Available online at www.sciencedirect.com
ScienceDirect
Journal of Memory and Language 58 (2008) 837–861
Journal of Memory and Language
www.elsevier.com/locate/jml

Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors ☆

Sieb Nooteboom *, Hugo Quené

Utrecht institute of Linguistics OTS, Janskerkhof 13A, Utrecht University, 3512BL Utrecht, The Netherlands

Received 28 January 2007; revision received 8 May 2007
Available online 13 August 2007

Abstract

This paper reports two experiments designed to investigate whether lexical bias in phonological speech errors is caused by immediate feedback of activation, by self-monitoring of inner speech, or by both. The experiments test a number of predictions derived from a model of self-monitoring of inner speech. This model assumes that, after an error in inner speech, (1) an early interruption of speech may be made when speech was initiated too hastily, (2) the error may be covertly repaired, leading to the correct target, (3) the error may be covertly replaced by another speech error, or (4) an error may go undetected, leading to a completed spoonerism. This model of self-monitoring was supported by the speech errors observed in two SLIP experiments. The pattern of results supports the idea that lexical bias has two sources, immediate feedback of activation and self-monitoring of inner speech.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Speech errors; Lexical bias; SLIP technique; Feedback; Self-monitoring

Introduction

Explanations of the lexical bias effect

Lexical bias is the effect that phonological speech errors, for example BARN DOOR inadvertently spoken as DARN BORE, result in real words more often than in nonwords, other things being equal. This was demonstrated in the laboratory over 30 years ago by Baars, Motley, and MacKay (1975). Lexical bias has also been convincingly demonstrated in spontaneous speech errors (Dell & Reich, 1981; Nooteboom, 2005a; but see Del Viso, Igoa, & Garcia-Albea, 1991; Garrett, 1976). Recently, it was found that in bilinguals, lexical bias does not discriminate between languages (Costa, Roelstraete, & Hartsuiker, 2006).

☆ Portions of this work were presented at the AMLAP, 5–7 September 2005, Ghent, at the workshop on Disfluency in Spontaneous Speech, Aix-en-Provence, 10–12 September 2005, and at the 10th winter conference of the Dutch Psychonomics Society, Egmond aan Zee, 16–17 December 2005. Our thanks are due to Theo Veenker for technical assistance, to Rob Hartsuiker and Gary Dell for sharing their thoughts on many aspects of the research reported here, to Harald Baayen for suggesting the use of bootstrap validation of logistic regression in the data analysis and to Huub van den Bergh for statistical guidance and assistance. The raw data of the experiments are currently available online in the form of an excel document at [http://www.let.uu.nl/~Sieb.Nooteboom/personal/Nooteboom&Quene_speecherrors.xls]. Those who are interested in the original sound files, comprising more than 42 h of speech, can contact the first author about the conditions.
* Corresponding author. Fax: +31 302536000.
E-mail address: sieb.nooteboom@let.uu.nl (S. Nooteboom).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.05.003

Basically, two competing explanations have been proposed for lexical bias, reflecting different models of the architecture of the mental production of speech. The original explanation by Baars et al. (1975) was in terms of pre-articulatory editing of inner speech. Baars et al. assumed that nonwords are more often detected, rejected and repaired in inner speech than real words. This would explain why overt phonological speech errors are more often real words than nonwords. This explanation is strongly supported by Levelt (1989) and Levelt et al., 1999. Levelt introduced his "perceptual loop" theory of self-monitoring, which claims that the "monitor" employs the same speech comprehension system that is also used in listening to other-produced speech. In self-monitoring, the speech comprehension system receives two different forms of input, inner speech allowing the speaker to detect, reject and repair speech errors before they are articulated, and overt speech, allowing the speaker to detect, reject, and repair speech errors after they have been articulated. Following Baars et al. (1975), Levelt assumes that self-monitoring of inner speech uses a criterion of lexicality ("Is this a word?"). Nonlexical speech errors are more easily covertly detected, rejected and repaired than lexical errors. This explains lexical bias. Self-monitoring is supposed to be a semi-conscious process, sensitive to context. This self-monitoring explanation of lexical bias would be supported by evidence that lexical bias is affected by context. Such evidence has been provided by Baars et al. (1975), who found that in an experiment eliciting spoonerisms nonword–nonword errors are suppressed in a "mixed" context with both word–word and nonword–nonword stimuli, and that word–word errors are suppressed in a nonword–nonword context. Motley and Baars (1976) demonstrated in a similar experiment that the probability of spoonerisms to be elicited increases dramatically when the target word pairs are preceded by word pairs that are semantically related to the spoonerisms. Motley, Camden, and Baars (1982) found that taboo words in elicited spoonerisms are more often suppressed than nontaboo words. The suppressed taboo words were also accompanied by increased Galvanic Skin Response, showing that the taboo words were actually present in inner speech before being edited out. Further support for the role of centrally controlled pre-articulatory editing comes from Hamm, Junglas, and Bredenkamp (2004) who showed that in an experiment eliciting spoonerisms a secondary cognitive task taxing the central control system increases the number of spoonerisms, and also that in girls suffering from anorexia nervosa, a secondary cognitive task leads to a sharp increase in the number of spoonerisms semantically related to their illness.

A second explanation of lexical bias has been proposed by Dell and Reich (1980, 1981), Stemberger (1985), Dell (1986), and Dell and Kim (2005). These authors assume that during the mental production of speech there is immediate feedback of activation between phonemes and word forms. This causes activation to reverberate between phonemes and word forms, giving speech errors that form real words an advantage over speech errors that have no corresponding lexical representations. A computational model implementing immediate feedback of activation neatly accounts for lexical bias and for some other well known properties of phonological speech errors, such as the so-called "mixed error" effect (phonological speech errors are more likely when error and target are not only phonetically but also semantically similar), and the "repeated phoneme" effect (two consonants are more easily substituted for each other when they are followed by the same vowel than when they are followed by different vowels). Because feedback between phonemes and words is supposed to be an automatic process internal to mental speech production, the feedback account of lexical bias cannot easily explain the earlier mentioned context effects.

It is important to realize that feedback and self-monitoring of inner speech are thought to be successive processes that do not exclude each other. Those who believe that feedback is responsible for lexical bias, do not deny that there is also self-monitoring of inner speech. They do, however, deny that self-monitoring employs a criterion of lexicality. Feedback leads to more word–word than nonword–nonword spoonerisms in inner speech, before self-monitoring operates, and the probability of such inner-speech errors to be detected, rejected and repaired would be the same for both word–word and nonword–nonword spoonerisms. In principle, though, both feedback and self-monitoring of inner speech could change the ratio between word–word and nonword–nonword spoonerisms. This is precisely what is proposed by Hartsuiker, Corley, and Martensen (2005) who report a well-controlled experiment eliciting word–word and nonword–nonword spoonerisms, in which the kind of context is varied from mixed (word–word and nonword–nonword priming and test word pairs) to nonlexical (nonword–nonword pairs only). The main finding in this study is that it is not the case that nonwords are suppressed in the mixed context, as claimed by Baars et al. (1975), but rather that word–word errors are suppressed in the nonlexical context. Hartsuiker et al. explain this suppression of real words in the nonlexical context by adaptive behaviour of the self-monitoring system. This explanation presupposes that there is an underlying pattern, before operation of the self-monitoring system, that already shows lexical bias. This underlying pattern would be caused by immediate feedback as proposed by Dell (1986). In an experiment eliciting lexical and nonlexical spoonerisms with bilingual subjects, Costa et al. (2006) explain lexicality effects on the nontarget lexicon as resulting from feedback between phonology and lexical items.

So now, not counting production-based monitoring (Laver, 1973, 1980; MacKay, 1992; Postma, 2000), we have at least three possible accounts of lexical bias: (a) feedback of activation between phonemes and word forms alone, (b) self-monitoring of inner speech employing a criterion of lexicality alone, and (c) a combination of feedback and self-monitoring. The main objective of this paper is to test predictions derived from these three competing accounts of lexical bias. A main obstacle when investigating the lexical bias effect is that both the immediate feedback between phonemes and words, and the self-monitoring of inner speech, are hidden from direct observation. We therefore set up a model of the underlying processes from which predictions of observable data can be derived.

The structure of this paper is as follows. First, we discuss the basic technique for eliciting spoonerisms, and some aspects of earlier findings that are relevant to our approach. Then we develop and test a simple model of self-monitoring of inner speech. With certain assumptions, to be discussed below, this model allows us to derive some quantitative predictions from each of the three alternative accounts of lexical bias. These predictions are then tested in two experiments eliciting lexical and nonlexical spoonerisms. In general, the results support the third account outlined above, viz. (c) a combination of self-monitoring and feedback.

The SLIP technique for eliciting spoonerisms and a brief meta-analysis of earlier findings

Most attempts to investigate the source of lexical bias have made use of the so-called SLIP (Spoonerisms of Laboratory-Induced Predisposition) Technique. This technique was introduced by Baars and Motley (1974), and used by Baars et al. (1975) to study lexical bias in phonological speech errors. The technique was inspired by the observation that inappropriate actions may result from anticipatory biasing: If one persons asks another to repeat the word "poke" many times, and then asks: "what is the white of egg called?", then the answer "yolk" may be elicited. This incorrect answer is induced by the rhyming relation with "poke" (Baars, 1980). The SLIP technique works as follows: Participants are successively presented visually, for example on a computer screen, with priming word pairs such as DOVE BALL, DEER BACK, DARK BONE, followed by a target word pair BARN DOOR, all word pairs to be read silently. On a prompt, for example a buzz sound or a series of question marks ("?????"), the last word pair seen, i.e. the target word pair, in this example BARN DOOR, has to be spoken aloud. Interstimulus intervals are in the order of 1000 ms, as is the interval between the test word pair and the prompt to speak. Every now and then the participant will mispronounce a word pair like BARN DOOR as DARN BORE, as a result of phonological priming by the preceding word pairs.

If the SLIP technique is used to study lexical bias, two types of stimuli are compared, viz. stimuli eliciting lexical, or word–word, spoonerisms, such as BARN DOOR turning into DARN BORE, and stimuli eliciting nonlexical, or nonword–nonword, spoonerisms, such as BAD GAME turning into GAD BAME. A common finding is that, although both types of stimuli are equally frequent, word–word spoonerisms are produced more frequently than nonword–nonword ones. This is the lexical bias effect.

A major problem in solving the long standing controversy about the source of lexical bias, is that the SLIP technique, while generating a somewhat higher percentage of speech errors of all possible kinds, is only marginally successful in generating spoonerisms of the primed-for kind. We conducted a survey of published experiments (see Nooteboom & Quené, in press) in terms of their yield (percentages of elicited full exchanges relative to the number of test stimulus presentations). The yield varies from 0.8% (Dell, 1986, 1000 ms) to 8.2% (Baars et al., 1975, Experiment 2).

From the very beginning in Baars et al. (1975) the inefficiency of the task has led to habits in analyzing the data that may have obscured important aspects of the participants' strategies. The first of these habits is the pooling of errors from different categories such as 'completed spoonerisms' (BARN DOOR > DARN BORE), other 'full exchanges' (BARN DOOR > DARK BOARD), 'partial exchanges' (BARN DOOR > DARN DOOR or DA... BARN DOOR), and 'other speech errors' (BARN DOOR > ROAD DIS). Second, researchers have often removed all intrusion errors (errors identical to words that had occurred earlier in the experiment), because they assumed that such intrusion errors were not caused by the mechanism under investigation, either immediate feedback from phonemes to words or self-monitoring. We will argue below that there are good reasons to keep these error categories separate, and we propose to add as a separate error category those errors, including intrusions, that start with the initial consonant of the second word. If our model of self-monitoring is valid, lexical bias should be investigated in the 'completed spoonerisms' of the type BARN DOOR > DARN BORE and BAD GAME > GAD BAME. Other 'full exchanges', 'interrupted spoonerisms', and 'other speech errors', should be investigated separately. This would lower the yield of the experiments considerably.

There are good reasons to keep apart so-called 'full' and 'interrupted' exchanges. We have attempted to look separately at the relative numbers of 'full' and 'partial' exchanges in a number of published experiments. However, it appears that the term 'partial' exchanges denotes different things in different publications. The definition

used by Baars et al. (1975), also used in most early publications, includes 'interruptions' but possibly also 'anticipations' and 'perseverations'. All publications by Hartsuiker and colleagues on SLIP experiments use the notion 'partial spoonerisms' for 'anticipations' and 'perseverations' only, and not for 'interruptions'. Only Dell (1986, 1990) reserves the term 'partials' for what Nooteboom (2005b) called 'interrupted' spoonerisms. Humphreys (2002) used the notion 'aborted' speech errors for 'interrupted' speech errors. So it appears that only Dell (1986, 1990), Humphreys (2002), and Nooteboom (2005b) have a separate and comparable category of 'interrupted' spoonerisms. Table 1 shows some relevant data of their experiments.

If the numbers of 'full' and 'interrupted' exchanges are pooled over all these 8 experiments, for the conditions with lexical and nonlexical outcomes separately, this results in the numbers shown in Table 2.

