The SLIP technique as a window on the mental preparation of speech: Some methodological considerations.

Sieb Nooteboom and Hugo Quené,

Utrecht University, Utrecht Institute of Linguistics OTS

Abstract:

This chapter proposes some improvements on a method for eliciting speech errors, the so-called SLIP technique, including the use of multi-level logistic regression for data analysis. This is demonstrated in an experimental test of a new theory of self-monitoring as the main cause of lexical bias in phonological speech errors.

Key words:

SLIP technique, speech errors, lexical bias, self-monitoring, logistic regression

1. Introduction

This chapter is about the so-called SLIP (Spoonerisms of Laboratory-Induced Predisposition) technique for eliciting spoonerisms. This technique was introduced in 1974 by Baars et al. (1975) and has since been used by a number of researchers to study aspects of the mental production of speech. Recently the technique has seen a revival. A major question in many of the published studies is: “What is the cause of lexical bias?” Lexical bias is the phenomenon by which phonological speech errors more often lead to real words than to nonwords, when a priori probabilities are equal. According to some, lexical bias is caused by feedback of activation between phonemes and words (Stemberger 1985, Dell 1986), whereas others claim that lexical bias is caused by self-monitoring of inner speech rejecting and correcting nonwords more often than real words (Levelt 1989; Levelt et al. 1999). This controversy is important because it reflects two different ways of looking at the architecture of the mental processes involved in human speech production. Researchers in the feedback camp believe that, although one may distinguish between different hierarchically related components of the speech production system, such as a component for retrieving and ordering lexical items and a component for retrieving and ordering phonemes, the hierarchical relation is only partial because there is interaction in the form of immediate feedback of activation between successive components. Researchers in the self-monitoring camp, on the contrary, are convinced that there is no immediate feedback between successive components of the production system.

For that reason, they have to account for lexical bias with a mechanism other than feedback. The alternative explanation is found in self-monitoring of inner speech detecting, rejecting, and repairing nonwords more often than real words. This long standing debate has not been settled, mainly because of some problems with the SLIP technique and how the technique is used. The aim of this chapter is three-fold: 1) to discuss some of these problems and make some suggestions as to how to deal with them in future research, 2) to present a new view of what might happen to elicited spoonerisms in inner speech that potentially leads to a new way of analyzing data obtained in SLIP experiments, and 3) to briefly present a re-analysis of some data obtained in an earlier experiment reported in Nooteboom (2005b) supporting the assumption that the main cause of lexical bias is to be found in self-monitoring. But first we will give some more information on the SLIP technique and the lexical bias effect.

2. The SLIP technique and the lexical bias effect

The SLIP technique was introduced by Baars and Motley (1974), and used by Baars et al. (1975) to study among other things lexical bias in phonological speech errors. The technique was inspired by the observation that inappropriate lexical responses may result from anticipatory biasing: A child asks another child to repeat the word “poke” many times, then asks: “what is the white of egg called?” In this way the (wrong) answer “yolk”, induced by the rhyming relation with “poke”, may be elicited

(Baars 1980). The SLIP technique works as follows: Participants are successively presented visually, for example, on a computer screen, with word pairs such as DOVE BALL, DEER BACK, DARK BONE, BARN DOOR, to be read silently. On a prompt, for example a buzz sound or a series of question marks ("?????"), the last word pair seen (the test word pair as opposed to the priming word pairs), 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 a word pair like BARN DOOR will be mispronounced as DARN BORE, as a result of phonological priming by the preceding word pairs.

