. However, we wish to point out that the frequencies of single other errors and of multiple errors were unknown to us until after we had made these predictions.

Experiment 1

This experiment was originally set up to investigate temporal aspects of internal and external detection and repair of segmental speech errors elicited with the SLIP technique, both with and without auditory feedback (Nooteboom & Quené, 2017). Results showed among other things (1) that the error-to-cutoff times of the elicited segmental errors had a bimodal distribution, the two peaks being separated by some 500 ms, (2) that the error-to-repair times (of course including the error-to-cutoff times) also had a bimodal distribution, the two peaks being separated by some 700 ms, (3) that the frequency of both internal and external error detection was not affected by the absence of auditory feedback. In the experiment many errors not elicited by the SLIP technique and many multiple errors were made. As these were not analyzed in Nooteboom and Quené (2017), they will be analyzed below, in an attempt to test our four predictions.

Method of Experiment 1

For a detailed description of the Method used in Experiment 1, we refer to Nooteboom and Quené (2017). Here we will briefly mention the main aspects of the method.

Speakers

There were 106 participating speakers, their average age being 23 years and 85 of them being female. All speakers were native speakers of Dutch and were paid for their participation.

Materials

There were two lists of stimulus items. Each item consisted of a pair of CVC words. In each list there were 32 test stimuli, 16 with the two initial consonants differing in one feature, either place or manner of articulation, and 16 with the two initial consonants differing in more than one feature. Each test stimulus was preceded by 5 precursor CVC CVC stimuli, the last 3 of which primed a reversal of the two initial consonants, as in *bouw jool, lijf deed, koet pop, kuur poet, kas piet*, precursors of the stimulus word pair *paf kiep*. There were also 23 filler stimuli in each list. These were preceded with a number of precursors varying between 0 and 4. These precursors did not prime a segmental reversal. After each test stimulus and each filler stimulus a sequence of *?????* was presented, as a cue to speak aloud the last word pair seen. After the *?????* there followed a presentation of the Dutch word for "repair?", to elicit sufficient repairs.

Procedure

Each speaker was tested individually in a sound-treated booth. The experiment was computer-controlled. The presentation of precursors, stimuli, *?????* and *repair?* cues always lasted 900 ms followed by a blank interval of 100 ms. Each speaker was presented with the two lists of stimuli, one with and one without auditory feedback, the order of the feedback conditions varying from one speaker to the next.

Scoring

All responses to all test stimuli were transcribed in orthography, or, where necessary, in phonetic transcription by the first author using the *Praat* computer program (Boersma & Weenink, 2016). For the current purpose responses were categorized as follows:

1. Fluent and correct responses of the type *bad game > bad game.*
2. Hesitations and omissions.
3. Single elicited segmental speech errors, i.e. completed reversals as *bad game > gad bame*, or completed anticipations as *bad game > bad bame*, interrupted elicited reversals or anticipations as *bad game > ga..bad game*, completed elicited perseverations as *bad game > bad bame*, and interrupted elicited perseverations as *bad game > bad b..bad game*.
4. Completed and interrupted other speech errors, i.e. non-elicited single segmental errors such as *bad game > bam game*, *bad game > bam..bad game*.
5. Completed or interrupted multiple errors such as *bad game > bam bame > bam b..bad game.*

All speech errors, except omissions, were also categorized as to whether the error was or was not repaired, and repairs were noted down.

Reliability of scoring

In order to assess the reliability of the transcription and categorization of responses, 11 out of 106 participants (about 10%) were selected at random, and all 704 responses to test stimuli provided by these 11 participants were transcribed and categorized independently by the second author as well. The two transcriptions did not match perfectly in 14 out of 704 cases (2.0%). Of these 14 discrepant cases, there were only 4 cases (0.6%) in which the difference was non-trivial (e.g. stimulus *vol teer*; response transcribed once as correct and fluent "vol teer" and once as elicited error "tol veer"). For the remaining 10 responses, the discrepancy was trivial (e.g. stimulus *peus kor*; response transcribed once as multiple error "keu poch" and once as multiple error "keuf poch"); 9 of these responses would be categorized identically with either transcription, and only 1 would end up in a different category (single vs multiple other error). These very low rates of divergence indicate that the transcriptions and classifications of the first author were indeed sufficiently reliable, and these were therefore used for further analysis.

Results of Experiment 1

A first breakdown of the observed speech errors is given in Table 2.1. In our further analyses we will mainly focus on the 859 single elicited, single other and multiple speech errors, 182 of which were repaired.

Interestingly, there are many more single elicited speech errors, and these are repaired far more often, than both either single other speech errors or multiple errors. This we had expected, because when single

Table 2.1
Numbers of responses broken down by response category and repair status, with percentages of repaired error responses.

Response category	Repair status		Total	%repaired
	Not repaired	Repaired		
Fluent and correct	5821	0	5821	0
Hesitations and omissions	64	40	104	38
Single elicited errors	298	115	413	28
Single other errors	187	31	218	14
Multiple other errors	192	36	228	16
Total	6562	222	6784	

segmental errors are successfully elicited by the SLIP technique, very likely the main competition is limited to the elicited error and the correct target, whereas in the case of single other and multiple errors presumably there is competition between more than two candidate forms. It is noteworthy that the percentages repaired are roughly equal for single other and multiple errors. One could have thought that multiple errors would be much more often repaired than single other errors, because in multiple errors there is a much stronger deviation from the correct form. This appears to be compensated, however, by the predicted relative degree of activation and number of the competitors, induced by the SLIP task. The reader may also note that in Table 2.1 in the row for hesitations and omissions there are 40 “repaired” cases. These are all hesitations of the type “bak zoon bak zoon”, “ba..bakzoon” or “ba..zak boon” for the stimulus *bak zoon*, where the initial response does not contain a speech error. These cases show that correct responses too can be subject to “repairing”, either with the correct target or with a competing incorrect form.

In the experiment we had stimulus word pairs where the interacting segments were either phonologically similar (one feature difference) or dissimilar (more than 1 feature difference), with equal numbers of stimuli. One expects the latter to be repaired more often than the former (cf. Nooteboom & Quené, 2008). This was confirmed by the numbers (and proportions) of repairs, as is evident from Table 2.2.

In Table 2.2 single elicited and other errors are collapsed. Possibly, the high percentages repaired are caused by the circumstance that all elicited errors, that is the majority of errors, are in initial position. We will come back to this when describing the results of Experiment 2. There, there was one contrast, between interacting vowels, not in initial but in second position. A loglinear analysis of the frequencies in Table 2.2 confirmed that there were fewer errors for dissimilar items than for similar items (in the loglinear model, this shows up as a main effect of similarity, $\beta = -0.385, Z = -3.63, p < .001$), and that there were relatively more repairs for dissimilar items than for similar items (in the loglinear model, this shows up as an interaction of similarity and repair status, $\beta = +0.657, Z = 3.00, p < .001$). (Detailed results of all regression models, including models having noise condition and session number as predictors, are reported in the Supplementary Materials online).

Because we made different predictions for repaired errors detected in internal speech (internally) and repaired errors detected in overt speech (externally), we analyze the distribution of error-to-cutoff times

Table 2.2
Numbers of responses broken down by phonological similarity and repair status, with percentages of repaired error responses.

Phonological similarity	Repair status		Total	%repaired
	Not repaired	Repaired		
Similar (1 feature)	403	81	484	17
Dissimilar (2 features)	274	101	375	27
Total	677	182	859	21

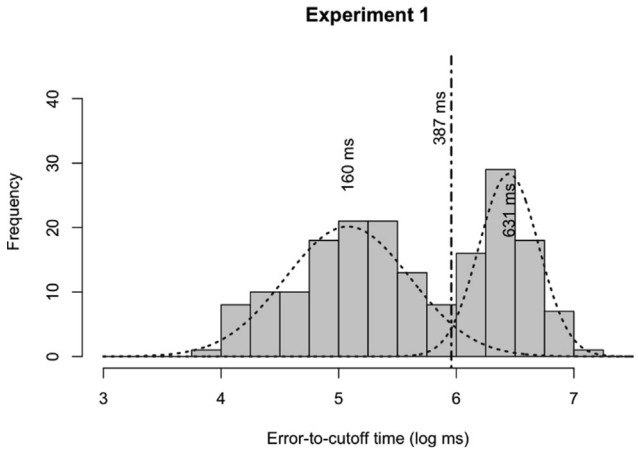

Fig. 2.1. Histogram of log-transformed durations of error-to-cutoff intervals in Experiment 1, for $N = 182$ repaired errors. Durations plotted with dotted lines indicate the estimated distributions from an uninformed gaussian mixture model (see text). The vertical dashed line indicates the interpolated boundary value (at 5.958 log ms, corresponding to 387 ms) between the two distributions.

of all repaired speech errors in the categories 'single elicited', 'single other' and 'multiple' together. Fig. 2.1 gives the relevant histogram.