The distributions differ significantly ($\chi^2(1) = 15.6$; $p < .001$). Note that the 'full exchanges' show a strong and highly significant lexical bias on a binomial test ($p < .001$), but the 'interrupted exchanges' do not differ ($p = .83$). This suggests that the 'interrupted' exchanges do not show a lexical bias effect. Nevertheless, Table 1 gives the impression that the lexical bias effect in 'interrupted' exchanges varies considerably from experiment to experiment, from negative (below 50%) to positive (above 50%). Possibly, the size and direction of the lexical bias effect in 'interrupted' speech errors depends on the specific features of the experiment. An informal comparison of the experimental methods suggests that this may be related to the amount of time pressure exerted on the participants, as well as on task structure, and instruction. Unfortunately, per experiment the data are too limited to investigate this possibility.

This brief meta-analysis of earlier findings suggests that 'full' and 'interrupted exchanges' should not be pooled into a single category, and also that it may be worth-while to ask what causes the variability of the lexical bias effect in 'interrupted spoonerisms'. It is noteworthy that virtually all 'interruptions' are cases where the expected spoonerism is 'early interrupted', i.e. after the initial consonant or initial CV. 'Early interruption' is clearly caused by monitoring inner speech (cf. Nooteboom, 2005b). Therefore it is not unreasonable to look for the cause of this variability in the operation of self-monitoring. If the positive lexical bias in 'completed spoonerisms' is compensated (to some extent) by a negative lexical bias in the 'interrupted errors', as in some experiments in Table 1, then this negative bias may be attributed to a Leveltian criterion of lexicality applied to inner speech, causing nonwords to be detected and rejected more often than real words. In those experiments in which 'interrupted errors' show a positive lexical bias, the criterion of lexicality is obviously not applied to those errors in inner speech that become overt as 'interrupted exchanges'. A positive lexical bias in 'full exchanges' could have been caused either by feedback, or by the monitor or by both. A positive lexical bias in 'interrupted exchanges', however, cannot easily be explained from monitoring inner speech. The reason is that repairs of 'interrupted exchanges' (interrupted spoonerisms) are overt, not covert (Nooteboom, 2005b). Therefore, a positive lexical bias in 'interrupted' exchanges would constitute stronger evidence in favour of feedback as a cause of lexical bias than a positive lexical bias in 'full exchanges' would. Thus, the relative frequencies of 'interrupted spoonerisms' are particularly relevant for the discussion of the cause of lexical bias.

If a negative lexical bias is found in 'interrupted exchanges' in some experiments and a positive lexical

Table 2
Numbers of 'full' and 'interrupted' exchanges, broken down by expected lexical and nonlexical outcomes, summed over eight published experiments

	Full exchanges	Interruptions
Lexical	234	172
Nonlexical	132	177

Table 1
Numbers of test trials, lexical and nonlexical 'full exchanges' and lexical and nonlexical 'interrupted exchanges'

Experiment	*N*	Lexical full exchanges	Nonlexical full exchanges	Lexical interrupted exchanges	Nonlexical interrupted exchanges
Dell (1986), 500 ms	880	21 (54)	18	35 (49)	37
Dell (1986), 700 ms	880	14 (54)	12	28 (56)	22
Dell (1986), 1000 ms	880	5 (71)	2	25 (57)	19
Dell (1990), Experiment 4, 600 ms	1260	8 (47)	9	6 (55)	5
Dell (1990), Experiment 4, 800 ms	1260	15 (88)	2	5 (83)	1
Humphreys (2002), Experiment 1	2000	51 (72)	20	29 (43)	38
Humphreys (2002), Experiment 4	1920	83 (62)	50	16 (50)	16
Nooteboom (2005b)	1800	37 (66)	19	28 (42)	39

The numbers of lexical 'full' and 'interrupted' exchanges are followed (within brackets) by percentages of the total numbers of 'full' or 'interrupted' exchanges, as an indication of the strength of positive or negative lexical bias.

bias in other experiments, then this also raises the question whether perhaps in the latter group of experiments the criterion of lexicality is directed elsewhere, for example to a class of responses that have so far escaped analysis. A possible candidate for this class of responses is formed by those errors that are not 'full' or 'interrupted' exchanges, but do start with the initial consonant of the second word. An example would be BAD GAME > GAS BAIT. Nooteboom and Quené (in press) demonstrated that the frequency of such errors is affected by the lexicality of the primed-for spoonerism. Such 'competing errors' (i.e. competing with the expected spoonerisms) were observed more frequently in the nonword–nonword than in the word–word priming condition. This suggests that at least some of those errors may be reactions to the elicited expected spoonerisms in inner speech.

A new model of self-monitoring

In Fig. 1, we present a simple flow chart that reflects our current model of monitoring inner speech during a typical SLIP experiment.

Target stimuli may elicit 'correct targets', 'elicited spoonerisms', or 'other speech errors' (here comprising all overt reactions in inner speech to target stimuli that are not 'elicited spoonerisms' or 'correct targets'). The main interest here is in what happens to 'elicited spoonerisms'. We speculate that a speech error like BAD GAME > G...BAD GAME originates in inner speech from competition between the 'correct target' and the 'elicited spoonerism' GAD BAME. The latter temporarily has the upper hand. Speech is initiated before the competition is resolved by the monitor. Meanwhile the monitor has detected the error in inner speech, whereupon the overt speech is interrupted before being completed. It is noteworthy that an interruption in cases like G...BAD GAME must be a reaction to inner and not to overt speech, because the speech fragment G... in such cases is shorter than a humanly possible reaction time. Frequently offset-to-repair times are also very short or even 0 ms, showing that not only the decision to stop but also the repair was prepared before articulation started (Blackmer & Mitton, 1991; Nooteboom, 2005b; see also Levelt, 1989: 473, 474; Hartsuiker, 2006).

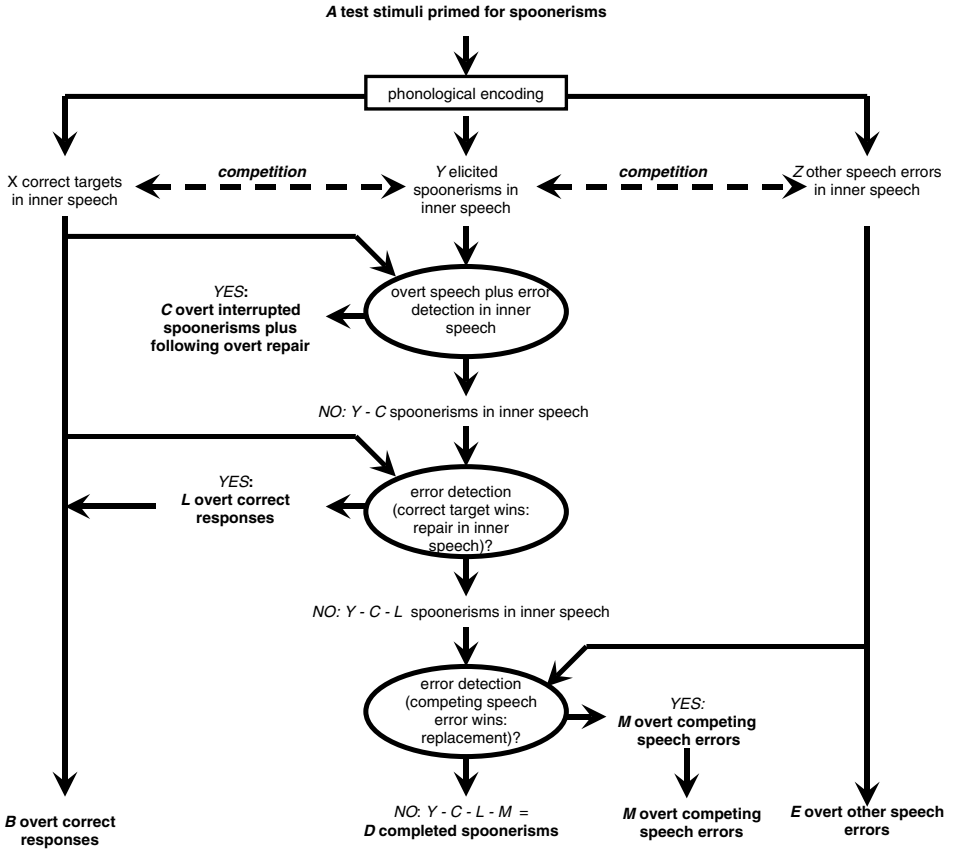

Fig. 1. Flow chart model of effects of monitoring inner speech during the development of spoonerisms on SLIP trials.

If an elicited error in inner speech is started to be spoken but interrupted, the repair has already been prepared in inner speech, and therefore the overt correct target can follow the interrupted spoonerism rapidly. From this we also predict that in the case of 'early interruption' the repair is virtually always the 'correct target' and not something else. This is so, because in the early phase of monitoring inner speech the 'correct target' was highly active anyway, and it competed with the elicited spoonerism: It is precisely this competition that the SLIP technique capitalizes on.

From our speculation that 'early interruptions' of 'elicited spoonerisms' are made if speech is initiated too hastily, further predictions can be derived. Before the monitor has resolved the competition between error and 'correct target', there are relatively more 'interrupted spoonerisms' if the participants are under time pressure and fewer under more relaxed conditions. Furthermore, it also follows that response times should be shorter for 'interrupted spoonerisms' than for 'other errors'.

Obviously, the 'correct target' may win the competition in inner speech with an 'elicited spoonerism'. This is accommodated by the 'repair' operation in Fig. 1, in which the 'elicited spoonerism' is replaced with the competing 'correct target', thereby resolving the competition in favor of the 'correct target'. However, such cases, being counted as 'correct responses', remain invisible in the experimental data.

However, the 'elicited spoonerism' may also be replaced by another word pair that is relatively active, for example a word pair that was part of the priming stimuli preceding the test stimulus. We assume here that such potential intruders often start with the initial consonant of the second word: BAD GAME was immediately preceded by the priming word pair GAS BAIT. Therefore the spoonerism GAD BAME immediately competes in inner speech with the still active GAS BAIT. The competition supposedly is enhanced by the sharing of the initial consonant. In Fig. 1 such cases follow the route marked as 'other speech errors'. The 'replacement' operation replaces the 'elicited spoonerism' with another highly active 'speech error'. From now on we refer to those speech errors that share the initial consonants with the 'elicited spoonerisms' as 'competing speech errors'. The reader may well ask why 'replacement' of an 'elicited spoonerism' should be limited to 'replacement' with errors that start with the initial consonants of the second word. Why not with arbitrary 'other errors'? The reason for this limitation is that the competition in inner speech is probably strongest between the spoonerism and those errors that start with the same consonant. But this limitation may be wrong. Fortunately, if the monitor happens to apply a lexicality criterion (as has been demonstrated by Nooteboom & Quené, in press), this is an empirical question. The frequency of errors starting with the same consonant as the 'elicited spoonerisms' should and the frequency other errors should not be sensitive to lexicality of the primed-for spoonerism.

Replacing an error by another error in inner speech should cost time. Thus one way to find out whether the current model makes any sense, is to measure response times for different error categories. Our model predicts, under the assumption that most errors of the type BAD GAME > GAS BAIT have competed with 'elicited spoonerisms', that the response times for such errors are considerably longer than the response times for the 'elicited spoonerisms' such as GAD BAME. Response times for 'competing speech errors' starting with the same consonant as the 'elicited spoonerism', should also be longer than those of other speech errors that presumably are not or less often involved in competition with the 'elicited spoonerism'. As argued above, response times for 'interrupted spoonerisms' of the form G. . .BAD GAME are predicted to be shorter than those for the GAD BAME cases.

The assumption that an intrusion error like GAS BAIT for BAD GAME is (in many cases) preceded by or has competed with an earlier 'elicited spoonerism' in inner speech is a very strong one. Such an intrusion may as well be independent of the 'elicited spoonerism', and should then be discarded, as has been common practice so far. However, our model of monitoring inner speech makes some testable quantitative predictions about the data obtained in a SLIP experiment. These predictions will be given shortly. Most importantly, Nooteboom and Quené (in press) found significantly more 'competing speech errors' in the nonword–nonword than in the word–word priming condition. This suggests that somehow the lexicality of the 'elicited spoonerisms' had played a role in the history of these 'competing errors'.

Obviously, the current model of monitoring of inner speech leads to some predictions that are independent of the three competing accounts of lexical bias. If these predictions hold, this would support the current model, which might then be used to test further predictions derived from each of the three competing accounts of lexical bias. We will first summarize our independent predictions, and then derive the predictions that follow from the three different accounts of lexical bias. So far, we have made the following predictions:

(a) Many errors in a typical SLIP task deviate from the elicited 'completed spoonerisms', but start with the initial consonant of the second word of the test stimulus word pair. An example would be BAD GAME, not turning into GAD BAME, but rather into GAS BAIT. These 'competing errors' are predicted to be frequent in the test condition, where spoonerisms are primed for, but not in the base-line condition, where no spoonerisms

are primed for. In this way, these 'competing errors' are supposed to differ from all kinds of other errors (not starting with the initial consonant of the second word) that bear no relation to the primed-for spoonerisms and that therefore may occur with equal frequency in the test and in the base-line conditions. In addition, we expect such 'competing errors' to result in real words, and not in nonwords.

(b) Response times for 'early interruptions' (BAD GAME > G...BAD GAME) are predicted to be shorter than response times for 'completed spoonerisms'.

