Published in final edited form as:

J Mem Lang. 2009 February 1; 60(2): 291. doi:10.1016/j.jml.2008.09.006.

More on Lexical Bias: How Efficient Can a “Lexical Editor” Be?

Nazbanou Nozari and Gary S. Dell

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

Abstract

The lexical bias effect (the tendency for phonological speech errors to create words more often than nonwords) has been debated for over 30 years. One account attributes the effect to a lexical editor, a strategic component of the production system that examines each planned phonological string, and suppresses it if it is a nonword. The alternative explanation is that the effect occurs automatically as a result of phonological-lexical feedback. Using a new paradigm, we explicitly asked participants to do lexical editing on their planned speech and compared performance on this inner lexical decision task to results obtained from the standard lexical decision task in three subsequent experiments. Our experimentally created “lexical editor” needed 300 ms to recognize and suppress nonwords, as determined by comparing reaction times when editing was and was not required. Therefore, we concluded that even though strategic lexical editing can be done, any such editing that occurs in daily speech occurs sporadically, if at all.

Keywords

lexical bias; self-monitoring; speech errors; feedback

INTRODUCTION

The lexical bias effect

More than three decades have elapsed since the *lexical bias effect* was first shown in speech errors, but after all these years it is still in the spotlight for the crucial role it plays in molding language production models. By definition, the lexical bias effect is the tendency for phonological speech errors to result in real words rather than nonwords at a higher rate than chance would predict (e.g., Baars, Motley & MacKay, 1975; Dell & Reich, 1981; Hartsuiker, Corley & Martensen, 2005; Humphreys, 2002; Nooteboom, 2005). For example, lexical bias should increase the chance of *BARN DOOR* slipping to *DARN BORE* relative to that of *BARN PORCH* slipping to *PARN BORCH*.

The initial interest in the lexical bias effect was mostly to determine its existence and robustness. Two groups of studies looked into this issue. Experimental studies all used the SLIP (Spoonerism of Laboratory Induced Predisposition) procedure, first developed by Baars et al. (1975). This method creates initial-consonant exchanges during the production of word pairs. In the SLIP task participants would see word pairs like *COULD GORE* on the screen following each other with short inter-trial intervals. Randomly a signal prompts the participants

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

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

to speak the last word pair they saw on the screen. Word-outcome target pairs (e.g., *DEEP COT*) could potentially slip into words (i.e., *KEEP DOT*) while nonword-outcome targets (e.g., *DEED COP*) would create nonwords if slipped (i.e., *KEED DOP*). The lexical bias effect is demonstrated if lexical-outcome slips exceed nonlexical-outcome slips. All the published experimental studies confirmed that this is indeed the case (Baars et al., 1975; Dell, 1986, 1990; Hartsuiker et al., 2005; Hartsuiker et al., 2006; Nooteboom & Quené, 2008; Oppenheim & Dell, 2008). The second group of studies used natural speech error corpora. The effect was initially discovered in studies of English, Dutch, and German (e.g., Berg, 1983; Dell & Reich, 1981; Stemberger 1984; Nooteboom, 2005) but not in Spanish (Del Viso, Igoa & Garcia-Alba, 1991). MacKay (1992), however, brings up the issue of statistical power for detecting lexical bias in the Spanish study because of the relative scarcity of errors in that study (182 errors as opposed to, e.g., 363 errors in Dell & Reich, 1981). Moreover, the existence of lexical bias in Spanish was recently confirmed by both analysis of the natural error data and experimental use of the SLIP procedure (Hartsuiker et al., 2006). Overall, the existence of the lexical bias effect, as a robust effect in speech errors, is no longer disputed. But as this debate is put to rest a new chapter opens: how is the effect created?

The lexical bias effect as a key concept in language production theories

Any language production model has to be able to explain, among other things, the error patterns in speech. Following the principle of parsimony, it is to a model's advantage if it provides a natural mechanism to account for such patterns rather than appealing to post-hoc explanations. The lexical bias effect happens to be one of the most intriguing speech-error effects in this respect, one that according to Hartsuiker (2006) has "divided the language production literature for decades."