The distribution of error-to-cutoff times is indeed bimodal, as confirmed by unsupervised clustering (Fraley & Raftery, 2002; Scrucca, Fop, Murphy, & Raftery, 2017) using R (R Core Team, 2019), and the distribution can be adequately described as being composed of two underlying gaussian distributions, one with a peak at 160 ms and one with a peak at 631 ms. The vertical dashed line in Fig. 2.1 corresponds to the best fitting separation of the two estimated distributions. Following Nooteboom and Quené (2017), we propose that the gaussian distribution with the shorter durations corresponds to the class of repaired errors that were detected in internal speech and the gaussian distribution with longer durations to the class of errors detected in overt speech. Below we will consider error-to-cutoff times shorter than 387 ms as corresponding to errors detected internally and those longer than 386 ms as corresponding to errors detected externally. Of course, because the two distributions overlap, there is some unavoidable statistical noise in this classification, but it seems good enough for our purposes.

So now we are in a position to test our four predictions. Our first prediction is:

(1) Error-to-cutoff times are longer after single non-elicited errors and after multiple errors than after single elicited errors.

The reader may remember that we made this prediction because we have assumed that in single elicited errors competition is mainly between the error form and the correct target form, whereas in other (single or multiple) errors there is competition between more than two forms, viz. the target form, the elicited error form, and at least one other form. Thus during self-monitoring of other errors more time would be needed than with elicited errors to resolve competition and to decide that an error has been made. In order to test this prediction, log-transformed error-to-cutoff times of the three categories of errors were compared using linear mixed-effects models (LMM), with participants as random intercepts (Hox, Moerbeek, & Van de Schoot, 2017; Quené & Van den Bergh, 2008), and with the error category as a fixed effect (see Table 2.1; omitting fluent and correct responses and hesitations and omissions, for which no valid error-to-cutoff time is available, and using single elicited errors as baseline category). [Here and in subsequent analyses, models with more complex random structures were over-specified and failed to converge; models with more complex fixed parts, including noise condition (0 = no noise, 1 = with noise), session

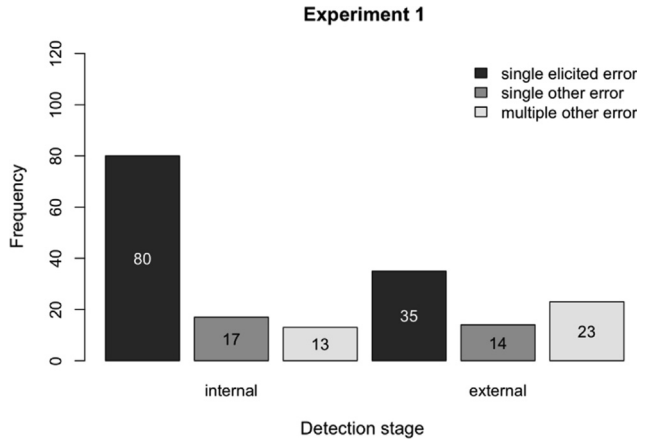

Fig. 2.2. Frequencies of single elicited, single other and multiple repaired errors in Experiment 1, separately for "internal" and "external" error detection.

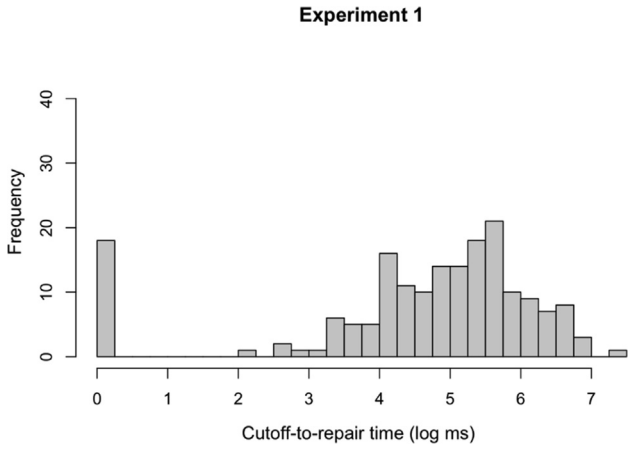

Fig. 2.3. Histogram of log-transformed cutoff-to-repair times in Experiment 1, with values of 0 ms converted to 1 ms before log transformation.

number (1, 2), and their interactions with response type, were also explored and are reported in the Supplementary Materials. For clarity we do not report those latter models here. Details of all regression models are however reported in the Supplementary Materials online; Lüdecke, 2019]. Error-to-cutoff times were found to be not significantly different between single elicited errors (mean 222 ms) and single other errors [mean 249 ms; $\beta = +0.115$, $t < 1$, n.s., 95%CI = $(-0.198, +0.394)$], but error-to-cutoff times were indeed significantly longer after multiple errors [mean 440 ms; $\beta = +0.683$, $t = 4.79$, $p < .0001$, 95%CI = (0.406, 0.991)]. This confirms our prediction (1) for multiple errors but not for single other errors.

That there is no significant difference between single elicited and single other errors, may be caused by the circumstance that distributions of error-to-cutoff times are far from normal, because these distributions include both internally and externally detected errors (see Fig. 2.1). This consideration leads to our second prediction:

(2) Single other errors and multiple errors are detected in internal speech relatively less often than single elicited errors are.

Fig. 2.2 gives the relevant breakdown of the data.

In order to test this prediction, the frequencies summarized in Fig. 2.2 were compared using a loglinear model [the low numbers of observations do not allow a GLMM with random effects], with response category (cf. Table 2.1, again using single elicited errors as baseline category) and detection stage (internal, external; using internal detection as baseline) as two predictors. Single elicited errors (baseline) were detected externally less often than internally (35:80, $\beta = -0.827$, $Z = -4.08$, $p < .001$). In a loglinear model, the predicted effect of detection stage on response category shows up as an interaction effect of these two predictors, with frequency being the dependent variable. Single other errors were also detected externally less often than internally (14:17), and this effect was, against prediction, relatively the same as for single elicited errors: interaction $\beta = +0.633$, $Z = 1.53$, $p = .126$. Multiple (other) errors were detected externally relatively more often than internally (23:13), and these were, as predicted, detected externally more often than single elicited errors were: interaction $\beta = +1.397$, $Z = 3.48$, $p < .001$. This confirms our prediction (2) for multiple errors but not for single other errors.

Our third prediction was:

(3) (a) Incorrect forms are used less often as repairs than correct forms are, and (b) this effect is less strong after external detection than after internal detection.

In order to test this prediction, the frequencies of correct and incorrect repairs were compared using a loglinear model [the low numbers of observations do not allow a GLMM with random effects], with detection stage (internal, external; using internal detection as baseline), and repair form (correct, incorrect; using correct repair as baseline) as two predictors. As predicted, for errors detected *internally*, repairs were mostly correct (81:29), and incorrect repairs were observed significantly less frequently ($\beta = -1.027$, $Z = -4.75$, $p < .001$). For errors detected *externally*, this effect was indeed significantly weaker (41:31; $\beta = +0.748$, $Z = 2.32$, $p = .020$). This confirms predictions (3.a) and (3.b).

Our fourth and last prediction was:

(4) (a) Cutoff-to-repair times are longer for single other errors and multiple errors than for single elicited errors, and (b) this effect is weaker after external detection than after internal error detection.

Fig. 2.3. is a histogram of all log-transformed cutoff-to-repair times (not to be confused with the error-to-cutoff times summarized in Fig. 2.1) in this experiment.