(c) Because 'competing errors' (BAD GAME > GAS BAIT) result from two consecutive operations instead of one, response times are predicted to be longer than those for 'completed spoonerisms' and than those for 'other speech errors'.

(d) Repairs of 'early interruptions' are virtually always formed by the 'correct targets', rarely by 'other errors'.

If the above predictions were confirmed by the data, then this would support the current model of monitoring inner speech (Fig. 1), and then we could attempt to use this model to derive and test some predictions from each of the three competing accounts of lexical bias. We will now formulate the predictions derived from each of these three competing accounts of lexical bias.

A first possibility is that lexical bias is caused by feedback only. This leads to the following predictions:

(1) If the dead-line for responding is not too short for feedback to work (e.g. about 1000 ms; cf. Dell, 1986), then for each error category separately, the numbers of errors are larger for the word–word than for the nonword–nonword priming condition. This is so because the number of internal spoonerisms (prior to monitoring) will be larger in the word–word than in the nonword–nonword condition. Because the monitor would not distinguish between these conditions, more errors that were underlyingly spoonerisms will become overt in the word–word condition, regardless of whether they are full spoonerisms, interruptions, or replacements.

Because the strength of feedback depends on the amount of time available (feedback builds up over time), certain differences between error categories in the size of the lexical bias effect can be explained by the 'feedback-only' account plus different response times for different error categories. If interrupted errors have the shortest, 'completed spoonerisms' intermediate, and 'competing errors' the longest response times, then one expects the lexical bias effect to increase in this order. Finally, the 'feedback-only' account does not predict a negative lexical bias for any error category under any circumstances.

(2) A 'feedback-only' account as proposed by Dell (1986) predicts a small interaction between the phonetic similarity of the two phonemes involved in a spoonerism and the lexicality of the spoonerism: Lexical bias is slightly stronger when the two phonemes are dissimilar than when they are similar. It is not immediately clear why this should be so, but this small interaction was found in a simulation study for phonemes that were or were not followed by the same vowel (Dell, 1986), and it was also found in a simulation with the same model for phonetically similar and dissimilar consonants (Dell, personal communication).

A second possibility is that the monitor alone is responsible for lexical bias by employing a lexicality criterion in the detection of speech errors in inner speech. This is the position taken by Levelt (1989) and Levelt et al. (1999). Although, the standard view is that detection of a speech error is generally followed by a 'covert repair', we assume here that detection may also be followed by 'interruption' or by an operation replacing the 'elicited spoonerism' with another speech error. From this view the following predictions can be derived:

(1′) As a result of the lexicality criterion applied by the monitor, there will be less 'interruptions' and/or 'competing errors' (and also less 'covert repairs', but these remain invisible), and as a consequence relatively more 'completed spoonerisms', in the word–word than in the nonword–nonword priming condition: The positive lexical bias in the 'completed spoonerisms' would be mirrored by a negative lexical bias in 'interruptions' and/or 'competing errors'. The reason is that nonword–nonword spoonerisms are more frequently detected in inner speech than word–word errors, and are subsequently either spoken and interrupted, or replaced by the correct target, or replaced by another speech error.

(2′) Predictions by the 'self-monitoring-only' account on the effect of phonetic similarity derive from Levelt's (1989) theory of self-monitoring, in particular its assumption that the monitor employs the same speech comprehension system that is used for the perception of other-produced speech. Part of the comprehension system is a system for word recognition. We assume that word forms in inner speech are fed to this word recognition system. The lexicality criterion works as follows: When no fitting lexical representation is found, an error is detected and a repair is initiated. If the nonword error form is phonetically similar to the target form it is likely that this target form is incorrectly recognized, because in the SLIP task this target form is pre-activated by the silent reading part of the task. When the target form is (incorrectly) recognized by the speech comprehension system, the error remains undetected. The probability that the target form will be recognized on the basis of the nonword error form decreases with increasing phonetic distance between error and target. Of course, the

lexicality criterion fails if the error is a real word. Detection of real-word errors must follow a different route, immediately comparing the error form with the intended target form (cf. Nooteboom, 2005a). It is as yet unclear how this comparison would be affected by phonetic similarity. We thus predict a modulating effect of phonetic distance. In the nonword–nonword priming condition there are relatively more 'interruptions' and/or 'competing errors' and relatively fewer 'completed spoonerisms' with dissimilar consonants than with similar consonants. Whether this would also be the case in the word–word priming condition is an open question.

Finally, as suggested by Hartsuiker et al. (2005), it is possible that lexical bias is caused by both feedback and a lexicality effect in monitoring inner speech. This leads to the following predictions, that are similar to but not identical with predictions 1′ and 2′:

(1″) There are fewer 'interruptions' and 'competing errors' (and fewer 'covert repairs') and more 'completed spoonerisms' in the word–word than in the nonword–nonword priming condition. Thus, the distributions of error categories would be significantly different for the two priming conditions. This difference would be caused by the monitor which, by employing a lexicality criterion, detects nonword–nonword errors more frequently than word–word errors. But this effect would be superimposed on a prior effect of feedback, that underlyingly causes there to be more word–word than nonword–nonword spoonerisms. Thus the lexical bias in 'interruptions' and/or 'competing errors' would be much smaller than that in the 'completed spoonerisms'. The lexical bias in 'interruptions' and 'competing errors' may be absent or even negative, depending on the relative strength of feedback and self-monitoring as sources of lexical bias (cf. Hartsuiker, 2006). A possible negative lexical bias in 'interruptions' and 'competing errors' would not compensate fully for the positive lexical bias in 'completed spoonerisms', as it would in the case of a 'self-monitoring only' account of lexical bias.

(2″) We again predict that in the nonword–nonword priming condition there are relatively more 'interruptions' and/or 'competing errors' and relatively fewer 'completed spoonerisms' with dissimilar consonants than with similar consonants. This is, as explained under (2′), because the monitor would miss errors that are similar more easily than errors that are dissimilar to the target.

There is one further prediction to be made, under the assumption that the monitor employs a lexicality criterion. This prediction relates to the variation of lexical bias in 'interruptions' over different published experiments. Let us hypothesize that the criterion of lexicality applied in monitoring inner speech, which is sensitive to attentional factors, is also influenced by time pressure. Then monitoring could be more directed towards the very early 'interrupted speech errors' under time pressure and more towards the later 'competing errors' under more relaxed conditions. This might explain the wide variation in strength and direction of lexical bias in 'interruptions' in the published experiments (see Table 1). This hypothesis predicts that under time pressure there is a negative lexical bias in the 'interruptions', but under more relaxed conditions there is either no lexical bias or a positive lexical bias in the 'interruptions', whereas the negative lexical bias in 'competing speech errors' should be stronger under more relaxed conditions than under time pressure. One would not expect a lexicality effect in the category of 'other speech errors' that do not begin with the same consonant as the expected spoonerism.

In order to test these predictions, two SLIP experiments were conducted, in which the numbers of 'completed spoonerisms', 'interrupted spoonerisms', 'competing errors' and 'other speech errors' were counted separately. Two key factors in both experiments were the lexicality of predicted outcome and phonetic distance between the to-be-interchanged consonants. In Experiment 1, participants were under considerable time pressure, and were explicitly urged to correct as fast as possible any speech error they would make. The second experiment mainly differed from the first in that there was little time pressure and no cue or explicit urge for correction.

Experiment 1

Experiment 1 was set up to test both the general predictions following from our simple model of monitoring inner speech, and, if the model is validated, to find out which of the three competing accounts of lexical bias in phonological speech errors gets most support.

Methods

The method was basically the same as in Nooteboom (2005b), but with some modifications, mainly intended to increase the time pressure, to improve on the design by using the same target word pairs as test stimuli and base-line stimuli, and to derive stimuli with nonlexical expected outcomes from those with lexical expected outcomes. In addition, several improvements were made in order to prevent participants from guessing the purpose of the experiment, or from predicting when a target stimulus would follow, thus forcing them to pay attention to each word presented.

Stimulus material

There were 18 target word pairs with expected nonword–nonword outcomes; these were derived from 18 pairs with expected word–word outcomes by changing only the coda of each word. This matching of stimuli with expected word–word and nonword–nonword out-

comes will be exploited in the data analysis. The precursor priming word pairs all had the reverse initial consonants as compared to the following test word pair. The last word pair priming for a spoonerism always had the same vowels as the target word pair. Each test and each base-line stimulus was preceded by five word pairs. For the test stimuli, the last three of these were priming an exchange of the initial consonants.

The initial consonants of priming word pairs and target word pairs were chosen from the set /f, s, x, v, z, b, d, p, t, k/ and each set of 18 target word pairs was divided in 3 groups of 6 word pairs with equal phonetic distance between initial consonants, viz. 1, 2 or 3 distinctive features.

To these test and base-line stimuli were added 46 filler stimuli, 4 of which had 4 preceding word pairs (no priming), 4 had 3 preceding word pairs (no priming), 12 had 2 preceding word pairs (6 of which were primed for spoonerisms by both preceding word pairs), 8 with 1 preceding word pair (4 primed for spoonerisms by the single preceding word pair), and 18 with 0 preceding word pairs. The idea was that the participants could not anticipate when a response had to be given, so that they had to pay full attention to each word pair, even to the first word pair of a trial sequence. In addition, 7 practice stimuli were constructed, with a variable number of nonpriming preceding word pairs. Two stimulus lists were constructed, with the two matching word pairs (yielding word–word and nonword–nonword outcomes) distributed complementarily over these lists. Practice and filler trials were identical in the two lists.

Participants

There were 102 participants, most of them students and employees of the Faculty of Humanities at Utrecht University, with no known or self-reported hearing or speech deficit.

Procedures

Each participant was tested individually in a sound-treated booth. The timing of visual presentation on a computer screen was computer controlled. The order in which test and base-line stimuli, along with their priming or nonpriming preceding word pairs, were presented was randomized and different for each pair of an odd-numbered and the following even-numbered participant. The order of the stimuli for each even-numbered participant thus was basically the same as the one for the immediately preceding odd-numbered participant, except that word–word outcome stimuli and derived nonword–nonword outcome stimuli were interchanged. Fifty-one participants were, after the practice word pairs, presented with list 1 immediately followed by list 2, the 51 other participants were presented with list 2 immediately followed by list 1. After the final word pair of each trial a "?????"-prompt, meant to elicit pronunciation of the last word pair seen (the target word pair), was visible during 900 ms and then immediately followed by a simultaneous loud buzz sound and blank screen, both of 100-ms duration. The participants were strongly encouraged to speak the last word pair seen before this buzz sound started. This was practiced during the practice items. The buzz sound was immediately followed by a cue consisting of the Dutch word for "correction", visible during 900 ms again followed by 100 ms with a blank screen. The participants were instructed to correct themselves immediately whenever they made an error. It was not necessary to wait for the "correction"-prompt. After the correction period and a 100-ms resetting period, the first word pair of the following trial sequence was presented.

All speech of each participant was recorded with a Sennheiser ME 50 microphone, and digitally stored on one of two tracks of DAT with a Grundig DAT-9009 Fine Arts DAT-recorder with a sampling frequency of 48000 Hz. The resulting speech was virtually always loud and clear. On the other track of the DAT two tones of 1000 Hz and 50-ms duration were recorded with each target stimulus, one starting at the onset of the visual presentation of the "?????"-prompt, the other starting at the onset of the presentation of the "correction"-prompt. These signals were helpful for orientation in the visual oscillographic analysis of the speech signals (and also for measuring response times). Whereas Baars et al. (1975) had their participants listen to white noise during the experiment, probably to make them focus on inner speech rather than overt speech, this was avoided in the current experiment. Testing took approximately 16 min for each participant.

Scoring the data

Responses to all test and stimulus presentations were transcribed either in orthography, or, where necessary, in phonetic transcription by the first author using a computer program for the visual oscillographic display and auditory playback of audio signals. Responses were categorized as:

(1) 'Fluent and correct responses' of the type BARN DOOR > BARN DOOR or BAD GAME > BAD GAME.
(2) 'Completed spoonerisms' of the type BARN DOOR > DARN BORE or BAD GAME > GAD BAME.
(3) 'Anticipations' of the type BARN DOOR > DARN DOOR.
(4) 'Interrupted spoonerisms' of the type BARN DOOR > D...BARN DOOR. There were very few interruptions after the first vowel of the elicited spoonerisms (cf. Nooteboom, 2005b). All interruptions were included.

(5) 'Competing errors' of the type BARN DOOR > DARK BOARD, BARN DOOR > DARK BORE, BARN DOOR > DARN BOARD, BAD GAME > GAS BAIT, BAD GAME > GAS BAME, OR BAD GAME > GAD BAIT. 'Competing errors' included all errors in which at least one of the two forms of the elicited spoonerism was replaced by something else, and the resulting error began with the initial consonant of the second word. The very few cases where the something else was one of the two words of the target stimulus were excluded.
(6) 'Perseverations' of the type BARN DOOR > BARN BORE.
(7) Miscellaneous errors, including BARN DOOR > GOAT BALL, but also (the very few) 'hesitation errors' such as BARN DOOR > uhh BARN DOOR.
(8) No responses.