Language production models, for the most part, fall under one of two categories with regard to their account of the underlying mechanism that creates lexical bias. The first group, who see production as a sequence of independent and serially ordered stages (e.g., Garrett, 1975; Laver, 1980; Levelt, Roelofs, & Meyer, 1999; Roelofs, 2004), look at lexical bias not as an effect arising during the speech formulation phase but as one resulting from monitoring the output of the formulation phase for its lexical status. That is, speech that is formulated, but not yet spoken, is assessed for whether it forms words or nonwords. This prearticulatory monitoring for lexical status -sometimes called "lexical editing"-was the first proposed mechanism for lexical bias (Baars et al., 1975) and this explanation for the effect continues to be influential (e.g., see Nooteboom & Quené, 2008, for a review). The second group, who believe in interactive rather than independent production stages, identify lexical bias as an automatic property of the production system, stemming from feedback between phonemic and lexical levels. In other words, they see the effect arising during speech formulation instead of in a post-formulation monitoring stage (e.g., Dell, 1986). More recently a mixed account of the lexical bias effect was proposed combining both interactive feedback and monitoring (Hartsuiker et al., 2005). Figure 1 summarizes these three accounts by presenting hypothetical error rates for word and nonword outcomes and attributing their differences to influences occurring during either formulation or editing. Notice that the assumed number of observed errors is the same in all cases and the differences lie in the number of formulated errors. In the interactive feedback account, there is no need for a monitoring process to explain the effect and thus what is formulated is what is observed. The influence of editing is maximal in the pure monitoring account, because all the difference in the number of lexical and nonlexical observed errors is ascribed to the monitoring process rather than to formulation.

In the research reported in this paper, we create an experimental analogue of a monitor that performs lexical editing, in order to evaluate how easy it is to make speech production contingent on lexical status. If this can be done quickly and accurately, then monitoring

explanations can be considered plausible. If not, then alternative, automatic accounts such as feedback gain in credibility. Specifically, in our main experiment participants are asked to produce single-syllable utterances. In one condition, they speak only if the prepared syllable is a word, that is, they do explicit lexical editing. In another, they speak only if the syllable is a nonword and, in a third condition, they produce the syllable regardless of its lexical status. By comparing response times when lexical editing (or its counterpart, “nonlexical editing”) is required to when it is not required, we can measure the difficulty associated with incorporating a lexical-status sensitive monitor into the speech production process. To prepare the ground for comparing different aspects of the above accounts, we first give an overview of the general monitoring accounts and then focus specifically on monitoring for lexical status and how this process differs from the interactive explanation of lexical bias.

Although proponents of the monitoring account of lexical bias all believe in the existence of a monitor, there are considerable differences among their views on the nature of this monitor and its properties. The phrase “self-monitoring in speech” refers to a wide variety of processes from error detection and correction to avoiding the utterance of socially inappropriate materials to deciding to restate a concept in a more informative way at the sign of listener’s confusion. The monitoring function that is relevant to the lexical bias effect would be a specific kind of monitor, a “lexical editor”, which would only be concerned with the lexical status of the potential utterances. This specificity is crucial to our purpose, as will be discussed later, but for now we will focus on reviewing monitoring of speech and the model of monitoring we created in our experiment.

Self-monitoring of speech production

Speakers appear to be able to monitor their speech both before and after it is articulated. There is enough evidence to believe that speakers have access to their “inner speech”, as well as their overt speech, and frequently intercept and repair errors through this internal channel (Blackmer & Mitton, 1991; Levelt, 1983). However, the precise nature of this inner speech has remained unclear. Some believe that all the intermediate cognitive processes in language production are subject to such monitoring while others conclude that only the final outcome of such processes is accessible to editorial operations (see Postma 2000 for a review). Moreover, there is little consensus on how this monitoring process is executed.

The most well-developed account of monitoring is the *perceptual loop monitor* (Levelt, 1983; Hartsuiker & Kolk, 2001). This theory assumes that speakers have no access to the components of the production system, but only to the final result of the process which happens to be either the inner or the overt speech (see also Bock, 1982). Monitoring takes place through parsing both potential and actual output by the comprehension system and treating it, more or less, as speech produced by someone else. Thus, the theory obviates the need for an additional monitoring component, aside from the comprehension system itself. There are two important assumptions to this theory: first, the whole monitoring process is a “central” process that involves working memory and is therefore subject to its limitations. A corollary to this assumption is that attention plays an important role in this way of monitoring. To further elucidate this point, Levelt (1989) states that self-monitoring is a controlled process that flexibly pays attention to specific characteristics of speech. Because of its controlled nature, engaging the perceptual loop monitor might be expected to complicate the speech production process in some manner. Of course, this complication is compensated for by the error correction (when the error has already been spoken) and prevention (when errors in planned speech are intercepted).