When the SLIP technique is used to study lexical bias, two types of stimuli are compared, stimuli generating lexical spoonerisms, such as BARN DOOR turning into DARN BORE, and stimuli generating nonlexical spoonerisms, such as BAD GAME turning into GAD BAME. Although both types of stimuli are equally frequent, responses commonly show that lexical spoonerisms are made more frequently than nonlexical ones. This lexical bias effect has not only been found in errors elicited in the laboratory with the SLIP technique, but was also clearly demonstrated in spontaneous speech errors (Dell 1986; Nooteboom 2005a), despite some failures to do so (Garrett 1976; Del Viso et al. 1991). Baars et al. (1975) explained lexical bias from "prearticulatory editing" of inner speech, more particularly from the assumption that in inner speech, speech errors forming nonwords are more often covertly detected and repaired than speech errors forming

real words. However, Stemberger (1985) and Dell (1986) explained lexical bias in phonological speech errors from immediate feedback of activation from phonemes to words for the mental preparation of speech. Such immediate feedback causes reverberation of activation between the planned phoneme string and all word forms containing this phoneme string or part of it. This gives an advantage in activation to phoneme strings corresponding to real words over those corresponding to nonwords, because the latter have no lexical representations. This advantage in activation would explain (among other things) the phenomenon of lexical bias.

The idea that there is immediate feedback from phonemes to lexical representations is not generally accepted. Notably Levelt (1989) and Levelt et al. (1999) assume that the mental preparation of speech is strictly serial and feedforward only, allowing no feedback between different levels of processing. These authors essentially support the original assumption by Baars et al. (1975) that lexical bias is caused by prearticulatory editing or self-monitoring of inner speech. Levelt assumes that this self-monitoring of inner speech employs the same speech comprehension system that is also used for self-monitoring of overt speech, and for listening to other-produced speech. The self-monitoring system is supposed to employ a general criterion of the form “Is this a word?” Therefore nonwords are more often covertly suppressed and repaired before overt production than real words. This would explain lexical bias in phonological speech errors.

To complicate matters, recently Hartsuiker et al. (2005) found experimental evidence that led them to assume that the relative frequencies of real-word and nonword speech errors are affected both by immediate feedback and by self-monitoring of inner speech. Their basic finding is that in a well-controlled experiment eliciting word-word and nonword-nonword spoonerisms with the SLIP technique testing for lexical bias, 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), 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. This suppression of real words in the nonlexical context is explained by adaptive behavior of the self-monitoring system, but 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). Thus now we have three competing views on the possible origin of lexical bias in phonological speech errors: 1) immediate feedback of activation alone, 2) self-monitoring of inner speech alone, 3) both immediate feedback of activation and self-monitoring of inner speech. It is important to realize that in principle feedback and self-monitoring are successive processes that do not exclude each other. Before self-monitoring operates, feedback has the effect that more real-word than nonword errors are made (Dell 1986). In the absence of feedback, numbers of real-word and nonword errors are

supposed to be equal before self-monitoring. We look for a way to count real-word and nonword errors in inner speech before self-monitoring operates.

3. Some problems with the SLIP technique

Obviously, it is still controversial whether or not there is feedback of activation from phonemes to words, and whether lexical bias is caused by such immediate feedback or by self-monitoring of inner speech, or by both. A major problem in putting an end to this long standing controversy one way or the other is that the SLIP technique is only marginally successful in generating spoonerisms of the primed-for kind. This has led to ways of analyzing the data that may have obscured important strategies used by participants in a SLIP task. Notably, instead of focusing on the predicted spoonerisms (BARN DOOR > DARN BORE), Baars et al. (1975), and in their wake most later researchers, in order to assess whether there are more speech errors made in the word-word than in the nonword-nonword priming condition, have collapsed “full exchanges” and “partial spoonerisms”. “Full exchanges” are exchanges of the two initial consonants without regard for what happens in the remainder of the two words as in BARN DOOR > DARN BORE or DARK BOARD or DARK BO.. where in the latter example one does not know how the second word would end or ended in inner speech. “Partial spoonerisms” are cases like BARN DOOR > DARN DOOR, but also DA.. BARN DOOR where overt speech was initiated too hastily and then interrupted.