Response times for all correct and incorrect responses, to both base-line and test stimuli, were measured by hand in a two-channel oscillographic display from the onset of the visual prompt (=the onset of the 50-ms tone) to the onset of the spoken response. The onset of the spoken response was in most cases defined as the first visible increase in energy that could be attributed to the spoken response. However, the voice lead in responses beginning with a voiced stop was ignored because in Dutch duration of the voice lead appears to be highly variable and unsystematic both between and within participants (Van Alphen, 2004), as confirmed by a range from 0 to roughly 130 ms observed for voice leads in the current experiment. Response times faster than 100 ms or slower than 900 ms were excluded from further analysis. This was done because response times shorter than 100 ms were considered anticipatory (i.e., not related to the prompt), and response times longer than 900 ms were initiated too late (i.e., after the response period within which participants were instructed to respond).

Results

Preliminaries

In this experiment, phonetic similarity was varied in terms of a difference of 1, 2 or 3 phonetic features between the two to-be-spoonerized consonants. During data analysis we found that most of our participants, mainly young Dutch students, had no voiced-voiceless opposition for word-initial fricatives. This agrees with a thorough study of devoicing of Dutch voiced fricatives in initial position in the period 1935–1993 (Van de Velde, Gerritsen, & Van Hout, 1995). Therefore word pairs were recoded as phonetically similar if the two consonants differed in only one feature, and phonetically dissimilar if the two consonants differed in more than one feature, ignoring the voiced-voiceless opposition for fricative consonants. After this recoding, the numbers of phonetically similar and dissimilar consonant pairs differed. However, this causes no problem with the main analysis applied here, viz. multinomial logistic regression, because the proportions in that analysis are always relative to the number of total responses in that condition, thus automatically normalizing for differences in the number of target stimulus presentations between conditions.

Data analysis

The first dependent variable in this SLIP experiment, viz. error rate in each response category, was analyzed by means of multinomial logistic regression (Hosmer & Lemeshow, 2000; Pampel, 2000), because this takes into account the interdependency of the distributions of responses over categories. In a logistic-regression analysis, the proportion P of each response category is converted to log-odd units [or logit units, i.e. to the logarithm of the odds of P; $\text{log-odd}(P) = \log(P/(1-P))$]. Negative log-odd values indicate $P < 0.5$. These log-odd values are then regressed on the independent factors and predictors.

However, the necessary assumption of independent observations was obviously violated, since multiple participants had responded to the same item. The random variation over items and over participants was simulated by performing bootstrap replications of the multinomial regression (Efron & Tibshirami, 1993), using a two-stage bootstrap-with-replacement procedure as recommended by Shao & Tu (1995, p. 247 ff). Recall that there are 18 pairs of matching target items (with expected word–word and nonword–nonword outcomes, respectively). In the first stage, a sample of 17 item pairs was drawn with replacement from the 18 of such pairs. One may note that, in this first stage, we could also have chosen to sample $102 - 1$ participants instead of $18 - 1$ item pairs. Indeed, results from both options were computed. Since inter-item variability was found to be larger than inter-participant variability, the analysis and results presented here are more conservative than those obtained through first-stage sampling over participants would be. These resampled items "brought along" their responses into the "pseudo" data set. In the second stage, a bootstrap sample was drawn with replacement from the "pseudo" data set, with the bootstrap sample having the same size as the "pseudo" data set.

The resulting data set was then analyzed by means of fixed-effects-only multinomial logistic regression, using a regression model containing an intercept, four dummy factors for the four main cells (defined by lexicality and dissimilarity), and the number of lexical neighbours (centered to its median value of 24) of the first stimulus word of each target word pair. We limited this

experimental variable to the first word of each target word pair because error probability is determined by the properties of the first word (Humphreys, 2002). This bootstrap-and-regression procedure was repeated 250 times. Post-hoc analysis of the regression coefficients showed that response distributions were only minimally affected by the neighbourhood density of the stimulus. The average coefficients of this predictor, for completed, interrupted, competing, and other errors were 0.002, 0.010, −0.050, and 0.005, respectively. Such effects, even if significant, are too small to be of any relevance. If the number of competitors would double, for example, from 24 to 48, then the average rate of 'completed spoonerisms' would change from 5.2% to 5.5%. Hence, we will further ignore these small effects of neighbourhood density on the error rates.

The four coefficients for the dummy factors may be regarded as estimated means for each cell, based on varying item pairs and participants for each replication. Differences between cells were evaluated by means of sign tests of the estimated means, using Bonferroni adjustment for multiple comparisons (see Quené, 2007, for further details of the procedures for bootstrap, analysis, and hypothesis testing).

Response times constitute the second dependent variable in this study. These were analysed by means of mixed-effects regression analysis, with dummy factors for each response category, as well as lexicality, dissimilarity, response category, and the number of lexical neighbours (again centered to its median value of 24) as fixed predictors. Both participants and matching item pairs were included as two crossed random factors (see Quené & Van den Bergh, 2004, for detailed arguments and simulations). Computations were done with the function *lmer* in the package *lme4* for *R* (Bates, 2005; Pinheiro & Bates, 2000), yielding estimated regression coefficients with associated standard errors. Differences among response categories were investigated by means of post-hoc contrasts among their estimates; the variances associated with these contrasts may be evaluated by means of χ^2 tests (using $\alpha = .05$; Goldstein, 1995; Pinheiro & Bates, 2000; Quené & Van den Bergh, 2004, submitted for publication).

Testing predictions from the current model of self-monitoring

A first breakdown of responses from Experiment 1 is given in Table 3.

This table suggests that the numbers of errors, in particular of 'completed spoonerisms', 'anticipations', 'interruptions' and 'competing errors', are strongly affected by the lexical outcome of the 'elicited spoonerisms' in the test condition, but not so in the base-line condition. Pooling, for the two priming conditions together, all those errors in the test condition in Table 3 that start with the initial consonant of the second word, i.e. 'completed spoonerisms', 'anticipations', 'interrupted spoonerisms' and 'competing errors', gives 294 errors. Pooling the remaining speech errors (perseverations, miscellaneous and no responses) gives 192 errors. When we do the same exercise for the base-line condition, we get only 33 errors starting with the initial consonant of the second word, and 270 remaining speech errors. These distributions differ of course significantly ($\chi^2(1) = 189$, $p < .001$), suggesting that many of the errors starting with the initial consonant of the second word are triggered by the priming of a consonant exchange.

Table 3
Numbers of responses from Experiment 1, broken down by response category over rows, by priming condition (test vs. base-line) and by lexicality condition (WW or word–word outcome vs. NN or nonword–nonword outcome)

	Test		Base-line	
	WW	NN	WW	NN
(1) Fluent & correct	1583	1603	1685	1684
(2) Completed spoonerism	54	32	5	0
(3) Anticipation	6	0	7	2
(4) Interrupted spoonerism	48	60	8	4
(5) Competing error	43	51	3	4
(6) Perseveration	0	1	0	0
(7) Miscellaneous	68	54	97	106
(8) No response	34	35	31	36
Total	1836	1836	1836	1836

We have pooled the very few anticipations with the 'completed spoonerisms', and concentrated the further analysis on 'completed spoonerisms', 'interruptions', 'competing errors' (starting with the same consonant as 'elicited spoonerisms'), and 'other speech errors' (not starting with the same consonant as the 'elicited spoonerisms').

Consistent with our first prediction (a) the so-called 'competing errors' are virtually limited to the test condition, having 94 of those against only 7 in the base-line condition (binomial: $p < .001$), in line with our suggestion that these errors may be reactions to earlier 'elicited spoonerisms' in inner speech. Of the 94 'competing errors' there are only 3 where there was a nonword response that was not one or both of the 'elicited spoonerism' candidates; all 3 of these occurred in the nonword–nonword priming conditions. For the category 'other speech errors' these numbers are very different: There are 123 'other speech errors' (miscellaneous and perseverations pooled) in the test condition (nonwords: 16 cases), and 203 in the base-line condition (nonwords: 24 cases).

Response times were analyzed by means of mixed-effects modeling, yielding the coefficients and variances in Table 4. As predicted (prediction b), average response times were shorter for 'interruptions' (539 ms) than for

Table 4
Estimated parameters for the mixed-effects regression of response times in Experiment 1

Fixed effects	Coefficients	*SE*	*t*
(Intercept)	495	9.6	51.49
Lexicality	1	4.4	0.18
Dissimilarity	−12	9.0	−1.29
Resp. completed	65	9.6	6.84
Resp. interrupted	44	8.7	5.00
Resp. competing	96	10.0	9.62
Resp. other	65	8.3	7.80
No. of neighbours	1	0.3	2.79
Lexic. × dissim.	−8	6.0	−1.30
Random effects	**Variance**	***SD***	**No. of obs.**
Participants	4793	69	102
Items	272	17	18
Residual	7355	86	3528

For fixed effects, regression coefficients are given, with standard errors and *t* values; for random effects, the variances and standard deviations are given (see text for details).

'completed spoonerisms' (561 ms), but not significantly so ($p = .083$). They were, as predicted, (prediction c) significantly longer for 'competing errors' (591 ms) than for 'completed spoonerisms' ($p = .026$). For the category 'other speech errors' (not sharing the initial consonant with the expected spoonerism), average response times were virtually identical to those of the 'completed spoonerisms' (560 ms, n.s.).

In accordance with prediction (d) 'interruptions' were virtually always followed by the correct target: Of the 128 interruptions, 5 were not repaired at all, 6 were 'repaired' with another error, and 117 were followed by the correct target.

These observations are consistent with our model of monitoring inner speech. Below we will see whether the data of this experiment allow to discriminate between the three competing accounts of lexical bias.

Competing explanations of lexical bias (1)

To test predictions from the three competing accounts of lexical bias, viz. feedback, self-monitoring or both, the data were analyzed with a multinomial regression analysis, as described above. Table 5 gives the raw counts of response categories and, between brackets, the fractions relative to the total number of overt responses.

Fig. 2 shows the estimated response rates in log-odds in Experiment 1, as obtained in the multinomial regression, broken down by priming condition, response category and phonetic similarity. As expected, the positive lexical bias in 'completed spoonerisms' is significant for both similar and dissimilar consonants (word–word and nonword–nonword outcome conditions were compared with sign tests, both $p < .001$).

The 'feedback only' account predicts (prediction 1) that, although there may be differences in the strength of the effect between similar and dissimilar consonants and between error categories, lexical bias is positive for all error categories. This should be so for similar and dissimilar consonants separately. A second prediction (prediction 2) is that lexical bias is slightly stronger for dissimilar than for similar consonants. We find a significant positive lexical bias in the 'completed spoonerisms' for both similar and dissimilar consonants, but in the 'interruptions' we find a significant positive lexical bias only for the similar consonants ($p < .001$). There is a significant negative lexical bias for the dissimilar consonants ($p < .001$). In the 'competing errors' lexical bias is absent for both similar ($p = .254$), and dissimilar consonants ($p = .259$). The significant negative lexical bias for dissimilar consonants in the 'interruptions' cannot easily be accounted for by the 'feedback only' account.

A 'self-monitoring only' account predicts (prediction 1′) that, both for similar and for dissimilar consonants, a positive lexical bias in 'completed spoonerisms' is fully compensated by a negative lexical bias in the 'interruptions' and/or 'competing errors'. The absence of a significant negative lexical bias in both 'interruptions' and

Table 5
Raw counts (pooled over participants and items), broken down by response category, by the main conditions (WW or primed for word–word, vs. NN or primed for nonword–nonword, and phonetically similar vs. dissimilar), for Experiment 1

Condition	Response category					
	Fluent	Completed	Interrupted	Competing	Other	Total errors
WW, sim	677 (0.846)	36 (0.045)	28 (0.035)	23 (0.028)	36 (0.045)	123 (0.153)
WW, diss	906 (0.93)	24 (0.024)	20 (0.02)	20 (0.02)	32 (0.032)	96 (0.096)
Sum	1583 (0.878)	60 (0.033)	48 (0.027)	43 (0.024)	68 (0.038)	219 (0.122)
NN, sim	717 (0.894)	18 (0.022)	20 (0.025)	25 (0.031)	22(0.027)	85(0.105)
NN, diss	886 (0.886)	14 (0.014)	40 (0.04)	26 (0.026)	33 (0.033)	113 (0.113)
Sum	1603 (0.89)	32 (0.018)	60 (0.033)	51 (0.028)	55 (0.031)	198 (0.11)

Also the totals of the error categories are given, and the sums over similar and dissimilar per priming condition. Fractions (in parentheses) are expressed relative to the total number of responses, as in the multinomial logistic regression analysis.