The second assumption of the perceptual loop monitor is the existence of two stages of monitoring: a conceptual monitoring phase in which, through a matching function, the message-to-be-spoken is compared to the speaker’s original intention. This stage is followed

by a criterion-based phase in which the planned speech at the phonetic or phonological level is evaluated according to a number of different criteria (syntax, lexicality, loudness, etc.) and ideally an error signal is sent to working memory if any of these criteria are not met (Hartsuiker, 2006). An indirect implication of this assumption is that a monitor only halts the erroneous utterance and sends a restarting signal to the system to perform the repair. Therefore the monitor, *per se*, has no active role in repair proper.

In attempting to implement an experimental monitor that comes as close as possible to the one described in the literature we used the properties of the perceptual loop monitor. The first assumption of the perceptual account of monitoring discussed above is that the monitor is “conscious and attentional” and “more or less ...deliberate” (Postma, 2000, pp.115). Therefore, in our experiment we explicitly *asked* people to monitor their speech before articulating it.

Secondly, we identified a criterion for the monitor. We instructed our participants to monitor for lexical status. In the next section we will talk in more detail about views on lexical editing and resume the discussion with picking a criterion of lexicality that would make our experimental monitor a lexical editor.

Views on a specific kind of monitoring: “the lexical editor”

In this section we delineate how monitoring and interactive (feedback) theories explain the lexical bias effect and discuss how our experimental lexical editor can be created. In Dell’s (1986, 1988) model lexical units (morphemes) and phonemic units are joined by two-way connections. When a morpheme node exists for a potential word, it can boost the activation of its phonemes. If the node does not exist (i.e., nonwords), the phonemes do not get this boost. It is worth mentioning that the feedback account readily explains another important finding in error data, the semantic bias effect (Motley & Baars, 1976), which highlights the fact that lexical errors tend to be semantically related to previously produced targets, simply because such nodes have already been primed by the semantic features they share with the target node. Furthermore, feedback was also proposed as the mechanism for the repeated phoneme effect, the tendency for phonemes to move between words that share phonemes (Dell, 1984) and for error effects at the word level such as the tendency for semantic errors to share phonology (e.g., Harley, 1984; Martin et al., 1996)

As an alternative to the feedback account, a “lexical editor” can be integrated into the monitoring component of the perceptual-loop monitor. So, in addition to monitoring for whether planned speech is correct, socially appropriate, and fully expresses the intended meaning, one can also monitor according to the criterion “is this planned string a word?” and then halt or edit the string if it is not. The actual mechanisms of lexical editing within a monitoring model have yet to be specified, aside from the assumptions that it can operate on planned speech, that the monitoring process is attention demanding, and that it can flexibly alter its criterion. Support for this flexibility comes from demonstrations that the lexical bias effect is curtailed when speakers expect their output to be nonwords. Such flexibility was first demonstrated by Baars et al. (1975). When most of the target strings in the SLIP procedure were nonwords, slips that created nonwords were about as likely as those creating words, and these rates were about equal to those for word outcomes when most targets were words. Hence, it appears that nonword outcomes are suppressed only when words are expected—exactly what would be expected from a flexible editor that suppresses nonword outcomes when the speaker expects to say words. Hartsuiker et al. (2005), however, found that lexical bias is observed when both word and nonwords targets are common, and suggested that automatic feedback contributes to the effect because, when both words and nonwords are expected, editing for lexical status serves no function. However, they also found that when nonword targets are expected, there is a tendency to suppress *word* outcomes, an effect of a hypothesized *non-*

lexical editor. Hartsuiker et al. thus suggested that both automatic feedback and a flexible ability to monitor output for lexical status contribute to lexical bias (Figure 1c).