Fig. 2.3 shows that the distribution of log-transformed cutoff-to-repair times deviates strongly from normal, mainly because immediate repairs, having a cutoff-to-repair times of 1 ms (converted from 0 ms to allow log transformation) are overrepresented, with 18/182 observations. In line with our fourth prediction, we first tested whether (a) relatively fewer of these *immediate* repairs occurred after single other errors and after multiple errors than after elicited errors, and (b) whether this effect was weaker for errors detected externally. The odds of immediate repair for elicited errors were 14:101, and these odds were indeed lower for single other errors (2:29, $\beta = -1.248$, $Z = -5.92$, $p < .001$) and for multiple errors (2:34, $\beta = -1.089$, $Z = -5.49$, $p < .001$), but these effects were not significantly different for errors detected internally or externally [LRT: $\chi^2(2) < 2$, n.s.]. This confirms prediction (4.a) but not (4.b) with regard to the occurrence of immediate repairs. Second, we tested the predicted effects on the log-transformed cutoff-to-repair times of *non-immediate* errors only (having cutoff-to-repair times greater than 1 ms), using an LMM with participants as random intercepts, and using the response category (again using single elicited errors as baseline category, cf. prediction 1 above) and detection stage (internal, external; using internal detection as baseline) as two fixed predictors. For internally detected errors, excluding those with immediate repairs, we find that cutoff-to-repair times after elicited single errors (baseline) are 94 ms on average. After single other errors, the cutoff-to-repair times are indeed significantly longer [mean 300 ms, $\beta = 1.157$, $t = 4.88$, $p < .001$, 95%CI (0.698, 1.616)], as they are after multiple errors [mean 173 ms, $\beta = 0.605$, $t = 2.34$, $p = .011$, 95%CI (0.105, 1.105)]. For *externally*

detected errors, excluding those with immediate repairs (single elicited: 161 ms), the effect of error category was significantly weakened for single other errors [238 ms; interaction $\beta = -0.771$, $t = -2.10$, $p = .020$, 95%CI $(-1.480, -0.062)$] but not for multiple errors (357 ms; interaction $\beta = 0.188$, $t < 1$, n.s.). This confirms our prediction (4.a), and confirms (4.b) for single other errors but not for multiple errors.

Discussion of Experiment 1

We tested four predictions derived from a theory of repairing segmental speech errors in which it is assumed that during speech preparation and during self-monitoring more than one candidate repair is competing for the same slot in the utterance. The predictions were made specifically for a SLIP task eliciting reversals of the initial consonants in pairs of CVC words, thereby boosting the activation of both the correct target forms and the specific elicited errors, but not of other errors.

Assuming that competition between highly activated forms is more easily resolved than competition between more than two less activated forms, and that error-to-cutoff time at least partly reflects the time needed for conflict resolution, we predicted that error-to-cutoff times are longer after both single other errors and multiple errors than after single elicited errors. This was confirmed for multiple errors but not for single other errors. We have assumed that the absence of a significant difference here might have been due to the strongly non-normal distributions of the error-to-cutoff times and/or lack of statistical power. However, if indeed single other errors have longer error-to-cutoff times than single elicited errors, the time available for internal detection will more often be exceeded for single other than for single elicited errors, and then detection shifts to the external stage. This led to our prediction (2) that single other and multiple errors are relatively less often detected in internal speech than single elicited errors are. This second prediction was borne out for multiple errors but not for single other errors. Apparently self-monitoring internal speech takes more time for multiple errors than for single elicited errors, but there is no significant difference between single elicited and single other errors. In this respect our second prediction is not borne out, possibly as a result of lack of statistical power. That error-to-cutoff times are significantly longer for multiple errors than for single elicited errors suggests that the hypothesized slowing effect of the difference in number of competing candidates overrides the speeding effect of phonetic distance. We interpret this as evidence that during self-monitoring there is competition between candidate word-like forms generated by the production system and that the number of competitors is a major determinant of temporal aspects of self-monitoring.

We also predicted that (3.a) incorrect forms are used less often as repairs than correct forms are, and (3.b) this effect is less strong after external detection than after internal detection. Both parts of this predicted were confirmed. This supports our assumption that activation of correct word form candidates is sustained from the lexical level, and this activation decreases during the delay between self-monitoring internal and overt speech.

Our fourth prediction was that (4.a) cutoff-to-repair times are longer for other single and multiple errors than for single elicited errors, and (4.b) that this effect is weaker after external detection than after internal error detection. This prediction was tested separately for immediate and nonimmediate repairs. The odds of immediate repairs were lower for both single other and multiple errors than for single elicited errors, but there was no interaction with detection stage. For non-immediate repairs we found that cutoff-to-repair times of both single other and multiple errors were significantly longer than those of single elicited errors. For single other errors, but not for multiple errors, the effect was significantly weakened after external detection. Based on the different patterns of results for immediate and non-immediate repairs, we assume that immediate repairs may be qualitatively different in their causes and effects from non-immediate repairs. Apparently, every now and then a repair does not have to be re-activated but is immediately available at interruption. In all other cases the times needed for re-activation form a single gaussian distribution. Re-activation of a repair takes more time after single other and after multiple errors than after single elicited errors. This effect is weaker after single other errors detected externally, but not after multiple errors detected externally.

Experiment 2

Experiment 2 is very similar to Experiment 1. This Experiment too was originally set up to investigate temporal aspects of internal and external detection and repair of segmental speech errors elicited with the SLIP technique, both with and without auditory feedback (Nooteboom & Quené, 2017). The main difference is that not only CVC CVC word pairs were used in which the initial consonants differed in place *and/or* manner of articulation, as in Experiment 1, but also word pairs were used in which the initial plosive consonants differed in the voiced-unvoiced distinction, and word pairs differing in the vowels. Again results showed among other things (1) that the error-to-offset times of the elicited segmental errors had a bimodal distribution, the two peaks being separated by about 500 ms, (2) that the error-to-repair times also had a bimodal distribution, the two peaks being separated by some 700 ms, (3) that the frequency of both internal and external error detection was not affected by the absence of auditory feedback. Also in Experiment 2 the many errors that were not specifically elicited by the SLIP technique, and that were not analyzed in Nooteboom and Quené (2017), will be analyzed below, in an attempt to test our four predictions.

Method of Experiment 2

For a detailed description of the Method used in Experiment 2, we refer to Nooteboom and Quené (2017). Here we will again briefly mention the main aspects of the method.

Speakers

There were 124 participating speakers, their average age being 23 years, and 103 of them being female. All speakers were native speakers of Dutch and were paid for their participation.

Materials

There were two lists of stimulus items. Each item consisted again of two CVC words. In each list there were 32 test stimuli eliciting interactions between initial consonants differing in place and/or manner of articulation, of which 16 with the two initial consonants differing in one feature, either place or manner of articulation, and 16 with the two initial consonants differing in more than 1 feature. Also each list contained 16 stimuli eliciting interactions between plosive consonants differing only in the voiced-unvoiced feature, and 16 stimuli eliciting interactions between vowels. In Experiment 2 there were not 23 but 46 filler stimuli, with a number of precursors varying between 0 and 4. The precursors of the fillers did not prime interactions. Further details of the materials, the procedure and the scoring were the same as in Experiment 1.

Reliability of scoring

In order to assess the reliability of the transcription and categorization of responses, 12 participants (about 10%) were selected at random, and all 1536 responses to test stimuli provided by these 12 participants were transcribed and categorized independently by the second author as well. The two transcriptions did not match perfectly in 29 out of 1536 cases (1.9%). Of these 29 discrepant cases, there were 14 cases (0.9%) in which the difference was non-trivial (e.g. stimulus *puik bof*; response transcribed once as elicited error “buik pof” and once as fluent and correct “puik bof”). For the remaining 15 responses, the

Table 3.1
Numbers of responses broken down by response category and repair status, with percentages repaired error responses.