Even given the practice of collapsing different error types, the numbers and percentages of spoonerisms obtained are not impressive and vary widely, from c. 8% "full exchanges" in one of the experiments described by Baars et al. (1975) to only 1% in one of the experiments by Dell (1986), 100% being the number of test trials. The variation in yield appears to be related to the method of stimulus presentation. More particularly, constant time pressure during the experiment appears to give more errors than more relaxed conditions. Notably, Baars et al. gave their participants no time for repairs, thus keeping time pressure on during the whole experiment. Dell, who wanted to make certain that elicited spoonerisms were not reading errors, followed the following procedure: After test stimulus presentation a sequence of question marks appeared on the screen as a signal for the participant to speak the last word pair seen aloud. The onset of the question marks was followed by a buzz signal after 500, 700, or 1000 ms. Participants were instructed to speak the last word pair seen before the buzz sound. Then, 500 ms after the buzz, for 2500 ms, the participant saw the phrase "Did you make an error?", and the participant answered this with "yes" or "no". Then the phrase "repeat words" was presented, also for 2500 ms, upon which the participant repeated the words that were, according to her/him, the correct words seen. The participant pressed a key on the terminal keyboard when she/he was ready to continue. This procedure made it possible to remove all errors from the data that might have been reading errors, under the assumption that, when participants were not aware of a discrepancy between the target seen and their spoken erroneous

response, the error might have been a reading error. The yield of this procedure was very low, certainly for the time limit of 1000 ms (0.8% of all test stimulus presentations). The same time limit was used in the original experiments by Baars et al. (1975), where the yield was, in experiment II, an order of magnitude higher (8.2%). The low yield in Dell's experiment may be attributed to the leisurely procedure with always more than 5000 ms between each test stimulus word pair and the following biasing word pair. It may also be relevant that it is uncertain how well the time pressure induced by the buzz sound following the test stimulus worked. Dell mentions that he removed responses starting after the onset of the buzz sound. This suggests that there were many cases where the participant certainly did not finish the response before the onset of the buzz sound, as they were instructed to do. As Dell's extensive measures to ensure that possible reading errors were removed from the responses apparently led to an extremely low yield, it seems more attractive to go along with the plausible assumption by Baars et al. that reading errors resulting in initial consonant exchanges are unlikely.

Similar to this comparison between two extremes, other differences between published experiments in percentages of elicited spoonerisms can potentially be explained by differences in the method of stimulus presentation and instruction to the participants. So the relatively low yield (4% and 2.29%) in the two experiments by Hartsuiker et al. (2005) is possibly related to their attempts to hide the purpose of the experiments from the participants by mixing the priming word pairs that preceded the

target word pair with non-priming word pairs. However, even the maximum yield of 8% “full exchanges” is low. The problem of scarcity of data is aggravated when there are good reasons not to collapse speech errors of different types. Of course, one way of dealing with this problem is simply to obtain more measurements. This can be done in two ways: presenting more test stimuli per participant, or running more participants. However, the requirements on the test stimulus language material usually restrict the range of possible word pairs to be used. This is so because a number of properties of words and word pairs affect the propensity of word pairs to elicit spoonerisms. This propensity has been shown to depend on the transitional probability of initial consonants and vowels in the word pair (Motley and Baars 1975), whether the vowels in the two words are the same (Dell 1986), frequency of usage of the stimulus words involved (Dell 1990), phonetic distance between the two initial consonants (Nooteboom 2005b), semantic relations between the biasing word pairs and the predicted spoonerisms (Motley and Baars 1976), and even phonological patterns particular to the stimulus word pairs in the experiment in question (Dell and Warker 2004). Although it is impossible to keep word pairs constant on all the properties mentioned, one should at least try to balance differences in these properties over different experimental conditions. This aspect of setting up SLIP experiments generally seems to get less attention than it deserves.