The purpose of this paper is to deliver quantitative assessments of the hypothetical lexical editor in speech production. How does such an editor work? So far we have introduced the general scheme of our experimentally-created monitor. Recall that perceptual loop monitors -which our experimental monitor is an example of- require a criterion of operation. So, to make our monitor a lexical editor we should pick a relevant criterion, a criterion of lexicality. The only real example that we have of an explicit and flexible process that assesses lexical status is the lexical decision task commonly used in studies of word input processing. Thus, we adopt this task as the model of the lexical editor and set the criterion to “is what I am about to say a word?” Our first experiment asks participants to do a lexical decision on their planned output, and to make their output contingent on the decision. If this inner lexical decision process can be quickly and accurately performed, then it may indeed occur in natural speech production and may explain at least some component of the lexical bias effect.

How quickly should an efficient lexical editor perform? In considering this question, it is useful to start with some actual estimates derived from data. The duration of the monitoring process for error detection (“does this string correspond to the target word?” as opposed to “is this string a word?”) has been estimated by Hartsuiker and Kolk (2001) by modeling actual error detection and repair data from speech samples. They showed that it would take no more than 150–200 ms to monitor each potential word for errors and about 200 ms is added for interception of the detected error, if any. The reasonably short duration of their estimate of monitoring time for error detection seems consistent with speech fluency. Moreover, given that error monitoring is clearly occurring in their speech data, we conclude that a monitor that works on this time scale is indeed compatible with fluency, or at least with the fluency present in their data. But can monitoring for lexical status also be carried out in this time frame?

The lexical decision literature provides a useful starting point for answering this question. Ratcliff et al. (2004) applied the diffusion model of choice responses (Ratcliff, 1978) to lexical decision data from visual stimuli to determine the duration of the “decision” component of lexical decision reaction times. This decision time, which can be assumed to correspond to the monitoring time for lexical status, was estimated to be around 165 ms for high frequency words and 225 ms for low frequency words (Ratcliff et al., 2004, experiment 9). Thus, the estimates from lexical decision during input processing are somewhat similar to Hartsuiker and Kolk’s error monitoring times, and so there is reason to expect that lexical editing on planned speech might also be associated with monitoring times under 200 msec. We will attempt to estimate the duration and accuracy of the process of lexical editing for words and nonwords in planned speech and to see if these estimates reveal a process that is as efficient as this.

Aside from our interest in evaluating the plausibility of the lexical editor in production, an investigation of people’s ability to make lexical decisions on planned speech is useful for another reason. The empirical base of cognitive psychology consists of studies of speeded (typically two-choice) responses to overt stimuli, such as the lexical decision task. Cognition, however, is more than just overt stimulus-response mappings. Many of the “stimuli” that we must categorize or otherwise deal with are provided by our own thought processes. Language production is the prime example. There is no imperative stimulus. Yet, the speaker must process a sequence of internally generated elements. We suggest that a systematic investigation of decision making to such elements can reveal useful properties of cognition. Our study looks at “lexical decisions” to generated and, in subsequent experiments, overt stimuli. By comparing decisions under these situations one can begin to discover the challenges of processing internal representations. For example, such a representation must be held in working memory in contrast to an overt stimulus that remains in view. Or, it may lack the detail that is present in

an overt stimulus. Thus, there is reason to think that processing internally generated items may be more difficult than processing those that are right in front of us.

We first present an experiment in which participants produce syllables and monitor them for their lexical status. Then in the following three experiments we compare our results with their equivalents in a visual and an auditory, as well as a go/nogo version, of the lexical decision task. In experiment 1 participants first hear an auditory syllable comprising a vowel and a consonant (e.g., /ʌn/) and immediately see a visual syllable on the computer screen (e.g., *GAP*). They are asked to combine the first letter of the visual stimulus (i.e., *G*) with the auditory syllable and construct a new syllable (outcome) in their minds (i.e., *GUN*). It is this newly constructed syllable that will be monitored for lexical status. The visual syllables that contribute to the onset of the constructed outcome syllable are words (e.g., *GAP*) half of the time and nonwords (e.g., *GACK*) half of the time (The manipulation of lexical status of the visual stimulus was simply an extra mechanism of control and should not be confused with the critical manipulation of lexical status of the outcome syllables that are monitored). The material is set up so that four combinations of stimulus/outcome exist for all the syllables: Stimulus Word/Outcome Word (SWOW), Stimulus Nonword/Outcome Word (SNOW), Stimulus Word/Outcome Nonword (SWON) and finally Stimulus Nonword/Outcome Nonword (SNON). In addition to these four material-based conditions, there are three procedural conditions. Based on random assignment to a condition, participants should either speak all the new outcomes (baseline condition), only word outcomes (the Word condition) or only nonword outcomes (the Nonword condition). By comparing the latter two conditions (the critical conditions) to the baseline, we can determine the time it takes to make speech dependent on lexical status.