Response category	Repair status		Total	%repaired
	Not repaired	Repaired		
Fluent and correct	13,069	0	13,069	0
Hesitations and omissions	228	67	295	23
Single elicited errors	956	184	1140	16
Single other errors	570	35	605	6
Multiple other errors	473	34	507	7
Total	15,296	320	15,616	

discrepancy was often trivial (e.g. stimulus *beur poos*; response transcribed once as multiple error "peul bool" and once as multiple error "peur bool"); 8 of these responses would be categorized identically with either transcription, and 7 would end up in a different category. These low rates of divergence indicate that the transcriptions and classifications of the first author were indeed sufficiently reliable, and these were therefore used for further analysis.

Results of Experiment 2

A first breakdown of the data obtained in Experiment 2 is given in Table 3.1. The reader may observe that again, as in Experiment 1, there are a number of repaired hesitations. We also see again that single other and multiple errors are much less often repaired than single elicited errors. The percentage repaired of single elicited errors is much lower in Experiment 2 than in Experiment 1. The explanation for this phenomenon is suggested by the numbers in Table 3.2.

In this experiment four phonetic contrasts were used. The number of stimuli was the same for each contrast, yet the number of speech errors and the percentages repaired are not. Most conspicuously, the voiced-voiceless contrast leads to many more errors than the other contrasts. This is related to the weak position of this contrast in Dutch: Voiced and voiceless plosive consonants are very similar (Van Alphen & McQueen, 2006; Van Alphen & Smits, 2004; Van Alphen, 2004). This implies that Dutch initial plosives only differing in voicing are easily confused, not only in perception but in phonological encoding too, leading to relatively many speech errors. Also, this implies for self-monitoring that the conflict between voiced and voiceless initial plosives is less than the conflict between other initial consonants, and therefore relatively less speech errors against voicing are predicted to be detected. Furthermore, it is predicted from a conflict-based theory of self-monitoring as proposed by Nozari et al. (2011) that, if relatively more speech errors are made in an experiment, this can affect the general state of the production system such that overall less errors are detected.

A loglinear analysis of these frequencies confirmed that, relative to the baseline category of one-feature items varying in place *or* mode of articulation, there were more *errors* for voiced-voiceless items ($\beta = +0.587$, $Z = 7.56$, $p < .001$), and fewer errors for vowel-contrast items ($\beta = -0.168$, $Z = -1.831$, $p = .067$) and for place *and* mode of articulation items ($\beta = -0.349$, $Z = -3.60$, $p < .001$). Relative to the baseline category of one-feature items varying in place or mode of articulation, there were relatively fewer *repairs* for voiced-voiceless items (interaction $\beta = -0.957$, $Z = -3.77$, $p < .001$), but the incidence of repairs was not significantly different for vowel-contrast items, or for place *and* mode of articulation items. Apparently, as one would expect, the voiced-voiceless contrast leads both to more errors and to fewer repairs. A perception-based theory of self-monitoring would predict this. (Detailed results of all regression models, including models having noise condition and session number as predictors, are reported in the Supplementary Materials online).

Table 3.2
Numbers of responses broken down by phonological contrast and repair status, with percentages of repaired error responses.

Phonological similarity	Repair status		Total	%repaired
	Not repaired	Repaired		
Voiced-voiceless	812	64	876	7
Vowels	360	52	412	12
Place or mode of artic (similar)	496	75	571	13
Place and mode of artic (dissimilar)	331	62	393	16
Total	1999	253	2252	11

For the two phonetic contrasts that are identical in Experiments 1 and 2, the relative numbers of repairs were lower in Experiment 2 (137/964 or 14%) than in Experiment 1 (182/859 or 21%). This does not follow from a perception-based theory of self-monitoring, but, as mentioned above, it is predicted by a conflict-based theory of self-monitoring. This point will be taken up in the general discussion.

We focus again in our further analysis on repaired single elicited, single other and multiple speech errors, and start with analyzing the distribution of error-to-cutoff times of all repaired speech errors in the above categories. Fig. 3.1 gives the relevant breakdown.

Again, the distribution of error-to-cutoff times is bimodal, and can be described as consisting of two underlying gaussian distributions, one with a peak at 240 ms and one with a peak at 644 ms. The vertical dashed line in Fig. 3.1 corresponds to the best fitting separation of the two estimated distributions. In line with Nooteboom and Quené (2017), we propose that the gaussian distribution with the shorter durations corresponds to the class of repaired errors that were detected in internal speech and the gaussian distribution with longer durations to the class of errors detected in overt speech. Below we will consider error-to-cutoff times shorter than 525 ms as corresponding to errors detected internally and those longer than 524 ms as corresponding to errors detected externally. Of course, because the two distributions overlap, there is some unavoidable statistical noise in this classification, but it seems good enough for our purposes.

So now we are in a position to test our four predictions. Our first prediction is:

(1) Error-to-cutoff times are longer after single non-elicited errors and after multiple errors than after single elicited errors.

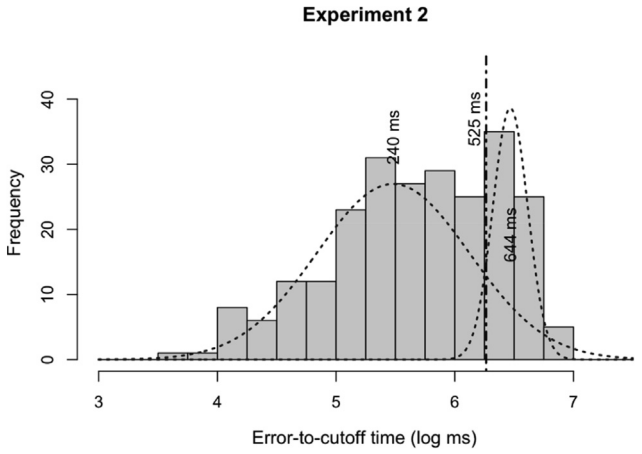

Fig. 3.1. Histogram of log-transformed durations of error-to-cutoff intervals in Experiment 1, for $N = 240$ repaired errors. Distributions plotted with dotted lines indicate the estimated distributions from an uninformed gaussian mixture model (see text). The vertical dashed line indicates the interpolated boundary value (6.264 corresponding to 525 ms) between the two distributions. (A number of cases were excluded because they had durations of 1 ms or less).

The reader may remember that we made this prediction because we have assumed that in single elicited errors competition is mainly between the error form and the correct target form, whereas in other (single or multiple) errors there is supposed to be competition between more than two forms. Thus during self-monitoring of other errors more time would be needed than with the single elicited errors to resolve competition and to decide that an error has been made. As in Experiment 1, this prediction was tested by means of LMM with participants as random intercepts, and with the error category as a fixed effect (cf. Table 3.1). Error-to-cutoff times were found to be approximately equally long after single elicited errors (mean 282 ms) and after single other errors [mean 246 ms; $\beta = -0.136$, $t = -1.03$, n.s., 95%CI $= (-0.410, +0.110)$], but error-to-cutoff times were significantly longer after multiple non-elicited errors [mean 439 ms; $\beta = +0.443$, $t = 3.33$, $p < .001$, 95%CI = (0.175, 0.707)]. This confirms our prediction (1) for multiple errors but not for single other errors.

Of course, the distributions of error-to-cutoff times deviate from normal, because error-to-cutoff times stem from both internal and external error detection, with different distributions (cf. Fig. 3.1). The difference in error-to-cutoff times between single elicited and multiple errors indicates that relatively more often single (elicited) errors are detected internally and multiple errors externally.

Our second prediction is:

(2) Single other errors and multiple errors are detected in internal speech relatively less often than single elicited errors are.

Fig. 3.2 gives the relevant breakdown of the data.

In order to test this prediction, the frequencies summarized in Fig. 3.2 were again compared using a loglinear model, with response category (cf. Table 2.1, again using single elicited errors as baseline category) and detection stage (internal, external; using internal detection as baseline) as two predictors. Single elicited errors (baseline) were detected externally less often than internally (46:137, $\beta = -1.076$, $Z = -6.23$, $p < .001$). In a loglinear model, the predicted effect of detection stage on response category shows up as an interaction effect of these two predictors, with frequency being the dependent variable. Single other errors were also detected externally less often than internally (5:29), and this effect was, against prediction, relatively the same as for single elicited errors: interaction $\beta = -0.647$, $Z = -1.25$, $p = .209$. Multiple errors were also detected externally relatively less often than internally (14:20), but, as predicted, multiple errors were detected externally relatively more often than single elicited errors were: interaction $\beta = +0.825$, $Z = 2.08$, $p = .037$. As in Experiment 1 this confirms our prediction (2) for multiple errors but not for single other errors.