With the possibilities for using many more test stimuli being thus limited, more data should be obtained by running more participants per experiment. However,

in doing this the number of elicited relevant speech errors per participant remains low. In fact, it is a common finding that in a SLIP experiment, whereas some participants make a satisfactory number of errors, there are others who never make a relevant speech error. Also, it cannot be excluded that participants differ not only in their tendency to spoonerize, but also in the kind of speech errors they make. If this is so, there is a fair chance that in a blocked design, apparent significant differences between experimental conditions or between relative numbers of error types may have their real source in differences between participants. As far as we see, the only way to deal with this problem is to use as many participants as possible, and test whether large enough non-overlapping subgroups in the same experimental condition show the same behavior. This is bad news for those who want to economize on these experiments.

These quantitative problems with the SLIP task are, of course, closely related with the chosen method of statistical analysis. In publications reporting SLIP experiments the most frequent statistical methods are (a) the Mann-Whitney U test, the Wilcoxon signed ranks test, and the binomial (or sign) test, when only the difference between two conditions is to be assessed, (b) the traditional analysis of variance (ANOVA), when an interaction between independent variables is of interest, and (c) the chi^2 (chi square) test, if one is interested in knowing whether distributions of error types are the same or different across conditions. Of these the Mann-Whitney U test, the Wilcoxon test, and the binomial test are the most straightforward, although

here also one should be aware that the scarcity and strong variability of data per participant may in some cases lead to spurious significance (capitalization on chance). Therefore it might be advisable to test whether non-overlapping subgroups give the same results. The use of traditional analysis of variance is not recommended for analyzing SLIP data, because the data is binomial (for which it was not designed) and the design matrix has more than 90% of the cells with zeroes. This causes the data to deviate strongly from the normal distribution. It also brings the power of the analysis down often to below 0.3, making spurious acceptance of the null hypothesis unacceptably probable. Finally, the chi^2 test requires that the observations counted in the contingency table are independent (Devore and Peck 2005). Given that the same participants contribute to each cell, and that participants tend to differ in the kind of errors they make, independence can not be guaranteed. Here also, the problem might be alleviated by repeating the test for non-overlapping subgroups of participants or by repeating the experiment with different participants.

Clearly, problems with the various statistical techniques in analyzing data obtained with the SLIP task arise mainly because of the variability of both stimuli and participants in their behavior, combined with the scarcity of data per participant and per stimulus. In our opinion, the most promising technique for analyzing such data is multi-level logistic regression, which we will briefly introduce in section 5. The main point in this section is that, if the SLIP task is to be used in further research on the possible origins of the lexical bias effect, each experiment requires a sufficient

number of participants to test whether the relevant patterns in the data occur in independent subgroups of participants.

4. A new view of what may happen to elicited spoonerisms in inner speech

In the self-monitoring account of lexical bias one would predict that 'early interruptions' of elicited spoonerisms, of the type D..BARN DOOR, and G..BAD GAME, are more frequent in the nonword-nonword priming condition than in the word-word priming condition. This is so because, according to Levelt's lexicality criterion, in inner speech nonwords are more easily detected than real words. Such early interruptions must be reactions to inner speech, not to overt speech, because the brief fragments of the elicited errors are shorter than humanly possible reaction times (see Levelt 1989: 473, 474 for an alternative argument). Not only the interruptions but also the following repairs, often with an offset-to-repair interval of 0 ms, presumably result from self-monitoring of inner speech (Blackmer and Mitton 1991, Nooteboom 2005b). Nooteboom (2005b) found that indeed 'early interruptions' are more frequent in the nonword-nonword than in the word-word priming condition. This finding suggests that 'early interruptions' should not be collapsed with full or partial (but not interrupted) exchanges, as has been done in many published reports.

As mentioned before, it has been standard practice among those employing the SLIP technique for eliciting spoonerisms in pairs of monosyllable words, to define as 'completed spoonerisms' or 'full exchanges' all those speech errors in

which the initial consonants of the words are exchanged, irrespective of what happens to the remainders of the two words. It has also been common practice to remove all speech errors from the data that can be interpreted as intrusions from earlier parts of the experiment. Now consider that one is studying lexical bias in elicited spoonerisms. A possible test stimulus word pair in the nonword-nonword outcome condition is BAD GAME, provoking the nonword-nonword error GAD BAME, but as it happens potentially also GAD BA..., where one does not know how the second error word would end (or ended in inner speech), or even GAS BAME, where one part of the pair is an intruding lexical item, or GAS BAIT, where both error words are real words. Those errors can hardly be taken for nonword-nonword spoonerisms.