Experiment 1

Method

Participants—Twenty-nine native English speaking undergraduates of the University of Illinois at Urbana-Champaign and twelve paid participants took part in the experiment. The first group received course credit while the second group received \$7 for their participation. There was no systematic difference between assigning participants to different conditions. The age range of all participants was 18–29 years.

Materials—There were 160 trials per participant in the study. Each trial was a combination of an auditory vowel-consonant (syllable) string (e.g., /ʌn/) followed by a visual monosyllabic letter string, which could be either a word or a nonword (e.g., *GAP* or *GACK*). The 160 trials were organized in 20 sets of 8. Each set contained two trials from each of the four combinations of stimulus/outcome in the experiment. The arrangement of these into sets of eight ensured that each auditory syllable participated equally in all four combinations and generally, that the materials were matched among combinations as illustrated in table 1, which presents an example set (the complete list can be found in appendix 1):

Lexical status was determined by the existence of the string's pronunciation in the Merriam-Webster online dictionary. Thus, /tʌt/ is a word regardless of whether its spelling is conceived of as “touch” or “tuch”. By this rule, there were 40 lexical outcome trials (e.g., *GUN*) and 40 non-lexical outcomes (e.g., *GUCH*) each to be produced twice, once through a word stimulus (e.g., *GUN* through *GAP*) and once through a nonword stimulus (e.g., *GUN* through *GACK*).

Procedure and data collection—Participants were tested individually in a quiet room. They were seated in front of a 17-inch monitor and were told to read the instructions of the task at their own pace. Both speed and accuracy were emphasized in the instructions. Each participant practiced 15 trials under the supervision of the experimenter, who provided feedback. In the test phase, trials followed each other with fixed inter-trial intervals. No

feedback was provided during the test phase. Each trial began with simultaneous presentation of a visual fixation target on the screen and an auditory syllable. The volume was adjusted for each participant individually during the first three practice trials. All syllables were pronounced unambiguously by the second author, a native speaker of American English. Immediately after the end of the auditory syllable presentation, the visual fixation target was replaced by the stimulus letter string which remained on the screen for 2000 ms and was then replaced by the fixation target again. Reaction times were recorded using a voice key with the starting point being the appearance of the visual stimulus. Participants were instructed to take the first letter of the visual stimulus (e.g., *G* in *GAP* or *GACK*) and put it at the beginning of the syllable they had just heard (e.g., /ʌt/ or /ʌn/) and make a new syllable (i.e., *GUCH* or *GUN*). From this point on, instructions differed for the different experimental conditions, although participants in all conditions were exposed to the same list of stimuli in randomized order. We created three such conditions: a baseline (13 participants), in which participants were instructed to speak all the new syllables, regardless of their lexical status, a Word condition (14 participants), in which participants were told to monitor their constructed syllables for their lexical status and speak them aloud only if they were meaningful words in English. In contrast with the latter condition, a Nonword condition (14 participants) was created so that participants would only speak outcomes that were nonwords. In all conditions, participants were instructed to respond as quickly and accurately as possible.