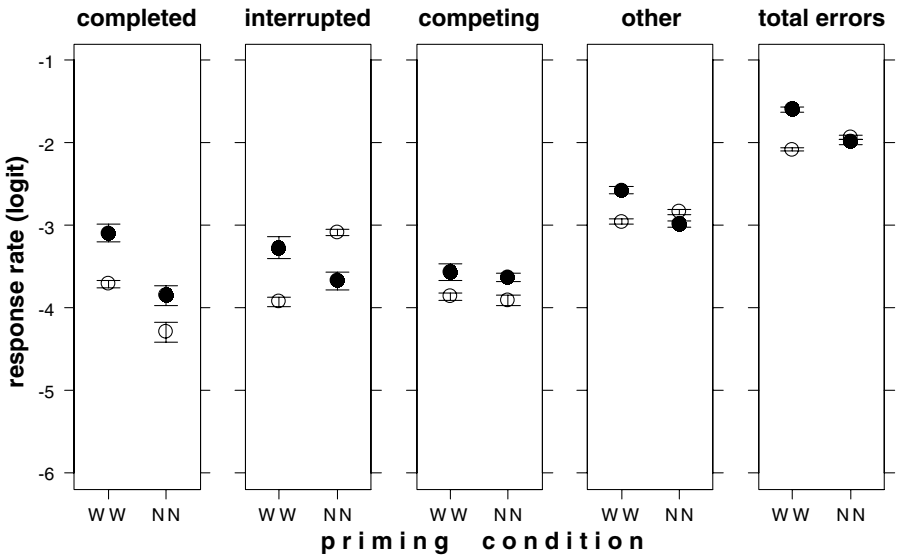

Fig. 2. Observed log-odds of estimated error rates in Experiment 1, broken down by priming condition (on the horizontal axis), by phonetic similarity or dissimilarity (filled and open symbols, respectively), and by response category (between panels). Error bars correspond to 95% confidence intervals of the bootstrapped logistic-regression coefficients (over 250 replications). Note that the mean of these bootstrapped coefficients may deviate from the observed error rate.

'competing errors' for similar consonants, where there is a positive lexical bias in the 'completed spoonerisms', remains unaccounted for by a 'self-monitoring only' account. The positive lexical bias for similar consonants in the 'interruptions' is particularly revealing, because this cannot be explained by 'covert error repairs' in inner speech, as interruptions are repaired overtly and not covertly. Thus the strong and significant positive lexical bias in the 'interruptions' pleads for an effect of immediate feedback of activation. This is very different for the conditions with dissimilar consonants where responses seem to behave as predicted by a 'self-monitoring' account.

A 'self-monitoring only' account also predicts (prediction 2′) for the nonword–nonword priming condition that the numbers of 'completed spoonerisms' are lower and the numbers of 'interruptions' and/or 'competing errors' are higher for dissimilar than for similar consonants (no such prediction was made for the word–word priming condition). Fig. 2 shows that, whereas in the word–word priming condition the error rate is systematically higher for similar than for dissimilar consonants for all three error categories (all $p < .001$), in the nonword–nonword priming condition the error rate is higher for similar than for dissimilar consonants in the 'completed spoonerisms' ($p < .001$), but indeed lower for similar than for dissimilar consonants in the 'interruptions' ($p < .001$), and about the same in the 'competing errors' (slightly but significantly higher for similar than for dissimilar consonants, $p < .001$). If these differences result from monitoring inner speech, they suggest that a lexicality criterion differentially affects the detection probability of errors with similar and dissimilar consonants. The lexicality criterion hardly operates on errors with similar consonants, but strongly affects detection probability of errors with dissimilar consonants. It seems that the lexicality criterion and the effect of similarity are not independent.

The 'feedback plus self-monitoring' account predicts (prediction 1″) that the lexical bias in 'interruptions' and/or 'competing errors' would be much smaller than that in the 'completed spoonerisms', absent or even negative, depending on the relative strength of feedback and self-monitoring as sources of lexical bias. A possible negative lexical bias in 'interruptions' and 'competing errors' would not compensate fully for the positive lexical bias in 'completed spoonerisms'. This prediction should hold for similar and dissimilar consonants separately. The results in Fig. 2 suggest that for similar consonants the positive lexical bias in 'interruptions' and/or 'competing errors' is smaller than in the 'completed spoonerisms'. This is compatible with the 'feedback plus self-monitoring account', under the assumption that the contribution to lexical bias of a lexicality criterion employed by the monitor is much weaker than the contribution of feedback. On the other hand, for the dissimilar consonants the positive feedback in the 'completed spoonerisms' is more than compensated by a negative lexical bias in the 'interruptions'. This suggests a very strong effect of a lexicality criterion in monitoring inner speech, but also suggests that a positive lexical bias in 'completed spoonerisms' is not simply mirrored by a

negative lexical bias in the 'interruptions' and 'competing errors'. Taken together, the current results support a 'feedback plus self-monitoring' account of lexical bias, as proposed by Hartsuiker et al. (2005).

As we have seen, in the 'competing spoonerisms' there is no significant positive or negative lexical bias. It seems that the 'competing spoonerisms' are not affected at all by the lexicality of the primed-for spoonerisms. How is this with 'other speech errors', not starting with the same consonant as the 'elicited spoonerisms', such as perseverations, seemingly unrelated other errors, and hesitations? For similar consonants there is a significant positive lexical bias ($p < .001$) in these 'other speech errors', for dissimilar consonants there is a significant negative lexical bias ($p < .001$). This is basically the same pattern as found for the 'interruptions'. Apparently, the frequency of these 'other speech errors' is sensitive to the lexicality of the primed-for spoonerisms.

Discussion

With respect to those predictions from our model of self-monitoring that do not relate to the three competing accounts of lexical bias, the data mainly support our model: So-called 'competing errors' are virtually limited to the test condition, 'other speech errors' are not. 'Interruptions' are virtually always followed by the correct targets, suggesting that in the early phase of self-monitoring the main competition is between expected spoonerisms and 'correct targets'. Response times for 'interruptions' are shorter (albeit not significantly) and for 'competing errors' longer than those for 'completed spoonerisms', suggesting that 'interruptions' may result from too hasty speech initiation, and that 'competing speech errors' may result from competition between expected spoonerisms with speech errors sharing the (primed-for) initial consonant with the spoonerism. That there is no difference between response times of 'completed spoonerisms' and 'other speech errors', suggests that the competition is mainly between 'completed spoonerisms' and those speech errors that share the initial consonant with the expected spoonerism. These findings support our suggestion that the model may be useful in an attempt to discriminate between the three competing accounts of lexical bias.

The data obtained in Experiment 1 cannot easily be reconciled with a 'feedback only' account due to the negative lexical bias for the dissimilar consonants in the 'interruptions' and in the 'other speech errors'. These effects can also not easily be explained by a 'self-monitoring only' account of lexical bias. This is so because for the similar consonants the positive lexical bias in the 'completed spoonerisms' is not at all mirrored by a negative lexical bias in the 'interruptions' and/or 'competing errors' and 'other speech errors', and not fully mirrored for the dissimilar consonants. The pattern in the data can most easily be accounted for by the 'feedback plus self-monitoring' account proposed by Hartsuiker et al. (2005).

The most striking result is the large difference between similar and dissimilar consonants. Similar consonants behave more or less as predicted by a 'feedback only' account, except for the absence of a positive lexical bias in the 'competing errors'. This suggests that, particularly for the 'interruptions' where there is little time, exchanges of similar consonants are rarely detected by the monitor. This agrees with Levelt's idea that the monitor employs the speech comprehension system. Detection of errors involving similar consonants would be much harder than detection of errors involving dissimilar consonants, particularly so under time pressure.

Dissimilar consonants, however, behave mainly as predicted by a self-monitoring account. Here the strongest negative lexical bias is in the 'interruptions', suggesting that errors involving dissimilar consonants are relatively easy to detect (even under time pressure), and that the criterion of lexicality is strongly affecting the early phase of self-monitoring (perhaps typically under time pressure). The effect of the lexicality criterion is absent in the 'competing errors'. This might mean that these 'competing errors' in reality had little or no competition with the expected spoonerisms. However, in view of the significant positive and negative lexical bias in the 'other speech errors' for similar and dissimilar consonants respectively, it is possible that the absence of a positive or negative lexical bias in the 'competing speech errors' effect is caused by two effects canceling each other out, viz. feedback and self-monitoring.

The negative lexical bias is also weaker in the 'other speech errors' than in the 'interruptions'. Whether or not this is related to the degree of time pressure, as hypothesized in the introduction, will be investigated in Experiment 2. The positive (for similar consonants) and negative (for dissimilar consonants) lexical bias in the 'other speech errors' comes somewhat as a surprise, given that the response times suggested little competition with the expected spoonerisms. It may be the case, however, that the effect seen in Fig. 2 is due to a minority of these 'other speech errors' and that most of the 'other speech errors' had little competition with the expected spoonerisms. This majority would then show no lexicality effect of feedback (because lexicality effect was defined in terms of the expected spoonerisms), and no effect of self-monitoring (for the same reason). These points will be taken up again in the discussion of Experiment 2.

Experiment 2

Experiment 2 was set up to test (a) the general predictions from our model of monitoring inner speech in a

further experiment, and (b) the hypothesis that the strength of the negative lexical bias in 'interruptions' is modulated by the time pressure on the participants. If participants are relieved from the considerable time pressure as exerted in Experiment 1, then the negative lexical bias is predicted to be much smaller or even absent in the 'interruptions', but much larger in the 'competing errors' and 'other speech errors'. This finding would suggest an explanation for the wide variation of lexical bias in so-called 'interrupted spoonerisms' in published experiments. It would also show that a positive lexical bias in 'interrupted spoonerisms' does not exclude that the monitor applies a lexicality criterion. It would rather show that a lexicality criterion is adaptable and, if circumstances change, can be directed at another phase of self-monitoring.

Methods

The method used was basically the same as the one applied in Experiment 1. However, some modifications were made that took away the need for correction and made the experiment more relaxed for the participants.

Stimulus material

Target word pairs for test stimuli and base-line stimuli, which in a few cases differ from those in Experiment 1, are given in Appendix B. The distribution over stimulus lists was identical to Experiment 1. Each word pair was either preceded by 3, 4, or 5 priming word pairs, chosen to prime a spoonerism, as in the sequence GIVE BOOK, GO BACK, GAS BAIT preceding the target word pair BAD GAME, or by 3, 4 or 5 nonpriming word pairs, providing a base-line condition. In this experiment the priming word pairs were not preceded by additional nonpriming word pairs, as was the case in Experiment 1 as an attempt to hide the purpose of the experiment from the participants. Note also that there were at least three precursor word pairs, whether priming (i.e. preceding test stimuli) or not priming (i.e. preceding base-line stimuli), so that participants might easily discover that they could relax during the first two precursor word pairs. In this experiment there were no fillers other than the base-line stimuli that were identical to the test stimuli in the other stimulus list. Practice items as in Experiment 1.

Participants

As in Experiment 1, but different individuals.

Procedures

The procedure was identical to the one in Experiment 1, including the new randomization of each stimulus list for each odd-numbered participant and the complementarity of odd-numbered and even-numbered participants, except for the following. After each target word pair the "?????"-prompt was visible during 900 ms, followed by a blank screen during 100 ms. There was no buzz sound before which participants had to respond, and no cue for correction. Participants were not urged to correct themselves. There were no fillers other than the base-line stimuli. Testing took approximately 8 min for each participant. Data were scored as in Experiment 1.

Results

The same feature coding and data analyses were used as in Experiment 1. A first breakdown of the numbers of responses obtained in Experiment 2 is given in Table 6. Pooling all errors, including 'completed spoonerisms', 'anticipations', 'interruptions' and 'competing errors', that start with the initial consonant of the second word on the one hand, and pooling all 'other errors' on the other hand, gives in the test condition 178:160 and in the base-line condition 13:206 errors. These distributions of course differ significantly ($\chi^2(1) = 128$, $p < .001$). This finding further supports the idea that many errors starting with the initial consonant of the second word are triggered by the priming of a consonant exchange. Again we have pooled the few anticipations with the 'completed spoonerisms', and concentrated further analysis on 'completed spoonerisms', 'interruptions', 'competing errors', and 'other speech errors'.

As in Experiment 1 and confirming our prediction (a), 'competing errors' (i.e. speech errors that are neither 'completed spoonerisms' nor 'interruptions', but that do start with the initial consonant of the second word) are far more numerous in the test condition than in the base-line condition: 67:3 (binomial: $p < .001$), suggesting that these also are triggered by priming a consonant exchange. Of these 67 'competing errors' in the test condition, only 6 showed a replacement with a nonword, 2 of which in the word–word and 4 in the nonword–non-

Table 6
Numbers of responses from Experiment 2, broken down by response category over rows, by priming condition (test vs. base-line) and by lexicality condition (WW or word–word outcome vs. NN or nonword–nonword outcome)

	Test		Base-line	
	WW	NN	WW	NN
(1) Fluent & correct	1682	1664	1733	1720
(2) Completed spoonerism	31	23	3	1
(3) Anticipation	0	1	2	0
(4) Interrupted spoonerism	30	19	2	2
(5) Competing error	26	36	2	1
(6) Perseveration	0	0	1	0
(7) Miscellaneous	38	64	66	76
(8) No response	29	29	27	36
Total	1836	1836	1836	1836

word priming condition. Again these numbers are very different for 'other speech errors', of which there are 102 in the test condition (13 nonword replacements) and 143 in the base-line condition (12 nonword replacements).

Response times were analyzed by means of mixed-effects regression modeling, yielding the coefficients and variances in Table 7. As predicted (predictions b and c) response times were significantly shorter for 'interruptions' (548 ms; $p = .012$) and significantly longer for 'competing errors' (675 ms; $p < .001$) than those for 'completed spoonerisms' (597 ms). Thus the idea that 'interruptions' result from too hasty articulation, and that 'competing errors' may be the results of two consecutive operations in inner speech, is supported in Experiment 2. As in Experiment 1, the response times for 'other speech errors' (587 ms, n.s.) are about the same as those of the 'completed spoonerisms'.