The real question is: Why would a primed-for nonlexical spoonerism like GAD BAME turn into a lexical or partly lexical spoonerism like GAS BAME or GAS BAIT, or GAS BA..? It turns out that, under certain conditions, this happens more often in the nonword-nonword than in the word-word priming condition (Nooteboom 2005c). This seems to be an effect of self-monitoring. In inner speech nonlexical spoonerisms would be detected and replaced more often than lexical spoonerisms, being replaced by either the correct targets (this would not become observable in the error data), or by other, but now lexical, errors. A possible rich source of these secondary, lexical, errors would be the lexical items that were recently encountered in the same experiment, and thus would be still relatively

active. These secondary errors provide valuable information on a possible strategy in self-monitoring, in which nonlexical errors in inner speech are replaced by lexical ones before speech is initiated. It should also be observed that an operation where one error (GAD BAME) is made in inner speech, which is then (partly) replaced by another error like GAS BAME or GAS BAIT, before pronunciation is started, should be time-consuming. Nooteboom (2005c) found that response times for the GAS BAIT cases are some 100 ms longer than those for the GAD BAME cases. This supports the hypothesis that errors like GAS BAIT for BAD GAME are indeed secondary errors, replacing the 'elicited spoonerisms', and made after the elicited speech error in inner speech has been rejected.

These reflections on error types lead to the following view of what may happen to 'elicited spoonerisms' in inner speech as elicited in a SLIP task. The basic assumption of this view is that all speech errors made in a SLIP task that start with the initial consonant of the second word have their origin as the 'elicited spoonerisms' in inner speech. If the speech errors deviate from the predicted spoonerisms this is caused by some operation of the self-monitoring system. Of course, the 'elicited spoonerisms' may go undetected, and then are articulated. This gives a category of predicted 'completed spoonerisms' of the type BARN DOOR > DARN BORE or BAD GAME > GAD BAME. Secondly, the 'elicited spoonerisms' may be detected in inner speech, but overt speech will be initiated too hastily, and then interrupted early, giving rise to a category of 'early interruptions' of the type

BAD GAME > G..BAD GAME. Thirdly, the 'elicited spoonerisms' will be detected in inner speech and then replaced by another speech error, also starting with the initial consonant of the second word, generating a category of so-called 'replacement errors' of the type BAD GAME > GAS BAME or > GAS BAIT. Fourthly, the 'elicited spoonerisms' will be covertly repaired before speech initiation: BAD GAME > GAD BAME > BAD GAME. Unfortunately these 'covert repairs' disappear from the error counts. Errors not beginning with the initial consonant of the second word are discarded.

Let us assume for a moment that there are no covert repairs. We will come back to these below. Under this assumption the current view of what may happen to 'elicited spoonerisms' in inner speech has an interesting property. The sum of the numbers of 'completed spoonerisms' (GAD BAME), 'early interrupted spoonerisms' (G..BAD GAME), and 'replacement errors' (GAS BAIT) equals the number of predicted spoonerisms in inner speech before self-monitoring operates. This is important, because the feedback account predicts that, before self-monitoring operates, there are more word-word than nonword-nonword spoonerisms (Dell 1986), but if there is no effect of feedback on lexical bias, this sum would be the same for the word-word and the nonword priming condition. Given this new view of self-monitoring inner speech, let us now see what predictions we can derive from each of the three accounts of the origin of lexical bias, keeping in mind that 'covert self-repairs' may cause elicited spoonerisms to become unobservable.