In addition to collecting response latencies with E-PRIME software (PST softwares, Pittsburgh, PA, USA), participants' responses were coded for errors once as they were performing the task (online coding) and once offline using their digitally-recorded responses. Error coding was conducted in the following manner: for the Word and Nonword conditions, aside from the mispronunciation errors, two types of errors were defined. Omission errors happened when a participant failed to speak a syllable he/she should have spoken in that condition (e.g., failing to say *GUN* in the Word condition or *GUCH* in the Nonword condition). Commission errors were coded when a syllable that should have been halted in a certain condition was spoken (e.g., saying *GUCH* in the Word condition or *GUN* in the Nonword condition). The purpose of online coding was to check why the error was generated in the first place. As implied by table 1, each outcome would be produced twice in the course of the experiment. If a participant made an error once but produced the correct form of that outcome at some other point in the experiment, the error would be counted as an error with no further checking. But if the participant made the same error twice on an outcome (e.g., failing to say *GUN* twice in the Word condition) then he/she would be asked to explain her decision right after the experiment was over. The error would only be counted as an error if the experimenter was convinced that the participant had constructed the outcome syllable correctly and was aware of its lexical status outside of the test condition, but had made a mistake in labeling the item correctly during the test. For example, if an error happened on the word *WRETCH* it would only be coded as an error if the participant confirmed that he/she had constructed the correct syllable and was able to describe the meaning of that word, but if he/she did not know the meaning that item would simply be dropped from his/her response pool. In the case of participants fluent in a second language (18% of the participants), all nonword outcomes were presented to them after the experiment to judge whether such syllables would mean anything in their second language and if so, those trials were discarded for those particular participants. 4% of responses had to be discarded for being recognized as a familiar word in another language. For the baseline condition, mispronunciation errors were thrown out of the list of reaction times. The checkings described above did not apply to this condition.

Results

Preliminaries—Two item sets were eliminated from all analyses because the auditory syllable /æθ/ was often misheard as /εθ/ due to dialectal variations. These sets were not included in the subsequent experiments.

Reaction time data—To eliminate the unfit items and participants we followed a two-step procedure. We set highly conservative criteria of exclusion to keep only those items and participants which showed successful monitoring, so we can evaluate the process of lexical editing in its ideal form. Step 1, the trial elimination procedure, was used for all three conditions in experiment 1. First, all trials with mispronunciations were discarded (0.1% for the Word condition, 0.3% for the Nonword condition and 0.6% for the baseline condition). And second, reaction times more extreme than the mean $\pm 2SD$ of the reaction time distribution for each participant were thrown away. By this method 3% of responses in the baseline condition, 3.6% from the Word condition and 3.8% from the Nonword condition were excluded.

Step 2 involved eliminating items and participants based on monitoring performance and hence was completed only for the Word and Nonword conditions. All commission errors were discarded from the reaction time analyses. The first part of this step consisted of making a distribution of the error percentages and eliminating those items with error rates of higher than 2SD above the mean of the distribution or 30% (in case that $>2SD$ was higher than 30%). In the second part, after discarding the unfit items from the previous part, the same was procedure done for the participants instead of items. These item and participant elimination steps were repeated sequentially until no more items or participants failed the criterion of exclusion.

After completing the above steps, 14% of items and 1 out of 14 participants from the Word condition and 11% of items and no participant from the Nonword condition were eliminated. The discarded items from these two conditions were also excluded from the baseline condition.

Mean reaction times for words and nonwords were much longer in the critical conditions than their counterparts in the baseline condition. Using participant means as the unit of analysis, the difference between the mean reaction time for words in the Word condition (995 ms, SE= 23 ms) and its baseline counterpart (702 ms, SE=24 ms) was 293 ms [$F_1(1,24) = 70.29$; 95%CI = ± 76 ms]. Using item means, the difference between the reaction times for words in the Word condition (1003 ms, SE=15 ms) and its baseline counterpart (699 ms, SE=6 ms) was 304 ms [$F_2(1,63) = 369.63$; 95%CI = ± 32 ms; min $F'(1,34) = 59.06$, $p < 0.01$]. The participant analysis returned a mean reaction time of 1143 ms (SE=31 ms) for nonwords in the Nonword condition and 724 ms (SE=28 ms) for nonwords in the baseline condition with a difference of 419 ms [$F_1(1,25) = 98.75$; 95%CI = ± 91 ms]. The item analysis returned 1144 ms (SE=12 ms) mean reaction time for nonwords in the Nonword condition and 718 ms (SE=11 ms) for the mean reaction time for nonwords in the baseline condition with a difference of 426 ms [$F_2(1,126) = 1361.78$; 95%CI = ± 41 ms; min $F'(1,29) = 92.07$, $p < 0.01$].