Because most of the single elicited errors are consonants in initial position and many of the single other errors are not in initial position, a priori it is not excluded that in this analysis there is a confound between an effect of initial position on the one hand and an effect of elicited versus not elicited on the other. However, the effect of being in initial position would strengthen, not weaken the predicted effect of being elicited by the SLIP procedure. Hence the possible confounding does not explain the absence of the predicted effect for the single other errors.

However this may be, with the vowel-contrast items there is no confounding between a position effect and an effect of error category. We also performed a separate analysis for the vowel-contrast items only. Single elicited errors on vowel-contrast items (baseline) were detected externally less often than internally (7:23, $\beta = -1.190$, $Z = -2.76$, $p < .001$). Single other errors were also detected externally less often than internally (3:11) and once more this effect was, against prediction, relatively the same as for single elicited errors: interaction $\beta = -0.110$, $|Z| < 1$, n.s. Multiple errors were also detected externally less often than internally (1:7), and this effect too was the same as for single elicited errors: interaction $\beta = -0.756$, $|Z| < 1$, n.s.

Our third prediction is:

(3) (a) Incorrect forms are used less often as repairs than correct forms are, and (b) this effect is less strong after external detection than after internal detection.

In order to test these predictions, the frequencies of correct and incorrect repairs were again compared using a loglinear model, with detection stage (internal, external; using internal detection as baseline), and repair form (correct, incorrect; using correct repair as baseline) as two predictors. As predicted, for errors detected *internally*, repairs were mostly correct (142:44), and incorrect repairs were observed significantly less frequently ($\beta = -1.206$, $Z = -6.77$, $p < .001$). For errors detected *externally*, this effect was weaker (45:20), but the interaction was not significant (interaction $\beta = +0.418$, $Z = 1.29$, $p = .196$). This confirms prediction (3.a) but not (3.b).

Our fourth prediction is:

(4) (a) Cutoff-to-repair times are longer for single other errors and multiple errors than for single elicited errors, and (b) this effect is weaker after external detection than after internal error detection.

Fig. 3.3 is a histogram of all log-transformed cutoff-to-repair times (not to be confused with error-to-cutoff times summarized in Fig. 3.1) in Experiment 2.

Fig. 3.3 shows that in this experiment too the distribution of log-transformed cutoff-to-repair times deviates strongly from normal, again

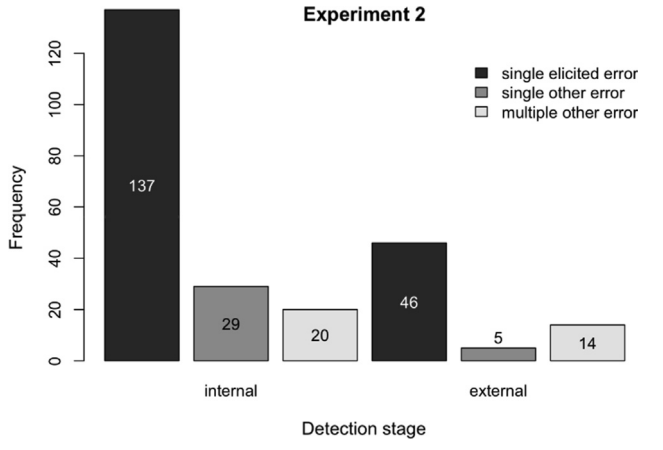

Fig. 3.2. Frequencies of single elicited, single other and multiple errors separately for "internal" and "external" error detection, in Experiment 2.

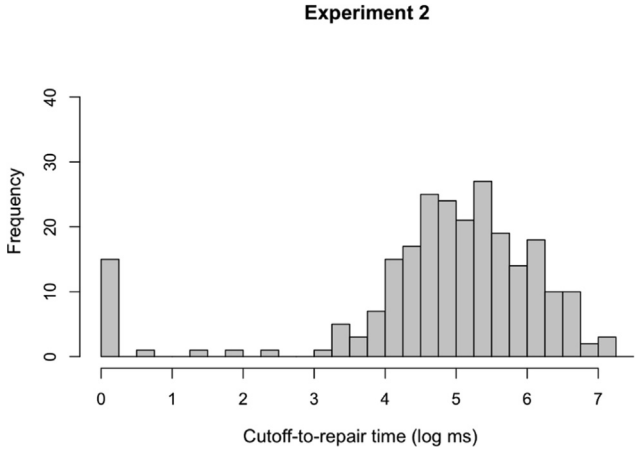

Fig. 3.3. Histogram of log-transformed cutoff-to-repair times in Experiment 2, with values of 0 ms converted to 1 ms before log transformation.

because immediate repairs, having a cutoff-to-repair times of 1 ms (converted from 0 ms to allow log transformation) are overrepresented, here with 15/251 observations. In line with our fourth prediction, we first tested whether (a) relatively fewer of these *immediate* repairs occurred after single other errors and after multiple errors than after elicited errors, and (b) whether this effect was weaker for errors detected externally. The odds of immediate repair for elicited errors were 11:171, and these odds were indeed lower for single other errors (2:33, $\beta = -1.674$, $Z = -8.82$, $p < .001$) and for multiple errors (2:32, $\beta = -1.705$, $Z = -8.87$, $p < .001$), but these effects were not significantly different for errors detected internally or externally [LRT: $\chi^2(2) < 2$, n.s.]. With regard to the occurrence of immediate repairs, this again confirms prediction (4.a) but not (4.b). Second, we tested the predicted effects on the log-transformed cutoff-to-repair times of *non-immediate* errors only (having cutoff-to-repair times > 1 ms), using an LMM with participants as random intercepts, and using the response category (again using single elicited errors as baseline category, cf. prediction 1 above) and detection stage (internal, external; using internal detection as baseline) as two fixed predictors. For internally detected errors, excluding those with immediate repairs, we find that after elicited single errors (baseline) detected in internal speech, cutoff-to-repair times are 124 ms on average. No effects of response category were observed [single other: mean 170 ms, $\beta = 0.314$, $t = 1.70$, n.s., 95%CI (-0.046, $+0.678$); multiple: mean 124 ms, $\beta = -0.00$, $|t| < 1$, n.s.], and these effects were the same for errors detected internally or externally [LRT: $\chi^2(2) < 2$, n.s.]. Hence, with regard to *non-immediate* repairs, neither prediction (4.a) nor (4.b) is confirmed.

Discussion of Experiment 2

In this experiment we used four phonetic contrasts between the two segments that were primed for interaction, viz voiced-voiceless, vowels, place or mode of articulation, place plus mode of articulation. We found that with the voiced-voiceless contrast significantly and considerably more errors were elicited than with the other three contrasts, and that also significantly and considerably fewer repairs were made with this contrast. Interestingly, for the two contrasts that were also used in Experiment 1, viz. place *or* mode and place *and* mode of articulation, fewer repairs were made in Experiment 2 than in Experiment 1, although the stimuli were essentially the same. In Experiment 1 (Table 2.2) the percentages were 17 and 27 for place *or* mode and place *plus* mode of articulation respectively, in Experiment 2 (Table 3.2) they were 13 and 16 respectively. We propose, in line with Nozari, Martin, and McCloskey (2019), that the difference between the two experiments is due to a difference in the general state of the production system. This finding, as will be explained in the general discussion, supports a conflict-based theory of self-monitoring of the kind proposed by Nozari et al. (2011).

We tested four predictions derived from a theory of repairing segmental speech errors in which it is assumed that during speech preparation and during self-monitoring more than a single word-like form is competing as candidate for the same slot in the utterance. The predictions were made specifically for a SLIP task eliciting reversals of the targeted segments in pairs of CVC words. Results of Experiment 2 mainly confirm the findings in Experiment 1. Again, in the overall pattern of results we do not find, as predicted, that single non-elicited errors are detected more slowly and more often after external detection than single other errors. For multiple errors this predicted effect was found. With respect to cut-off-to-repair times we find a different pattern for immediate than for nonimmediate repairs. Immediate repairs were more frequent both for single other and multiple errors than for single elicited errors, but the frequencies of immediate repairs did not show the predicted interaction with detection stage. In contrast to Experiment 1, we do not find a significant difference in cutoff-to-repair times of non-immediate repairs between single elicited and either single other or multiple errors, and no interaction with detection stage.