The first possibility is that immediate feedback alone is responsible for lexical bias. This would not mean that there is no self-monitoring, but rather that self-monitoring would not employ a lexicality criterion and therefore would not cause a bias towards word-word spoonerisms. This is the position taken by Stemberger (1985) and Dell (1986). From this position together with our new view of self-monitoring we derive the following predictions:

1) The sum of 'completed spoonerisms', 'interrupted spoonerisms', and 'replacement errors' is larger for the word-word than for the nonword-nonword priming conditions, corresponding to the lexical bias effect caused by feedback before self-monitoring operates.

2) For each error type separately the number of errors is also larger for the word-word than for the nonword-nonword priming condition. Note that only in this 'feedback only' account the lexical bias effect is equally strong in all error types.

A second possibility is that self-monitoring 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). We will assume that the lexicality criterion equally affects all three response types following detection of a speech error. From this the following predictions can be derived:

3) The sum of 'completed spoonerisms', 'interrupted spoonerisms', and 'replacement errors' is the same or only slightly larger for the word-word than for the nonword-nonword priming conditions. Before self-monitoring operates, the number of 'elicited

spoonerisms' would be equal. However, there may or may not be more 'covert repairs' in the nonword-nonword than in the word-word priming condition, but if there are, this effect would be notably smaller than the lexical bias effect in the 'completed spoonerisms'. This is so, because of prediction 4):

4) As a result of the lexicality criterion in self-monitoring inner speech there should be fewer 'interruptions' plus 'replacement errors', and as a consequence more 'completed spoonerisms', in the word-word than in the nonword-nonword priming condition. Note that if this indeed were so, even when the sums of all error types were equal for the two priming conditions, this would mean that both feedback and covert repairs contribute little or nothing to lexical bias.

Finally, as suggested by Hartsuiker et al. (2005), it is possible that lexical bias is caused by both feedback and self-monitoring. This leads to the following predictions, that are basically the same as predictions 1 and 4:

1') The sum of all errors is larger in the word-word than in the nonword-nonword priming condition.

4') There are, relative to the sum of all errors, fewer 'interruptions' plus 'replacement errors' and more 'completed spoonerisms' in the word-word than in the nonword-nonword priming condition.

Although a significant difference in the sum of all errors combined with a significant difference in the distribution of 'interruptions' and 'replacement errors' in the predicted direction would be consistent with Hartsuiker's hypothesis, such a

finding should be interpreted with care. A difference in the sum of all errors between priming conditions could be attributed either to feedback or to the unobservable 'covert repairs' being more frequent in the nonword-nonword than in the word-word priming condition, as a result of the lexicality criterion employed in self-monitoring.

5. A re-analysis of some earlier data

We will briefly apply the above reasoning to data obtained in an experiment reported by Nooteboom (2005b) which gives more details on the experiment. The observed numbers and percentages of 'completed spoonerisms', 'early interruptions', and 'replacement errors' are given in Table I, for the word-word and nonword-nonword priming conditions separately.

_____

Table I about here

_____

The data in Table I shows a strong and significant lexical bias in the 'completed spoonerisms' (binomial test, p=.012). The data also shows a mirror image of the lexical bias effect: There are fewer interrupted and replaced spoonerisms in the word-word than in the nonword-nonword priming condition, as predicted by a self-monitoring account of lexical bias. The distributions of error types are significantly different for the two priming conditions ($chi^2(2)=12.44$; p=.002). Importantly, the sum of error types does not differ significantly between priming conditions (binomial

test, p=.173), suggesting that there is no contribution of immediate feedback or of the number of 'covert repairs' to lexical bias. These results are not in line with the predictions of the 'feedback only' account, and lend support to the 'self-monitoring only' explanation of lexical bias. Lexical bias in this experiment can be completely accounted for by self-monitoring interrupting and replacing nonword-nonword spoonerisms more frequently than word-word spoonerisms.