These differences represent the core data of this paper, monitoring times for lexical status. From this point on we take 304 ms and 426 ms to be the monitoring times for words (filtering out nonwords) and for nonwords (filtering out words) respectively. The choice of these item-based differences over those derived from the participant analyses is justified by their conceptual advantage. These values are the differences calculated by subtracting appropriate mean baselines for each item from the critical-condition item means and averaging the differences. We chose item- rather than participant means here because the Word/baseline-word and Nonword/baseline-nonword differences are within-item differences, but not within-participant differences because of the between-participant design of the Word, Nonword, and baseline conditions. Please note, though, that the item- and participant-based differences are quite similar, anyway. Although our conservative method of item/participant elimination

ensures capturing monitoring times in ideal cases, still we calculated a monitoring time of 312 ms for words and 464 ms for nonwords if only step 1 of the elimination process would be completed (of course, commission errors never entered the analysis). These results are quite compatible with our main findings and remove doubts about a significant discrepancy as the result of a highly conservative elimination process.

Another finding was that nonwords took longer than words to be produced in both the critical and baseline conditions. Between the critical conditions, nonword production was 148 ms longer than word production by the participant analysis [$F_1(1,25) = 13.29$, 95% CI = ± 83 ms] and 141 ms longer using the item analysis [$F_2(1,61) = 53.04$, 95% CI = 39 ms; min $F'(1,38) = 10.63$, $p < 0.01$]. In the baseline condition, nonwords were 22 ms longer than words using the participant analysis [$F_1(1,12) = 7.89$, 95% CI = 16 ms] and 19 ms longer using the item analysis [$F_2(1,61) = 7.28$, 95% CI = 17 ms; min $F'(1,42) = 3.79$, $p > 0.05$]. However, the difference was much larger between the Word and Nonword conditions compared to word and nonwords in the baseline as a significant interaction between condition (baseline vs. critical) and lexical status (word vs. nonword) implies [$F_1(1,49) = 5.07$, $p < 0.05$; $F_2(1,268) = 43.74$, $p < 0.01$; min $F'(1,61) = 4.54$, $p < 0.05$].

Finally, we found a small interaction between the lexical status of the stimulus (e.g., *GAP* vs. *GAP* in table 1) and that of the outcome (e.g., *GUN* vs. *GUCH* in table 1) in the critical conditions [$F_1(1,25) = 23.64$, $p < 0.01$; $F_2(1,58) = 8.61$, $p < 0.01$; min $F'(1,83) = 6.31$, $p < 0.05$]. Figure 2 shows the nature of this interaction. As indicated by the figure, nonwords were generated faster in response to word stimuli, while it was the opposite for words. This is in line with findings of Gordon and Baum (1994) that a lexical prime facilitates recognition of a nonword but not a word target. Note that this interaction does not affect our conclusions about monitoring times, that is, comparisons between the Word and Nonword conditions, and their baselines.

Error data—As mentioned before, after removing doubts about irrelevant sources of errors (unfamiliarity with the meaning of a certain word, having mistaken a nonsense syllable in English for a meaningful one in other languages, etc.) we divided errors into two groups: commissions and omissions. Recall that commissions consist of utterances that should have been halted according to a lexical criterion, but were not, while omissions were trials in which the syllables were incorrectly not spoken. The results of the error analysis are presented in table 2. Commission errors are of particular interest, since these show where the monitoring attempt has failed. These were far more likely in the Nonword condition (16%) than in the Word condition (3%), [$t(26) = 8.19$, $p < 0.01$]. Omission errors showed the opposite pattern being more likely in the word condition [$t(26) = 3.62$, $p < 0.01$].

Discussion

Our lexical editor took a long time to edit out both nonwords (304 ms) and words (426 ms). Let us consider how we determined these values. A schema of our experimentally-created monitoring process is shown in detail in figure 3. Processing the auditory syllable is not included in this scheme, since we are only presenting what is actually being measured in the experiment and recording response latencies begins as soon as the visual stimulus appears on the screen. In order to complete the task participants needed to scan the visual stimulus for its first letter (visual processing). The next step was to combine this letter with the just-heard syllable and to make a new syllable (assembling). For the participants in the Word and Nonword conditions, the next two steps (lexical decision and go/nogo) constitute monitoring; these are skipped in the baseline condition. The final step consists of articulating the assembled syllable if appropriate (responding). So, with the baseline reaction times we are capturing three of the

five steps displayed in figure 3. On the other hand, participants in the other two conditions had to complete the monitoring phase as well.