We have argued that due to priming by the precursor word pairs, in a SLIP task the activation of both the correct forms and the elicited error forms containing the reversed segments is boosted, thereby relatively suppressing activation of other candidates. However, in a SLIP task we generally do not only find the elicited errors but also non-elicited errors, suggesting that despite the priming due to the precursor word pairs, every now and then the main competition is not between the elicited error form and the correct form, but between more than two candidate forms, i.e. another error form and at least both the elicited error form and the correct form. Assuming that competition between highly activated forms is more easily resolved than competition between more than two less activated forms, and that error-to-cutoff time at least partly reflects the time needed for conflict resolution, we predicted (Prediction 1) that error-to-cutoff times are longer for single other and multiple errors than for single elicited errors. This was confirmed for multiple errors, but not for single other errors. This means that, despite the fact that in multiple errors conflict with the correct target items is greater than in single errors, resolution of the competition takes more time after multiple errors than after both single elicited and single other errors. The effect of more than two word form candidates competing with each other apparently is greater than the effect of the difference with the target items.

That we did not find the predicted effect for single other errors possibly is partly caused by the strongly non-normal distributions of error-to-cutoff times. The distributions of error-to-cutoff times are not necessarily gaussian for two reasons. One is that the distributions possibly are truncated in the lower tail, because cases in which the internal speech process is interrupted before overt speech is initiated are not included, although such cases seem to be rare, as predicted from Hartsuiker and Kolk (2001) and confirmed by Nooteboom and Quené (2019). The other reason is that the distribution of error-to-cutoff times contains repaired errors detected both in internal and in overt speech. These two classes of repaired errors have different temporal properties (cf. Fig. 3.1). But if indeed nevertheless conflict resolution takes more time for single other errors than for single elicited errors, then we expect that the time available for self-monitoring before speech initiation runs out more often for single other errors than for single elicited errors. In such cases self-monitoring may be shifted to overt speech. This gave our second prediction, i.e. that after single other errors there are relatively more errors detected in external speech than after single elicited errors, and of course the same prediction holds for multiple errors. This prediction was borne out for multiple errors, but again not for single other errors. If indeed, as suggested earlier, the absence of the predicted effects is due to lack of statistical power, the findings do not contradict that self-monitoring of single other and multiple errors takes more time than self-monitoring of single elicited errors. We still believe that during self-monitoring there is competition between candidate forms and that the amount of competition is a major determinant of temporal aspects of self-monitoring. In line with this reasoning, the finding that the predicted effect is stronger for multiple than for single other errors suggests that in multiple errors there are relatively more competing candidate repairs than in single errors. This is, of course, entirely reasonable.

If indeed, as we have assumed, correct repair candidates are sustained from the lexical level and incorrect repair candidates are not, we predict that correct repairs are far more frequent than incorrect repairs. However, we also have assumed that during the time delay between internal and external error detection, the activation of the correct candidate repair decreases. From this one would expect that the advantage of the correct candidate over other candidates, diminishes. Therefore it is predicted that the difference between correct and incorrect repairs is less for externally than for internally detected errors. Although this effect was indeed somewhat weaker in externally than in internally detected speech errors, this difference was not found to be significant.

After speech is interrupted because an error was detected, a repair

has to be made. As we have seen, every now and then a repair is already available at the moment of interruption, either because the repair is rapidly available or because interruption has been postponed until a repair is available. In those cases repairs are immediate, i.e. cutoff-to-repair times have a duration of 0 ms (which we changed to 1 ms before taking the logarithm). But in most cases making a repair available has not been completed at the moment of interruption. After a single elicited error has been detected, the main competition is between the error form and the correct target form. After another single error or a multiple error has been detected, the main competition is between the error form and one or more other error forms. Activation of the correct candidate repair will generally be higher than activation of an incorrect candidate repair, because activation of the correct is candidate is both boosted by the SLIP procedure and is sustained from the lexical level, whereas activation of a not-elicited incorrect candidate is not boosted and not sustained. This led to our fourth prediction, viz. that cutoff-to-repair times are longer for single other errors and for multiple errors than for single elicited errors, and that this difference decreases from internal to external self-monitoring. However, because of the over-representation of immediate repairs, the distributions of cutoff-to-repair times deviate strongly from normal. For this reason we also looked separately at the odds for immediate repair for single elicited, single other and multiple errors, and at the cutoff-to-repair times for the nonimmediate errors. The odds of immediate repairs were indeed significantly lower for single other and multiple errors than for single elicited errors, confirming prediction 4.a. This suggests that single elicited errors are repaired more rapidly than other errors. However, this difference was not significantly reduced between internally and externally detected speech errors. This does not confirm that activation of candidate repairs decreases between internal and external error detection. We also looked at the difference in cutoff-to-repair times between single elicited, single other and multiple errors after exclusion of immediate repairs. In contrast with Experiment 1, the effect for both single other errors and multiple errors was not significant after exclusion of immediate repairs. We suspect that this is due to the different temporal properties of repaired speech errors against the four contrasts used.

General discussion

The current attempt to investigate how segmental speech errors are repaired during self-monitoring capitalizes on the possibility that, at least statistically, such errors can be classified as detected in internal speech and detected in overt speech. The distribution of error-to-cutoff times happens to be bimodal, and to be a composite of two underlying gaussian distributions. This was found in Nooteboom and Quené (2017) and is confirmed in the current experiments. This bimodal distribution strongly supports the important proposal by Levelt (1983, 1989) and Levelt et al. (1999) that speech errors can be detected both in internal speech and in overt speech. The estimated time delay of about 470 ms in Experiment 1 and about 400 ms in Experiment 2 between internal and external detection is considerably longer than one would have expected, for example, from the computational implementation by Hartsuiker and Kolk (2001) of Levelt's model. Hartsuiker and Kolk predict an average error-to-cutoff time of 270 ms for the internal loop and 393 ms for the external loop, the difference being in the order of 120 ms. However, the differences estimated by the computational model (Hartsuiker & Kolk, 2001) may be incomparable to the difference estimated from the response times in the SLIP experiments as analyzed here. Whereas our experiments are limited to segmental errors, the experimental database Hartsuiker and Kolk used to tune their model contained both segmental and higher order errors, and also had relatively few errors. It is as yet not clear how these factors would affect the time delay between internal and external error detection.

Also, Hartsuiker and Kolk were in no position to classify empirically the errors in their corpus as being detected internally or externally. Their classification was entirely based on theory. Yet, the time delay found by Nooteboom and Quené (2017), and also in the current analysis is so long that it asks for an explanation. We propose that after external error detection interruption is postponed much more often and perhaps for a much longer time (for postponement of interruption see Seyfeddinipur et al., 2008; Tydgat et al., 2012) than after internal error detection, because at that late stage of self-monitoring fewer repair candidates are still sufficiently activated. This ties in with our proposal that activation of candidate repairs decreases during the time delay between internal and external detection.

Another property of the dual perceptual loop theory by Levelt and his colleagues, is that speech errors are detected by employing the same speech comprehension system that is also used in listening to other-produced speech. This proposal has come under serious attack for example in Nozari et al. (2011) and Nozari et al. (2016). One reason is that in aphasics one has found a double dissociation between self-monitoring and perceptual abilities (see a.o. Vigliocco & Hartsuiker, 2002). A second reason is that there is an attractive production-internal alternative in the form of conflict-based monitoring (Botvinick, Braver, Barch, Carter & Cohen, 2001; Yeung, Botvinick, & Cohen, 2004). As we have seen in the introduction to this paper, in the case of speech, a proposal of conflict-based monitoring capitalizes on the assumption that during speech preparation there potentially is, both at the level of semantic features and at the level of phonological features, competition between more than a single candidate for the same slot in the string being generated. If two or more candidates are highly activated, a conflict signal is generated and sent to a domain general executive center (Nozari et al., 2011). A third reason for preferring a production-based theory of self-monitoring is that there is a growing amount of evidence for competition between candidate word forms during speech preparation (Frisch & Wright, 2002; Goldrick & Blumstein, 2006; Goldstein et al., 2007; McMillan & Corley, 2010; Mowrey & MacKay, 1990; Slis & Van Lieshout, 2016). Nozari et al. (2016) investigated the locus of control processes in selection control and post-monitoring control. Their results support a model in which selection control operates separately at lexical and segmental stages, but post-monitoring control operates on the segmentally-encoded outcome.