The frequency-based chi^2 analysis above, however, assumes that all cell values are independent, which can not be guaranteed here. Moreover, it is quite likely that individual participants and items vary in their propensity to produce speech errors. These sources of random variation should not be ignored, because they may result in capitalization on chance (Quené and Van den Bergh 2004). To remedy these problems, the same data were also analyzed by means of multi-nomial logistic regression[1)] (Hosmer and Lemeshow 2000; Pampel 2000). The proportion P of responses within each error category is converted to logit units (i.e., to the logarithm of the odds of P, or logit(P) =log(P/(1-P))). These logit units, in the model functioning as regression coefficients, can here be seen best as estimated cell means with all participants and all items kept apart. Negative values indicate proportions lower than 0.5. Lower values stand for lower cell means. There are three response variables, corresponding to the three error categories in Table I. These three variables are then regressed simultaneously on the lexicality factor. For the resulting logit values see Table I. This analysis was chosen, because logistic regression is perfectly

suitable for regression analysis with one or more dichotomous dependent variable(s). The multi-nomial logistic regression was done by means of a mixed-effects model, with participants and items as additional random factors (Goldstein 1995; Snijders and Bosker 1999; Luke 2004; Quené and Van den Bergh 2004). This type of model is similar to a repeated-measures, within-participant, analysis of variance.

This analysis confirms the effects summarized above. First, the lexical bias yields a significant contrast between word-word and nonword-nonword priming conditions in the 'completed spoonerisms' ($F(1,36)=6.83$, $p=.013$)[2)]. Second, this bias is reversed in the 'interrupted' and 'replaced' spoonerisms, yielding a significant interaction effect between the two priming conditions and the three error categories ($F(2,36)=9.49$, $p=.004$). Thus the effects observed in the raw counts aggregated over participants and over items (Table I) hold out under more advanced statistical analysis, in which the random variation over response categories, between participants and between items, is taken into account. This multi-level multi-nomial regression overcomes the methodological problems discussed in section 3 above (in chi^2 analyses or ANOVA). The analysis suggests that the results cannot be ascribed to random variation between participants or between items, although some uncertainty remains because of the scarcity of observations per participant. In addition, the power of the analysis is not very good because of the limited number of participants.

Although this re-analysis raises hopes about the suitability of this method for SLIP data, and although its results are certainly suggestive, our current interpretation awaits confirmation from further experiments with many more participants. Analyses of further data sets are underway at the time of writing.

6. Conclusions

The SLIP task for eliciting spoonerisms has inherent shortcomings as an experimental technique, stemming from its low yield in terms of spoonerisms per participant, combined with variability of both participants and items in their propensity to spoonerize. This may easily lead to spuriously significant effects. This inefficiency of the technique has also led to a standard practice in coding and analyzing the data that may obscure important strategies of participants in this task. Notably it has become customary to collapse error types that are better kept apart. A new view of self-monitoring inner speech is presented, which makes different predictions for ‘completed spoonerisms’ on the one hand, and for both ‘early interruptions’ and other errors beginning with the initial consonant of the second word on the other. This view is briefly tested here on data obtained in an earlier experiment, using two analysis methods. The results suggest that lexical bias in phonological speech errors is caused by ‘elicited spoonerisms’ in inner speech being more often interrupted after speech initiation, or replaced by other speech errors before speech initiation, in the nonword-nonword than in the word-word priming

condition. The results also suggest that there is no contribution of immediate feedback and no contribution of 'covert repairs' to lexical bias.

Footnotes:

[1)] We are aware that many readers may be unfamiliar with this type of analysis, which is unfortunately computationally complex. We refer to Quené and Van den Berg (2004) for a readable introduction.

[2)] In logistic regression analysis it is customary to give diagnostic chi^2 values where in ANOVA F ratios are given. In order to avoid confusion with the earlier chi^2 analysis in this chapter, we have recalculated the chi^2 values of the logistic regression analysis to the more familiar F ratios.

ACKNOWLEDGEMENT

We are grateful to Huub Van den Bergh for his assistance in running the logistic regression analyses.