'Interruptions' were followed by the correct target (prediction d): This happened in 50 out of 54 'interruptions' in the test condition. The remaining four were not followed by anything. Again, these results support our model of monitoring inner speech.

Table 7
Estimated parameters for the mixed-effects regression of response times in Experiment 2. For fixed effects, regression coefficients are given, with standard errors and *t* values; for random effects, the variances and standard deviations are given (see text for details)

Fixed effects	Coefficients	*SE*	*t*
(Intercept)	502	8.0	62.3
Lexicality	−11	4.9	−2.29
Dissimilarity	−26	7.4	−3.53
Resp. completed	94	13.1	7.21
Resp. interrupted	46	13.9	3.32
Resp. competing	172	12.6	13.71
Resp. other	84	9.8	8.60
No. of neighbours	1	0.3	2.18
Lexic. × dissim.	10	6.9	1.44
Random effects	**Variance**	***SD***	**No. of obs.**
Participants	2611	51	102
Items	297	17	18
Residual	8719	93	3507

Competing explanations of lexical bias (2)

The data of Experiment 2 were analyzed as those in Experiments 1, in order to answer the same questions and to compare results across these experiments. Table 8 gives the raw counts of response categories and the fractions relative to the total number of overt responses.

Fig. 3 gives the estimated response rates in log-odds in Experiment 2, as obtained in the bootstrapped multinomial regression, broken down by priming condition, response category and phonetic similarity. Apart from other effects, lack of time pressure decreases the overall number of errors. The lower response rates in Experiment 2 may also explain the larger confidence intervals. Apparently, in those cells of the matrix where there are relatively few error responses, the average response rate cannot be determined very precisely, not even with the bootstrapped multinomial regression analysis. Nevertheless, the data allow some relevant conclusions.

The positive lexical bias in 'completed spoonerisms' is not significant for similar consonants ($p = .046$, n.s. after Bonferroni correction) but is significant for dissimilar spoonerisms ($p < .001$).

In the 'interruptions' in Experiment 1 we found a significant positive lexical bias for similar and a significant negative lexical bias in the dissimilar consonants. This difference in direction of lexical bias between similar and dissimilar consonants is not replicated in Experiment 2. Here, we find a significant positive lexical bias in the 'interruptions' for both similar ($p < .001$) and dissimilar ($p < .001$) consonants. If there is a lexicality criterion at work in this experiment, it certainly does not

Table 8
Raw counts (pooled over participants and items), broken down by response category, by the main conditions (WW or word–word, vs. NN or nonword–nonword primed, and phonetically similar vs. dissimilar), for Experiment 2

Condition	Response category					
	Fluent	Completed	Interrupted	Competing	Other	Total errors
WW, sim	709 (0.912)	21 (0.027)	14 (0.018)	13 (0.017)	20 (0.026)	68 (0.088)
WW, diss	931 (0.942)	10 (0.01)	16 (0.016)	13 (0.013)	18 (0.018)	57 (0.058)
Sum	1640 (0.929)	31 (0.018)	30 (0.017)	26 (0.015)	38 (0.022)	125 (0.07)
NN, sim	702 (0.905)	18 (0.023)	10 (0.013)	14 (0.018)	32 (0.041)	74 (0.095)
NN, diss	921 (0.93)	6 (0.006)	9 (0.009)	22 (0.022)	32 (0.032)	69 (0.07)
Sum	1623 (0.919)	24 (0.014)	19 (0.011)	36 (0.02)	64 (0.036)	143 (0.08)

Also the totals of the error categories are given, and the sums over similar and dissimilar per priming condition. Fractions (in parentheses) are expressed relative to the total number of responses, as in the multinomial logistic regression analysis.

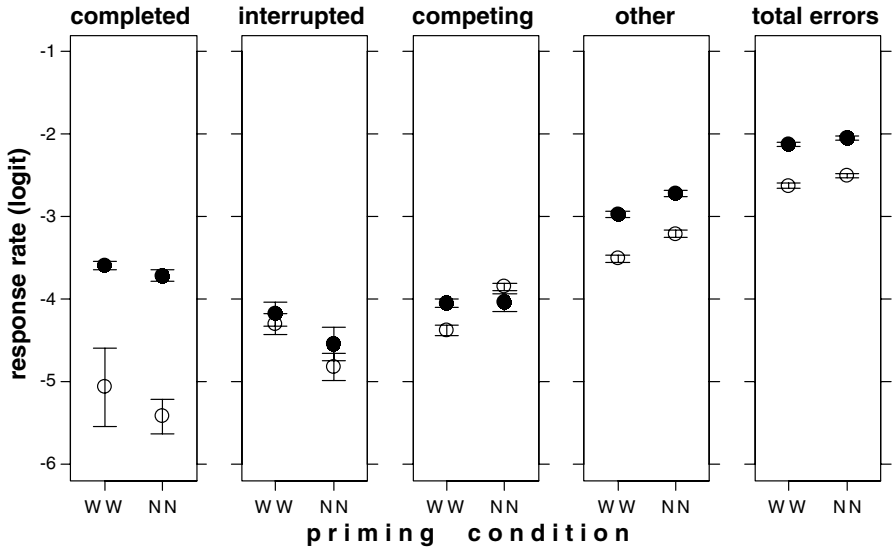

Fig. 3. Observed log-odds of estimated error rates in Experiment 2, broken down by priming condition (on the horizontal axis), by phonetic similarity or dissimilarity (filled and open symbols, respectively), and by response category (between panels). Error bars correspond to 95% confidence intervals of the bootstrapped logistic-regression coefficients (over 250 replications). Note that the mean of these bootstrapped coefficients may deviate from the observed error rate.

affect the 'interruptions' strongly. This is different for the 'competing errors'. In Fig. 3, we find no positive or negative lexical bias for similar consonants ($p = .9$), but, in contrast with Experiment 1, a strong and significant negative lexical bias for dissimilar consonants ($p < .001$).

Could the work load for the lexicality criterion have been switched from being directed mostly at 'interruptions' with dissimilar consonants in Experiment 1 to being directed mostly at 'competing errors' in Experiment 2? This is indeed suggested by the considerable positive lexical bias for both similar and dissimilar consonants in the 'interruptions' plus the strong and significant negative lexical bias for dissimilar consonants, in the 'competing errors'.

The category of 'other speech errors' shows, for both similar and dissimilar consonants, a strong and highly significant negative lexical bias. This is different from Experiment 1, in which these 'other errors' showed a positive lexical bias in similar and a negative lexical bias in dissimilar consonants. For both similar and dissimilar consonants the lexical bias is considerably more negative in Experiment 2 than in Experiment 1, which suggests that the lexicality criterion in monitoring inner speech has a stronger effect on these 'other errors' without than with time pressure.

Discussion

In Experiment 2 as in Experiment 1, testing those predictions from our model of self-monitoring that do not relate to the three competing accounts of lexical bias supports the model. 'Competing errors' are virtually limited to the test condition, other errors are not. Response times for 'interruptions' were again shorter and response times for 'competing errors' again longer than those for 'completed spoonerisms', now both significantly. 'Interruptions' are practically always followed by the correct target.

The main findings in Experiment 2 with respect to lexicality and similarity are the following. There is a positive lexical bias in the 'completed spoonerisms' which is insignificant for similar consonants but significant for dissimilar consonants. Where in the 'interruptions' in Experiment 1 we found a positive lexical bias for the similar and a negative lexical bias for the dissimilar consonants, in Experiment 2 we find a positive lexical bias for both similar and dissimilar consonants in the 'interruptions'. This cannot be explained from self-monitoring, because repairs or replacements of interrupted spoonerisms are made overtly, not covertly. We interpret this as evidence for a contribution of feedback to lexical bias, and also as evidence that under relaxed conditions a criterion of lexicality is not or only weakly directed at an early phase of self-monitoring, where it would affect the frequency of interruptions.

Whereas in the 'competing errors' in Experiment 1 we found no lexical bias for both similar and dissimilar consonants, in Experiment 2, a negative lexical bias was found in Experiment 2 for both similar and dissimilar consonants. This suggests that the absence of lexical bias in Experiment 1 may have been due to two mutually counteracting effects, viz. feedback causing a positive lexical

bias effect and self-monitoring causing a negative lexical bias effect. We interpret the significant negative lexical bias in the 'competing errors' of Experiment 2 as evidence that under more relaxed conditions the lexicality criterion has a stronger effect on a later phase of self-monitoring than it has under time pressure.

The pattern in the data of Experiment 2 cannot easily be reconciled with either a 'feedback only' or a 'self-monitoring only' account. The results show strong lexicality effects of a process that must operate before self-monitoring but also of a process that can best be identified with the monitor employing a lexicality criterion. This pattern of results is most compatible with a 'feedback plus self-monitoring' account as suggested by Hartsuiker et al. (2005).

The results of Experiment 2 again suggest that the items that may replace 'elicited spoonerisms' in inner speech are not limited to those speech errors that begin with the initial consonant of the second word. Apparently other active speech errors also regularly replace an 'elicited spoonerism'. The frequency of these 'other errors' is sensitive to the lexicality of the primed-for spoonerisms, causing a significant positive lexical bias in 'other errors' in the condition with similar consonants in Experiment 1 and causing a negative lexical bias in all other cases. This indicates that the primed-for spoonerisms have played a role in their history: Any lexicality effect in these experiments is defined in terms of the lexicality of the expected spoonerisms. The reader may also note that the patterns in both experiments show a stronger negative lexical bias for dissimilar than for similar consonants, in agreement with the self-monitoring account of lexical bias. For the 'other speech errors' this is even more suggestive than in case of the 'competing errors' because the 'other speech errors' themselves do not contain the similar or dissimilar consonants. The effect of similarity must be explained from completely hidden processes.

General discussion

The current approach to investigating possible causes of lexical bias in phonological speech errors, although drawing heavily on earlier work by many researchers, differs in some important respects from earlier attempts. Our main innovations here are (1) the simple flow chart model of monitoring inner speech described in the introduction and in Fig. 1, (2) separate analyses for 'completed spoonerisms', 'interruptions', 'competing errors' and 'other errors', (3) the use of phonetic distance between the two to-be-spoonerized consonants as an experimental factor, (4) measuring and analyzing response times as a function of error category, and (5) using multinomial logistic regression plus bootstrap replications for statistical analysis of the error rates.

As we have seen, the main properties of our flow chart model are supported by the data. In particular the idea that 'interruptions' reflect a relatively early phase and 'competing errors' a relatively late phase of self-monitoring is supported. The data suggest that in the early phase of self-monitoring the main competition is between expected spoonerisms and correct targets, whereas in the later phase of self-monitoring the main competition is between expected spoonerisms and both 'competing' and 'other speech errors'. Response times suggest that relatively many 'competing errors' (sharing the initial consonant with the expected spoonerism) compete with the expected spoonerisms, whereas a much smaller percentage of the 'other speech errors' (not sharing the initial consonant with the expected spoonerism) is affected by competition with the expected spoonerisms. This may seem in conflict with the finding that the lexicality effects in this latter category in absolute terms are not necessarily smaller than those in the 'competing speech errors'. However, the error rate for the 'other speech errors' is considerably higher. Therefore the same absolute effect can be obtained with a much smaller proportion of the errors.

The data in both experiments show, as expected, a positive lexical bias in the 'completed spoonerisms'. The question that concerns us here is what causes this positive lexical bias: Feedback, self-monitoring or both. We will first limit our discussion to the 'interruptions', because their role in the discussion of the controversy between different accounts of lexical bias is in some sense least controversial. It is relevant to the current discussion and worth repeating that 'interruptions' reflect the operation of monitoring inner speech and not overt speech. The reason is that the speech fragments before interruption are nearly always shorter than humanly possible reaction times (cf. Blackmer & Mitton, 1991; Nooteboom, 2005b; Hartsuiker, 2006). It is also relevant that interruptions are often followed by overt repairs with offset-to-repair times of 0 ms, showing that not only the interruptions but also the repairs were planned before speech initiation (Blackmer & Mitton, 1991; Nooteboom, 2005b).

Before we go into any more detail it may be good to point out that 'interruptions' are not rare. This means that many speech errors are detected in inner speech. It is also good to notice that whatever the effects of lexicality and phonetic dissimilarity on the probability that errors are detected and then interrupted, to be discussed in a moment, these factors have only relatively small effects on a basic rate of error detection. Imagine that, as suggested by Levelt (1989), the basic criterion for error detection was nonlexicality. In that case lexical errors would either not be detected at all, or by more time-consuming syntactic and/or semantic valuation. This is not what happens, because in the current experiments many real-word errors are detected even though

there is no useful syntactic or semantic context. In addition, it has been demonstrated elsewhere that real-word phonological errors are treated by the monitor as phonological and not as lexical errors (Nooteboom, 2005a). Lexical and nonlexical, similar and dissimilar errors are all detected relatively frequently and relatively quickly. This suggests that the monitor mainly relies on comparing the form in inner speech directly with the still active intended form (cf. Hartsuiker, 2006; Nooteboom, 2005a, 2005b). The relatively weak effects of nonlexicality and dissimilarity are superimposed on the relatively high basic detection rate, as apparent in the interruption frequencies.