There is one aspect of our data that in the context of the conflict-based theory of self-monitoring, asks for specific attention because it is informative of the mental mechanism leading to error detection. The conflict-based theory for error detection was proposed and further investigated by Nozari et al. (2011), Nozari et al. (2016) and Nozari et al. (2019). This theory has an interesting property that apparently, so far, has not been put the test. As discussed in the introduction, the main proposal of the theory is that during speech preparation there can be conflict on a representational level between simultaneously active and competing options for a particular slot in the string being generated. When an error has been made, there may be multiple competing items with high levels of activation. In such cases conflict information is passed on to a domain general executive center. Particularly Nozari et al. (2019, p.1230) argued that there is a “consistent relationship between error and repair probabilities, disentangled from position, compatible with a model in which greater control is recruited in error-prone situations to enhance the effectiveness of repair”. Interestingly, in the context of other properties of this theory, it is predicted that in a high-conflict situation, when conflict is relatively high even on correct trials, the system is less able to tell the difference between a correct high-conflict trial and an error trial. This has two consequences. More errors will be made, and fewer errors will be detected. This prediction can be tested in comparing our Experiments 1 and 2. The reason is the inclusion of the relatively weak voiced-voiceless contrast in Experiment 2. With “weak” we here mean that the contrast between a voiced and a voiceless initial plosive is less easily perceptible than other phonetic contrasts and also potentially carries less conflict than other contrasts.

The inclusion of voiced-voiceless contrast increases considerably the relative number of segmental errors in Experiment 2 as compared to

Experiment 1, and also the relative number of high-conflict correct trials in Experiment 2 as compared to Experiment 1. This would affect the general state of the monitoring system in such a way that in Experiment 2 more errors are made and fewer errors are detected than in Experiment 1. Of course, a perceptual account of self-monitoring would also predict that segmental errors that are less easily perceptible would be less often detected, but this prediction would be limited to the less perceptible errors themselves. It would not generalize to other, more easily perceptible, errors, as it would in the conflict-based theory of self-monitoring. Comparing error and repair rates in our two experiments shows (a) that voiced-voiceless errors are both much more frequent and much less often repaired than other errors, and (b) that in Experiment 2 repair rates are much lower than in Experiment 1, also for those contrasts that are the same in both experiments. However, for those contrasts that are the same in both experiments, we did not find that more errors are made. Our results in this respect partly support the conflict-based theory of self-monitoring.

Those who argue for production-based monitoring of internal speech generally believe that self-monitoring of overt speech is audition-based (cf. Huettig & Hartsuiker, 2010; but see Nozari et al., 2016). However, Nooteboom and Quené (2017) demonstrated that self-monitoring of both internal and overt speech does not depend on audition. They concluded that self-monitoring of overt speech can be based on somatosensory and proprioceptive information from the articulators, as suggested by Hickok (2012), Lackner (1974), Lackner and Tuller (1979) and Pickering and Garrod (2013). This does not necessarily imply that audition never plays a role in self-monitoring overt speech. However, under time pressure participants in the experiments in Nooteboom and Quené (2017) did not rely on audition, presumably because they employed the first information available, and that would be information from the articulators. It should also be pointed out that some studies that showed a contribution of audition to self-monitoring (e.g. Postma & Kolk, 1992; Postma & Noordanus, 1996), had collapsed segmental and higher order errors. It seems reasonable that detection of higher order errors can be more dependent on audition, because potentially a longer time window is involved.

We had predicted that error-to-cutoff times would be shorter for repaired single elicited errors than both for repaired single other errors and repaired multiple errors, and also that both single other errors and multiple errors would be more often externally detected than single elicited errors. These predictions were borne out for multiple errors but not for single other errors. Possibly, the effect of the SLIP procedure on boosting versus not boosting the activation of single segmental errors is too small to lead consistently to the predicted effects. One may note that the predicted effects are supposed to be caused mainly by fewer versus more candidate repairs competing for the particular slot in the utterance. For single segmental errors perhaps there are not consistently more candidate repairs when these are not elicited than when these are elicited. Our evidence would have been stronger with significant differences between single elicited and single other repaired errors.

The situation is different for the multiple errors. Although one may argue that these are very different from the single elicited repaired errors, and therefore they are not really comparable, yet we argue that these are interesting because in the absence of competing candidates one would expect that multiple errors would be detected faster, not slower, than single elicited errors simply because greater deviation of the error form from the target form would lead to greater conflict and therefore to faster detection. But this is not what happens. Of course, in the case of multiple errors it is reasonable to suppose that there are more competing forms than in the case of single errors. Resolving the conflict between more than two candidate forms would be more complicated and take more time than resolving the conflict between only two candidate forms. This is what our theory of repair predicts, and this is what we find.

Our assumption that correct candidate repairs are sustained from the lexical level and incorrect repairs are not, is supported by our finding in both experiments that correct repairs far outnumber incorrect ones. Our hypothesis that activation of candidate repairs decreases during the delay between internal and external error detection, is confirmed in Experiment 1. In Experiment 2 the general prevalence of correct repairs is confirmed, but the difference between internally and externally detected errors, although in the predicted direction, did not reach significance. In Experiment 1 we find that the difference between single other errors and multiple errors is significantly weakened going from internal to external error detection. This decrease in the relative number of correct repairs during the time delay between detection in internal and overt speech is evidence that after error detection repairs are not generated by re-compilation but rather by re-activation of already available candidate forms. After error detection of errors in overt speech, repairs consisting of incorrect forms are no exception; this suggests that in SLIP experiments quite often there is competition between more than two candidate word forms. Not only in single elicited errors, where correct repairs are to be expected, but also in other errors, where supposedly the main competition often is between two or more incorrect forms, we nevertheless often find correct repairs; this supports the idea that correct candidate repairs are sustained from the lexical level whereas incorrect repairs are not.

Our fourth and final prediction was that cutoff-to-repair times are shorter for single elicited than for other errors, and that this difference is weaker after external than after internal detection. The reason for this prediction is that, due to the SLIP technique, in single elicited errors the main competition is between two highly activated candidates, viz. the correct form and the elicited error form. In other errors competing candidates are supposedly less activated, and there may be more easily competition between more than two candidate forms. The predicted difference between internal and external detection stems from our assumption that the activation of candidate repairs decreases during the delay between internal and external detection. Looking at the relative number of immediate repairs we find, as predicted, in both experiments that there are fewer immediate repairs after single other and multiple errors than after single elicited errors. The difference was, however, not significantly weaker after external than after internal detection. This suggests that after single elicited errors more often a repair is immediately available at the moment speech is interrupted. This is in line with our suggestion that in the case of single elicited errors there are on average less competing candidate repairs than in the case of other errors. The predicted difference between internal and external detection was not confirmed. Looking at the distributions of error-to-cutoff times after exclusion of immediate repairs we found in Experiment 1, but not in Experiment 2, that error-to-cutoff times were indeed longer after both single other and multiple errors and that this difference was significantly weakened after external compared to internal detection for single other but not for multiple errors. The results of Experiment 1 suggest that when repairs are not immediate, they have to be re-activated. After internal detection, this re-activation takes less time after single elicited than after single other errors. This again is in line with our assumption that after detection of a single elicited error there are on average fewer competing candidates than after other errors. We have currently no good explanation that we do not find a similar difference between single elicited and multiple errors. In Experiment 2 we do not find significant differences between single elicited, single other and multiple errors, or between internal and external error detection. As the temporal properties between the four stimulus categories were rather different, they may have obscured the effects we were looking for.