[Word count, including footnotes and Table I, but not including title page, 50 words abstract, 5 key words, and references: 5533].

REFERENCES

Baars, B.J. (1980). ‘Eliciting predictable speech errors in the laboratory’, in V. A. Fromkin (ed.), *Errors in linguistic performance*. New York: Academic Press, 307-318.

Baars, B.J., and 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., and 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.

Blackmer E.R., and Mitton J.L., 1991. ‘Theories of monitoring and the timing of repairs in spontaneous speech’, *Cognition* 39: 173-194.

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: 313-349.

Dell, G.S. and Warker, J.A. (2004). ‘The tongue slips into (recently) learned patterns’, in H. Quené & V. Van Heuven (eds.), *On Speech and Language; Studies for Sieb Nooteboom.* Netherlands Graduate School of Linguistics: Occasional Series, 47-56.

Del Viso, S., Igoa, J.M., and 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.

Devore, J and Peck, R. (2005). *Statistics: The exploration and analysis of data* (5th ed.). Belmont, CA: Brooks/Cole.

Garrett, M. F. (1976). 'Syntactic process in sentence production', in R.J. Walker, Walker and E.C.T. Walker (eds.), *New approaches to language mechanisms*. Amsterdam: North-Holland Publishing Company, 231-256.

Goldstein, H. (1995). *Multi-level Statistical Models*. New York: Halstead Press.

Hartsuiker, R., Corley, M., and 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., and Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley.

Levelt, W.J.M. (1989) *Speaking. From intention to articulation*. Cambridge, MA: The MIT Press.

Levelt, W.J.M., Roelofs, A., and Meyer, A.S. (1999). 'A theory of lexical access in speech production', *Behavioral and Brain Science*s 22: 1-75.

Luke, D. A. (2004). *Multi-level Modeling*. Thousand Oaks, CA: Sage. Series: Quantitative Applications in the Social Sciences, Vol. 143.

Motley. M.T. and Baars, B.J. (1975). ‘Encoding sensitivities to phonological markedness and transitional probabilities’, *Human Communication Research* 2: 351-361.

Motley, M.T. and Baars, B.J. (1976). ‘Semantic bias effects of verbal slips’, Cognition 4: 177-187.

Nooteboom, S.G. (2005a). ‘Listening to one-self: Monitoring speech production’, in R. Hartsuiker, Y. Bastiaanse, A. Postma, and F, Wijnen (eds.), *Phonological Encoding and Monitoring in Normal and Pathological Speech*. Hove: Psychology Press, 167-186.

Nooteboom, S.G. (2005b). ‘Lexical bias revisited: Detecting, rejecting, and repairing speech errors in inner speech’, *Speech Communication* 47 (issue: 1-2): 43-58.

Nooteboom, S.G. (2005c). ‘Lexical bias re-revisited. Secondary speech errors as a major source of lexical bias’, in J. Véronis and E. Campione (eds.), *Disfluency in Spontaneous Speech* (An ISCA Tutorial and Research Workshop). Équipe DELIC, Université de Provence, 139-144.

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

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

Snijders, T. and Bosker, R. (1999). *Multi-level Analysis: An introduction to basic and advanced multi-level modeling*. London: Sage.

Stemberger, J.P. (1985). ‘An interactive activation model of language production’, in A.W. Ellis (ed.), *Progress in the psychology of language*, Vol 1, London: Erlbaum, 143-186.

Table I.

Numbers of completed, interrupted, and replaced spoonerisms separately for the word-word and nonword-nonword priming conditions as observed in a SLIP task with N=1800 test stimuli (Nooteboom 2005b). Regression coefficients from the mixed-effects, multi-nomial logistic regression (in logit units, see text) are given in parentheses.

	word-word	nonword-nonword	total
completed	39 (-3.10)	19 (-3.84)	58
interrupted	27 (-3.48)	45 (-2.94)	72
replaced	22 (-3.69)	30 (-3.37)	52
total	88	94	182