One may, entirely reasonably, ask why we assume that in an error like G. . .BAD GAME, G. . . is the overt part of the spoonerism GAD BAME, and not of any other speech error starting with G. . . We have no proof for this, but we observe that the experiments were designed in such a way that the main competition is expected between correct target and the elicited spoonerism (consonant exchange). Second, it has been shown in an earlier experiment (Nooteboom, 2005b) and in the current experiments that the frequency of such 'interruptions' is highly sensitive to the lexicality of the 'elicited spoonerisms'. If the 'elicited spoonerism' were not involved, its lexicality could hardly play a role. Thirdly, it may be observed that virtually always the 'interruptions' are followed by the correct target, confirming that the main competition is between 'elicited spoonerism' and 'correct target'. Our suggestion that 'interruptions' are made when speech is initiated too hastily is supported by the finding that, at least in Experiment 2, response times are significantly shorter for 'interruptions' than for 'completed spoonerisms'. The error may be spoken so hastily because the priming in that case was very successful, so that the error is relatively active, although still in competition with the also very active correct target. ''Too hasty'', then means that speech initiation does not wait until the competition between error and target has been resolved by the monitor.

Unfortunately, earlier published experiments have never provided an analysis from which it could be concluded that there was either a significant positive or a significant negative lexical bias separately in the 'interruptions'. Our experiments show both a significant positive lexical bias, for similar consonants in Experiment 1 and for both similar and dissimilar consonants in Experiment 2, and a strong and significant negative lexical bias for dissimilar consonants in Experiment 1. The significant positive lexical bias in the 'interruptions' is particularly revealing, because it cannot be explained from the monitor covertly repairing nonlexical errors more frequently than lexical ones. 'Interruptions' are the immediate products of detecting errors in inner speech, but the repairs are overt, not covert. For all we know a positive lexical bias in the 'interruptions' can only be explained from a process preceding monitoring inner speech. Thus this positive lexical bias may be interpreted as evidence in favour of a contribution of feedback to lexical bias. The reader may note that nearly always when there is only a positive lexical bias, the error rate is higher for similar than for dissimilar consonants. This reflects the well-known phenomenon that in phonological speech errors similar phonemes substitute more easily for each other than dissimilar ones (Fromkin, 1971; MacKay, 1970; Nooteboom, 1973).

However, the strong and significant negative lexical bias in the 'interruptions' for the dissimilar consonants in Experiment 1 cannot be explained from feedback. This is best interpreted as betraying an effect of the monitor employing a lexicality criterion. This monitoring effect is absent for similar consonants, suggesting that under time pressure the monitor easily misses errors that are phonetically similar to their targets. This effect is absent in Experiment 2 for both similar and dissimilar consonants, suggesting that under more relaxed conditions detection of nonlexicality of errors in inner speech is turned away from an early phase of monitoring inner speech.

The variation in frequency of 'interruptions' may be ascribed to two mechanisms. On the one hand there is immediate feedback of activation between phonemes and word forms, affecting relative frequencies of lexical and nonlexical spoonerisms. On the other hand a criterion of lexicality is also employed in monitoring inner speech. The monitor in its turn follows two different routes. The most important route is direct comparison between error and intended target and for all we know this route is hardly affected by either lexicality or similarity. The other route is detection of nonlexical errors. Under time pressure, as in Experiment 1, detection of nonlexical errors is already active in the very early phase of monitoring inner speech, thereby affecting the frequency of 'interruptions'. However, if error and target are phonetically similar, detection nearly always fails. Only if error and target are phonetically dissimilar, detection of nonlexicality is so frequent that it completely overrules the underlying positive lexical bias, and then it causes a strong negative lexical bias in the 'interruptions'. Under more relaxed conditions, as in Experiment 2, detection of nonlexicality fails for both similar and dissimilar consonants, at least in the early phase of monitoring inner speech that is reflected in the interruption frequency.

In case a spoonerism in inner speech is not interrupted, it may yet be detected and then be replaced with something else. This may be the 'correct target' with which the spoonerism competes, certainly in the early phase of monitoring. This leads to a 'covert repair'. Relative frequency of 'covert repairs' of lexical and nonlexical phonological speech errors form the standard

explanation of lexical bias as suggested by Baars et al. (1975), Levelt (1989), Levelt, Roelofs, and Meyer (1999). Unfortunately, 'covert repairs' remain invisible in the error counts. However, we have speculated that an elicited error in inner speech may also be replaced with another speech error, a plausible candidate being a speech error sharing the initial consonant with the 'elicited spoonerism'. Admittedly, we can not be certain that all these 'competing errors' are reactions to 'elicited spoonerisms' in inner speech. Our finding that response times are significantly longer for 'competing errors' than for 'completed spoonerisms' may be taken to support the idea that at least many of these 'competing errors' are the result from two consecutive operations, first producing a spoonerism and subsequently replacing this spoonerism with another, competing, speech error. But there may also be other explanations for these longer response times. Because 'competing errors' virtually always involve real words, there is not only competition between error and correct target on the phonological level, but also on the lexical level. This lexical competition might be responsible for the longer response times. Also the finding that under time pressure there are relatively more 'interruptions' and under more relaxed conditions relatively more 'competing errors' does not necessarily mean that our 'competing errors' were preceded by 'elicited spoonerisms' in inner speech. However, they might. As in the case of the 'interruptions', the most convincing evidence is the finding that the frequency of 'competing errors' is sensitive to the lexicality of the expected spoonerisms. This can only be explained by assuming that these expected spoonerisms somehow played a role in the history of at least part of the 'competing errors'.

We find in both Experiment 1 and Experiment 2 that the basic rate of 'competing errors' is relatively high. For both similar and dissimilar consonants a lexical bias, positive or negative, is absent in the 'competing errors' in Experiment 1. Of course, this is not strong evidence in any way, because it might reflect that all or most of the 'competing errors' have nothing to do with the 'elicited spoonerisms'. Contrarily, the absence of any effect may be due to two underlying mutually counteracting effects, viz. feedback from phonemes to words causing a positive lexical bias and monitoring inner speech causing a negative bias. The interesting part is in Experiment 2. Although in that experiment again there is no positive or negative lexical bias for the similar consonants, there is a strong and significant negative lexical bias for the dissimilar consonants. Apparently, in Experiment 2 for the dissimilar consonants the frequency of 'competing errors' is sensitive to lexicality of the primed-for spoonerisms. It would be unlikely that this sensitivity arises only for the dissimilar consonants in Experiment 2, and not for the dissimilar consonants in Experiment 1, nor for the similar consonants in Experiment 1 and 2. On these grounds, the absence of lexical bias in the other three cases is tentatively interpreted as a composite of positive and negative lexical biases canceling each other out.

This tentative interpretation gets support from results in the category of 'other speech errors'. Under time pressure, in Experiment 1, these show a positive lexical bias in the condition with similar consonants, and a negative lexical bias in the condition with dissimilar consonants. This suggests once more that nonlexicality of the primed-for spoonerism is easily detected with dissimilar error and target, and not so easily with similar error and target. Under more relaxed conditions, in Experiment 2, we find a strong and significant negative lexical bias in the 'other speech errors' both in the condition with similar and the condition with dissimilar consonants. Remember that in Experiment 2 we found a positive lexical bias in the 'interruptions' for both similarity conditions, which suggests that without time pressure detection of nonlexicality is turned away from the early phase of monitoring inner speech (reflected by the 'interruptions'). We can now see that detection of nonlexicality has turned to both the 'competing errors', showing a significant negative lexical bias in the dissimilar condition, and, even stronger, to the 'other speech errors', showing a strong and significant negative lexical bias in both similarity conditions, thus compensating for the absence of any noticeable monitoring effect in the 'interruptions'. Apparently, detection of nonlexicality of errors in inner speech is a dynamic process, that can either be focused on an early or on a later phase of monitoring inner speech, depending on the amount of time pressure.

As far as we see now, the current results can most easily be interpreted by assuming that, as proposed by Hartsuiker et al. (2005) and recently defended by Hartsuiker (2006), lexical bias in phonological spoonerisms has two sources, viz. feedback of activation from phonemes to words and monitoring inner speech employing a lexicality criterion. We admit that this comes as a surprise to us. We had expected to find that monitoring inner speech is the sole source of lexical bias. The latter stand-point is defended in Nooteboom (2005b) and Nooteboom and Quené (in press). In some sense our current conclusion is unfortunate. As pointed out by Hartsuiker (2006), the need to assign two sources to lexical bias in phonological speech errors limits the conclusions one can draw from speech error patterns. Such patterns are used as evidence in the long standing debate on modularity versus interactivity in language production (Dell, 1986; Levelt et al., 1999; Rapp & Goldrick, 2000; Vigliocco & Hartsuiker, 2002). The problem is that it will not always be easy to know which quantitative aspects of the data are caused by feedback and which are caused by the mon-

itor. The problem is aggravated because of the very different behaviour in our experiments of errors that are phonologically similar and errors that are phonologically dissimilar to the targets. In most earlier experiments these two categories have been collapsed. This may well have hidden important quantitative aspects of the data that reflect the operation of underlying mechanisms.

If there is indeed feedback of activation from phonemes to words, the question arises where this feedback comes from. Stemberger (1985), Dell (1986), and Dell and Kim (2005) assume that there is immediate and automatic feedback of activation within the production system proper. Alternatively, it has been suggested that when the monitor employs speech perception, this possibly leads to feedback from phonemes to word forms via the perception of inner speech (Roelofs, 2004). Recently, evidence was found that, although production and perception processes do not share representations on the form level and the phonological level, there are indeed close links between production and perception, among other things feeding activation from perception back to production both on the phonological level and the word form level (Özdemir, Roelofs, & Levelt, in press, submitted for publication; Roelofs, Özdemir, & Levelt, in press). Thus feedback could be a side-effect of the inner perceptual loop employed by the monitor. Such feedback via inner speech would also explain why the monitor does not fully depend on global criteria of the form "is this a word?", as proposed by Levelt (1989). Feedback of this kind could enable the monitor to compare an error form in inner speech more or less directly with the intended form.

In our experiments we focused only on phonological errors. Levelt (1989) proposed that the monitor may also be directed at syntactic or semantic well-formedness and social appropriateness. Monitoring inner speech, however, mainly depends on phonological information. Slevc and Ferreira (2006) showed in a speech halting paradigm which supposedly taps important aspects of monitoring inner speech that monitoring success depends on phonological dissimilarity between error and target, but not on semantic dissimilarity (although emotional valence of the words involved did affect monitoring success). This result agrees with the strong effects of phonetic similarity in our experiments. The reason for the importance of phonological information possibly is that it is very rapidly available. Monitoring inner speech is under time pressure, because it attempts to detect and repair speech errors before they are spoken. This is different for repairs of overt lexical errors. It has been shown in a study of overt phonological and lexical speech errors and their repairs in spontaneous speech that speakers need more time before they stop speaking, and that they backtrack further, after a lexical than after a phonological speech error (Nooteboom, 2005a). Possibly, in monitoring inner speech nonphonological criteria are used rarely, simply because there is not enough time.

With respect to the role of phonological or phonetic information in inner speech, a remarkable result was recently obtained by Oppenheim and Dell (in press). Using a paradigm eliciting overt phonological errors or phonological errors in inner speech not prepared to be spoken, the authors found that whereas overt speech errors show both a strong lexicality effect and a strong phonological similarity effect, errors in inner speech not prepared to be spoken, show the lexicality effect only. The authors conclude that (silent) inner speech is impoverished at lower (featural) levels, but robust at higher (phonemic) levels. If this is correct, it implies that the effect of sound dissimilarity on monitoring success in our experiments is a lower level phonetic effect rather than a higher level phonological effect. This would also explain why Wheeldon and Levelt (1995), who had participants monitor their own internal speech without a speaking task, did not find lower level phonetic effects. They concluded that in their experiments participants monitored their internal generation of an abstract syllabified phonological representation. The findings by Oppenheim and Dell suggest that this result cannot be generalized to inner speech prepared for being spoken.

Part of our results we have attributed to the operation of monitoring inner speech. Recently, Hartsuiker (2006) has formulated three prerequisites for a satisfactory account of any monitoring bias, viz. (1) the proposed account poses functional monitoring criteria; (2) the bias can be altered by manipulations affecting monitoring performance; (3) the monitoring bias occurs also in perception. The monitoring bias we are dealing with here, is a lexicality bias. We find only weak evidence for a lexicality bias caused by the monitor with similar consonants, but strong evidence with dissimilar consonants. This in itself can be explained from assuming that nonlexicality is detected by the speech perception system (i.e. by Levelt's speech comprehension system), employing word recognition. When error and target are similar in sound, the difference may be easily missed; when they are dissimilar in sound, detection of nonlexicality is much more likely. The resulting bias for detecting nonlexicality (when error and target are dissimilar) is functional in the sense that it allows rapid detection of an error, given that nonlexical forms are very rare in the context.