A major feature of our results is that in most regards our predictions were confirmed in the comparisons between single elicited and multiple errors, but often not in the comparisons between single elicited and single other errors, with a few notable exceptions. For those who believe that the difference between single elicited and multiple errors is too great to make comparisons interesting, our attempt to confirm basic aspects of our theory of repairing segmental speech errors may seem

little successful. However, we believe that our comparisons between single elicited and multiple speech errors are revealing because our predictions are contrary to what one would expect in the absence of our theory. We have shown that multiple speech errors are detected slower than single elicited errors, that correct repairs prevail over incorrect repairs, that they do so somewhat less after internal than after external detection, that, in case there is a significant difference, repairing takes less time after single elicited than after other errors. These findings support our theory of repairing segmental speech errors.

Conclusion

In this paper we investigated some aspects of repairing segmental speech errors in a SLIP task. Predictions were derived from our proposal (Nooteboom & Quené, 2017, 2019) that repairs do not stem from re-compilation of the correct form but rather from candidate forms competing for the same slot in the utterance with the misspoken forms both during speech preparation and after speech is initiated. To this end we analyzed repaired segmental speech errors obtained in two SLIP experiments, keeping apart errors detected in internal speech and errors detected in overt speech, and also keeping apart single errors elicited in the SLIP task by priming reversals of two corresponding segments in two CVC monosyllable words, single not-elicited or other and multiple errors. The results of our experiments show that (a) error-to-cutoff times are considerably and significantly shorter for single than for multiple errors; (b) relatively more single elicited errors than multiple errors are detected in internal speech; (c) single elicited errors are relatively more often than multiple errors repaired with the correct forms, and that this effect is stronger after internal than after external error detection; finally (d) repairing takes more time after single other than after single elicited errors. Some of the predicted effects with respect to single other errors did not reach significance. Yet, together, these findings support a theory of repairing in which repairs stem from candidate forms that compete with the misspoken form during speech preparation and during self-monitoring of internal and overt speech.

Author note

Sieb Nooteboom, Utrecht institute of Linguistics OTS, Utrecht University; Hugo Quené, Utrecht institute of Linguistics OTS, Utrecht University.

The authors are grateful to the Utrecht institute of Linguistics OTS for providing the technical and logistical facilities that enabled us to do the two experiments. The authors are also grateful for the pertinent comments by two reviewers, one of them being Nazbanou Nozari, who both contributed greatly to the end result. The raw data, analysis scripts and supplementary materials of the two experiments can be found online at: https://osf.io/jahqe/.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2019.104069.

References

Boersma, P., & Weenink, D. (2016). Praat: Doing Phonetics by Computer (Computer Program). Version 6.0.18. Available at: http://www.praat.org/ (accessed 31-01-2016).
Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*(3), 624–652.
Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104*(4), 801–838.
Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language, 43*(2), 182–216.
Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.
Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics, 30*, 139–162.
Fromkin, V. A. (1973). Appendix (A sample of speech errors). In V. A. Fromkin (Ed.). *Speech errors as linguistic evidence* (pp. 243–269). The Hague: Mouton.
Garnham, A., Shillcock, R. C., Brown, G. D. A., Mill, A. I. D., & Cutler, A. (1982). Slips of the tongue in the London-Lund corpus of spontaneous conversation. In A. Cutler (Ed.). *Slips of the tongue and language production* (pp. 251–263). Amsterdam: Mouton.
Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes, 21*, 649–683.
Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition, 103*, 386–412.
Hartsuiker, R. J., & Kolk, H. H. J. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology, 42*, 113–157.
Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*, 135–145.
Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). New York: Routledge.
Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: 'external' but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes, 25*, 347–374.
Lackner, J. R. (1974). Speech production: Evidence for corollary discharge stabilization of perceptual mechanisms. *Perceptual and Motor Skills, 39*, 899–902.
Lackner, J. R., & Tuller, B. H. (1979). Role of efference monitoring in the detection of self-produced speech errors. In W. E. Cooper, & E. C. T. Walker (Eds.). *Sentence processing* (pp. 281–294). Hillsdale, N.J.: Erlbaum.
Laver, J. D. M. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.). *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 287–306). New York: Academic Press.
Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.
Levelt, W. J. M. (1989). *Speaking. From intention to articulation.* Cambridge, Massachusetts: The MIT Press.
Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–75.
Lüdecke, D. (2019). sjPlot: Data visualization for statistics in the social sciences. R package 2.6.2. https://doi:10.5281/zenodo/1308157.
MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills.* Berlin: Springer-Verlag.
McMillan, C. T., & Corley, M. (2010). Cascading influences on the production of speech: Evidence from articulation. *Cognition, 117*, 243–260.
Mowrey, R., & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence. *Journal of the Acoustical Society of America, 88*, 1299–1312.
Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.). *Errors in linguistic performance: Slips of the tongue, ear, pen and hand* (pp. 87–95). New York: Academic Press.
Nooteboom, S. G. (2005). Listening to one-self: Monitoring speech production. In R. Hartsuiker, Y. Bastiaanse, A. Postma, & F. Wijnen (Eds.). *Phonological encoding and monitoring in normal and pathological speech* (pp. 167–186). Hove: Psychology Press.
Nooteboom, S. G., & Quené, H. (2008). Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language, 58*, 837–861.
Nooteboom, S. G., & Quené, H. (2013). Parallels between self-monitoring for speech errors and identification of the misspoken segments. *Journal of Memory and Language, 69*, 417–428.
Nooteboom, S. G., & Quené, H. (2017). Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language, 95*, 19–35.
Nooteboom, S. G., & Quené, H. (2019). Temporal aspects of self-monitoring for speech errors. *Journal of Memory and Language, 105*, 43–59.
Nozari, N., Dell, G., & Schwartz, M. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology, 63*, 1–33.
Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, Cognition and Neuroscience, 31*(7), 886–903.
Nozari, N., Martin, C. D., & McCloskey, N. (2019). Is repairing speech errors an automatic or a controlled process? Insights from the relationship between error and repair probabilities in English and Spanish. *Language, Cognition and Neuroscience, 43*(9), 1230–1245. https://doi.org/10.1080/23273798.2019.1637007.
Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences, 36*(4), 329–347.
Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition, 77*, 97–131.
Postma, A., & Kolk, H. H. J. (1992). The effects of noise masking and required accuracy on speech errors, disfluencies, and self-repairs. *Journal of Speech and Hearing Research, 35*, 337–544.
Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech, 39*(4), 375–392.
Pouplier, M., & Goldstein, L. (2005). Asymmetries in the perception of speech production errors. *Journal of Phonetics, 33*, 47–75.
Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modelling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*, 413–425.
R Core Team (2019). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rossi, M., & Peter-Defare, E. (1998). *Les Lapsus ou notre Fourche a Langué.* Paris: Presses universitaires de France.
Schlenck, K., Huber, W., & Wilmes, K. (1987). "Prepairs" and repairs: Different monitoring functions in aphasic language production. *Brain and Language, 30*, 226–244.
Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2017). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal, 8*(1), 205–233.
Seyfeddinipur, M., Kita, S., & Indefrey, P. (2008). How speakers interrupt themselves in managing problems in speaking: Evidence from self-repairs. *Cognition, 108*(3), 837–842.
Slis, A., & Van Lieshout, P. (2016). The effect of phonetic context on the dynamics of intrusions and reductions. *Journal of Phonetics, 57*(3), 430–445.
Tydgat, I., Stevens, M., Hartsuiker, R. J., & Pickering, M. J. (2012). Deciding where to stop speaking. *Journal of Memory and Language, 64*, 359–380.
Van Alphen, P. M., & McQueen, J. M. (2006). The effect of voice onset time differences on lexical access in Dutch. *Journal of Experimental Psychology: Human Perception and Performance, 32*(1), 178–196.
Van Alphen, P. M., & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: The role of prevoicing. *Journal of Phonetics, 32*(4), 455–491.
Van Alphen, P. M. (2004). *Perceptual relevance of prevoicing in Dutch.* Unpublished doctoral dissertationNijmegen, The Netherlands: Radboud University.
Van Wijk, C., & Kempen, G. (1987). A dual system for producing self-repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology, 19*, 403–440.
Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in language production. *Psychological Bulletin, 128*, 442–472.
Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: Conflict monitoring and the error-related negativity. *Psychological Review, 111*(4), 931–959.