Hartsuiker's second prerequisite is also met: We have seen that the positive or negative lexical bias can be altered by manipulating the time pressure for the participants: Under time pressure detection of nonlexicality mainly affects the frequency of 'interruptions',

under more relaxed conditions it affects mainly the frequency of both 'competing errors' and 'other speech errors'. This can be seen as an attentional effect. Attentional differences might influence the efficiency of the monitor. However, Rapp and Goldrick (2004) noted that the most obvious prediction from a less efficient monitor is that there are fewer corrections across the board (something we also found), and not that the corrections would be qualitatively different. This attentional explanation leaves the qualitative differences between our Experiments 1 and 2 unexplained. It also leaves unexplained that there are qualitative differences in the pattern of error detection frequencies between inner speech and overt speech (Nooteboom, 2005a). We propose that the monitoring strategy of detecting nonlexicality, being superimposed on a monitoring strategy that compares error with intended form, can, under the influence of time pressure, be directed at or be directed away from the early phase of monitoring inner speech.

Hartsuiker's third prerequisite is that the monitoring bias should also occur in speech perception. This prerequisite has the qualification (stated by Hartsuiker) that listening to others and monitoring oneself differs. Because of this qualification, it is safest to interpret this prerequisite as meaning that the same bias should occur in monitoring inner speech and monitoring overt speech for speech errors. Monitoring one's own overt speech clearly involves speech perception. Nooteboom (2005a) found no lexicality effect in self-monitoring of overt speech. Perhaps Hartsuiker's requirement for a perceptual analogue of a monitoring bias employed in monitoring inner speech is too strong. This may be so because (a) the function of monitoring inner speech is very different from the function of monitoring overt speech (cf. Nooteboom, 2005a), (b) the information available to the monitoring system is very different in the two situations (Hartsuiker, 2006), and (c) the time-constraints are very different. Monitoring inner speech must be relatively fast because speech errors should be detected and repaired before speech initiation. Therefore the route via detection of nonlexicality may be a helpful strategy for speeding up error detection. Monitoring overt speech is more relaxed, and should be focused primarily on detecting and repairing those errors that otherwise would harm communication. This goal may be very similar to detection of errors in other-produced speech.

Our starting-point in this paper was in a simple flow chart model of what may happen to an 'elicited spoonerism' in monitoring inner speech, exemplified in Fig. 1. The data in Experiments 1 and 2 support this model. We then used this model to derive a number of predictions from three alternative accounts of lexical bias in phonological speech errors. The data to our surprise rather strongly support the proposal by Hartsuiker et al. (2005) and Hartsuiker (2006) that lexical bias has two sources, (1) feedback of activation from phonemes to words and (2) self-monitoring of inner speech employing a criterion of lexicality. Furthermore the results show that the criterion of lexicality in self-monitoring affects detection frequencies far more strongly when error and target are dissimilar than when these are similar: Varying phonetic similarity yields different patterns of detection frequencies. We also found evidence that the precise pattern in the data strongly co-varies with the degree of time pressure under which the participants operate. These findings make it clear that patterns of error detection frequencies in inner speech are variable and not always easy to interpret. However, if our interpretations are valid, then errors that are similar to the targets reveal more about the effects of feedback on speech production, and errors that are dissimilar reveal more about the effects of self-monitoring on speech production.

Appendix A. Target and filler word pairs in Experiment 1

See Tables A1–A3.

Table A1
Target word pairs ("w1" and "w2") used in Experiment 1, in Dutch orthography

Lexical outcomes (WW)				Nonlexical outcomes (NN)			
w1	nbd	w2	nfeats	w1	nbd	w2	nfeats
bak	36	zoon	2	ban	23	zool	2
beuk	20	pol	1	beun	13	por	1
dom	23	gaar	3	dol	36	gaaf	3
doos	26	bel	1	doof	18	bed	1
fuik	12	bit	2	fuif	07	bil	2
geit	11	been	3	gijn	11	beet	3
kaal	33	duif	2	kaap	26	duim	2
kan	24	peer	1	kam	26	peen	1
keek	24	baas	2	keel	23	baat	2
ken	30	zooi	3	kef	19	zoog	3
paf	18	kiep	1	pal	28	kiem	1
pier	28	vaal	2	piek	24	vaag	2
pin	28	tof	1	pit	33	tos	1
tol	33	veer	3	top	31	veeg	3
vijl	24	kat	3	vijg	04	kap	3
voet	26	zeen	1	voer	22	zeep	1
zaal	25	boom	2	zaag	06	boot	2
zoen	23	puil	3	zoek	19	puin	3

The same word pairs were used as test word-pairs (primed for spoonerisms) and as baseline stimuli (not primed for spoonerisms). For each word pair, the difference in the number of classical distinctive features between the two initial consonants ("nfeat") and the number of nearest neighbours of the first word ("nbd") is also provided.

Table A2
Filler word pairs ("w1" and "w2") used in Experiment 1, in Dutch orthography

Lexical outcomes (WW)		Nonlexical outcomes (NN)	
w1	w2	w1	w2
git	dek	gil	dep
haan	lijs	haam	lijp
kit	waan	kir	waag
loog	haat	loof	haar
rib	wen	rif	weg
rik	loot	ring	loon
ruim	liep	ruin	lies
ruis	heet	ruik	heem
wak	hel	was	hef
wijn	ruit	wijf	ruig
woef	leen	woed	looi

These word pairs were preceded by 1 to 4 priming or non-priming word pairs (see text for details).

Table A3
Filler word pairs ("w1" and "w2") used in Experiment 1, in Dutch orthography

w1	w2
baar	vief
deeg	biet
deur	bies
dijn	koor
heil	noor
hoop	laai
hor	weef
hos	gup
jaag	hof
look	haas
maak	juk
mik	reeg
moet	neut
mom	vit
puim	boef
riem	dof
ris	meel
ros	feil
vaam	kien
vaat	tip
vet	pot
vim	kil
ving	kog
wieg	keus

These word pairs were not preceded by priming or non-priming word pairs (see text for details).

Appendix B. Target and filler word pairs in Experiment 2

See Table B1.

Table B1
Target word pairs ("w1" and "w2") used in Experiment 2, in Dutch orthography

Lexical outcomes (WW)				Nonlexical outcomes (NN)			
w1	nbd	w2	nfeats	w1	nbd	w2	nfeats
bad	38	pol	1	bar	29	por	1
bak	36	zoon	2	ban	23	zool	2
dom	23	gaar	3	dol	36	gaaf	3
doos	26	bel	1	doof	18	bed	1
fuik	12	bit	2	fuif	07	big	2
geil	11	been	3	gijn	11	beet	3
kaal	33	duif	2	kaap	26	duim	2
kan	24	peer	1	kam	26	peen	1
keer	30	baas	2	keel	23	baat	2
ken	30	zooi	3	kef	19	zoog	3
paf	18	kies	1	pal	28	kiem	1
pier	28	vaal	2	piek	24	vaag	2
pin	28	tof	1	pis	31	tos	1
tol	33	veer	3	top	31	veeg	3
vijl	24	kast	3	vijg	04	kant	3
voet	26	zuil	1	voer	22	zuid	1
zaal	25	boom	2	zaag	06	boot	2
zoen	23	puil	3	zoek	19	puin	3

The same word pairs were used as test word-pairs (primed for spoonerisms) and as baseline stimuli (not primed for spoonerisms). For each word pair, the difference in the number of classical distinctive features between the two initial consonants ("nfeat") and the number of nearest neighbours of the first word ("nbd") is also provided.

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jml.2007.05.003.

References

Baars, B. J. (1980). On eliciting predictable speech errors in the laboratory. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 307–317). New York: Academic Press.

Baars, B. J., & Motley, M. T. (1974). Spoonerisms: Experimental elicitation of human speech errors. *Journal Supplement Abstract Service, Fall 1974. Catalog of Selected Documents in Psychology, 3*, 28–47.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 14*, 382–391.

Bates, D. (2005). Fitting linear mixed models in R: Using the lme4 package. *R News, 5*, 27–30.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition, 39*, 173–194.

Costa, A., Roelstraete, B., & Hartsuiker, R. J. (2006). The lexical bias effect in bilingual speech production: Evidence for feedback between lexical and sublexical levels across languages. *Psychological Bulletin & Review, 13*, 612–617.

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

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5*(4), 313–349.

Dell, G. S., & Kim, A. E. (2005). Speech errors and word-form encoding. In R. Hartsuiker, Y. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech* (pp. 17–41). Hove: Psychology Press.

Dell, G. S., & Reich, P. A. (1980). Toward a unified model of slips of the tongue. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 273–286). New York: Academic Press.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: an analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior, 20*, 611–629.

Del Viso, S., Igoa, J. M., & Garcia-Albea, J. E. (1991). On the autonomy of phonological encoding: evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research, 20*, 161–185.

Efron, B., & Tibshirami, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fromkin, V. F. (1971). The nonanomalous nature of anomalous utterances. *Language, 47*, 27–52.

Garrett, M. F. (1976). Syntactic process in sentence production. In R. J. Walker & E. C. T. Walker (Eds.), *New approaches to language mechanisms* (pp. 231–256). Amsterdam: North-Holland.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). Edward Arnold: London.

Hamm, S., Junglas, K., & Bredenkamp, J. (2004). Die Zentrale Exekutive als präartikulatorische Kontrollinstanz [The central executive as a prearticulatory control device]. *Zeitschrift für Psychologie, 212*, 66–75.

Hartsuiker, R. J. (2006). Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes, 21*, 856–891.

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

Hosmer, D. W., & Lemeshow, S. (2000). In *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Humphreys, K. (2002). Lexical bias in speech errors. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Laver, J. D. M. (1973). The detection and correction of slips of the tongue. In V. A. Fromkin (Ed.), *Speech errors as linguistic evidence* (pp. 132–143). The Hague: Mouton.

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. (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.

MacKay, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia, 8*, 323–350.

MacKay, D. G. (1992). Constraints on theories of inner speech. In D. Reisberg (Ed.), *Auditory imagery* (pp. 274–285). Hillsdale: Lawrence Erlbaum Associates.

Motley, M. T., & Baars, B. J. (1976). Semantic bias effects of verbal slips. *Cognition, 4*, 177–187.

Motley, M. T., Camden, C. T., & Baars, B. J. (1982). Covert formulation and editing of anomalies in speech production: Evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior, 21*, 578–594.

Nooteboom, S. G. (1973). The tongue slips into patterns. In V. A. Fromkin (Ed.), *Speech errors as linguistic evidence* (pp. 144–156). The Hague: Mouton.

Nooteboom, S. G. (2005a). 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. (2005b). Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech. *Speech Communication, 47*, 43–58.

Nooteboom, S. G., & Quené, H. (in press). The SLIP technique as a window on the mental preparation of speech: Some methodological considerations. In M. J. Solé, P. Beddor & M. Ohala (Eds.), *Experimental Approaches to Phonology*. Oxford: Oxford University Press.

Oppenheim, G. M., & Dell, G. S. (in press). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect.

Özdemir, R., Roelofs, A., & Levelt, W. J. M (in press). Perceptual uniqueness point effects in monitoring internal speech. *Cognition*.

Özdemir, R., Roelofs, A., & Levelt, W. J. M. (submitted for publication). The locus of phonological facilitation from spoken distractors in picture naming.

Pampel, F. C. (2000). *Logistic regression: A primer, Quantitative applications in the social sciences; 07-132*. Thousand Oaks, CA: Sage.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-plus*. New York: Springer.

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

Quené, H. (2007). Analyzing multinomial repeated-measures data with two-stage bootstrapped logistic regression. Manuscript in preparation.

Quené, H., & Van den Bergh, H. H. (submitted for publication). Examples of mixed-effects modeling with crossed random effects and with binomial data.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication, 43*, 103–121.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 460–499.

Rapp, B., & Goldrick, M. (2004). Feedback by any other name is still interactivity: A reply to Roelofs (2004). *Psychological Review, 111*, 573–578.

Roelofs, A. (2004). Comprehension-based versus production-internal feedback in planning spoken words. A rejoinder to

Rapp and Goldrick (2004). *Psychological Review, 111*, 579–580.

Roelofs, A., Özdemir, R., & Levelt, W. J. M. (in press) Influences of spoken word planning on speech recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer.

Slevc, L. R., & Ferreira, V. S. (2006). Halting in single word production: A test of the perceptual loop theory of speech monitoring. *Journal of Memory and Language, 54*, 515–540.

Stemberger, J. P. (1985). An interactive activation model of language production. In A. W. Ellis (Ed.). *Progress in the psychology of language* (Vol. 1, pp. 143–186). London: Erlbaum.

Van Alphen, P. M. (2004). Perceptual relevance of prevoicing in Dutch. Unpublished doctoral dissertation, Radboud University, Nijmegen, The Netherlands.

Van de Velde, H., Gerritsen, M., & Van Hout, R. (1995). De verstemlozing van de fricatieven in het Standaard-Nederlands. Een onderzoek naar taalverandering in de periode 1935–1993 [The devoicing of fricatives in Standard Dutch. An investigation of language change in the period 1935–1993]. *De Nieuwe Taalgids, 88*, 422–445.

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in language production. *Psychological Bulletin, 128*, 442–472.

Wheeldon, L. R., & Levelt, W. J. M. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language, 34*, 311–334.