Therefore, all five steps shown above are captured in the response latencies for the Word and Nonword conditions. Since what is of interest to us is the cost of monitoring, we can compute this by simply subtracting mean response times for words and nonwords in the baseline from their counterparts in the Word and Nonword conditions. This subtraction gives us the 304 ms difference for words and the 426 ms difference for nonwords which represent the extra time needed if monitoring for words and nonwords, respectively, is to be included in speech production. In other words these are estimates of the cost of adding a lexical editor to the process of production (see Figure 4).

The cost of the editor that is derived by this subtraction procedure should not necessarily be solely identified with the summed durations of the hypothesized lexical decision and go/nogo steps. The total editorial cost could also include any delays in transitions to and from the monitoring process that are incurred because of the inclusion of the editorial steps. Inserting any extra process into the middle of another process might be expected to slow the original process down near the insertion point. Note that this “dual-task” cost should apply to any non-automatic lexical editing that occurs in natural speech production as well as in our experiment. As we are largely interested in determining the total cost of editing out nonwords (and words), and not in estimating the precise durations of single stages, determining this cost by subtraction is appropriate for our purpose.

In experiment 1 we investigated a model of lexical editing which resembles a lexical decision task in the speaker’s mind, with its outcome guiding the process of speech production. In the following section, we will present three actual lexical decision studies in order to explore two of the key findings in experiment 1: the very long monitoring times and the relative difficulty of filtering out words compared to nonwords. In these studies, the part in which the speaker has to mentally construct the target syllable is eliminated. The participant is just given the outcome syllables used in experiment 1 and performs a lexical decision on those syllables. The three studies vary in the nature of the stimulus and the response. We begin with a go/nogo version of the lexical decision task, which echoes the go/nogo nature of experiment 1, except that the task is not embedded in the process of speech production. Then we present two canonical (two-choice button-press) lexical decision tasks (on visual and auditory stimuli), which in comparison to the go/no lexical decision task, allow for an assessment of the decision process. In the end, we will use a mathematical model to extract the “decision times” from the canonical lexical decision studies. Recall that the previous estimates of such times are well under 200 ms, at least for frequent words (e.g., Ratcliff et al., 2004). These decision times will allow us to assess the ease of our materials for lexical decision. More importantly, they will be directly comparable to what we have labeled “monitoring costs” in experiment 1,

Experiment 2: the lexical decision task

The go/nogo lexical decision task

Method

Participants: Twenty-eight participants (all between 19 and 27 years of age) participated in the study for \$5. None of them had taken part in experiment 1.

Materials: The list for all the lexical decision tasks comprised all outcomes, words and nonwords, of experiment 1, presented to the participants in random order. After removing the ambiguous sets containing the auditory syllable /æθ/, 72 items were left. (Remember that in experiment 1 each outcome was produced twice, once using a lexical stimulus and once using a nonlexical stimulus. Each outcome appeared only once in the lexical decision tasks, however.)

Procedure and data collection: Participants were divided into two groups: a Word condition (14 participants) and a Nonword condition (14 participants). The first group was instructed to push the space bar of the keyboard only if the syllable appearing on the screen was a word. Opposite directions were given to the second group. Participants were instructed to be as quick and as accurate as possible. The first 15 trials were practice trials, in which participants received feedback for their performance. In the test phase, a fixation cross appeared at the center of the screen for 500 ms, followed by the visual stimulus (the outcome syllables of experiment 1) that remained on the screen for 2000 ms. Response times were automatically recorded by the E-PRIME software from the moment that the visual stimulus appeared on the screen. The data of each participant was skimmed through for errors right after his/her performance and he/she was given a chance to explain the errors. If an error was due to any sort of semantic confusion that item was dropped from that participant's profile. Reaction times for the correct responses entered the analysis.

Results—10% of words and 12% of nonwords were eliminated in accordance with the exclusion criteria used in Experiment 1. Also, one participant in the Nonwords condition was removed due to his close-to-random performance on the test. The participant analysis returned a mean response latency of 609 ms (SE=14 ms) for words and 690 ms (SE= 25 ms) for nonwords. The 81 ms difference was significant [$F_1(1,25)=8.0$, 95%CI = ± 58 ms]. The item analysis generated a mean of 610 ms (SE=13 ms) for words and 678 ms (SE=15 ms) for nonwords, with a difference of 68 ms [$F_2(1,28)=16.97$, 95%CI = ± 43 ms; Min F' (1,45)= 5.44, $p<0.05$]. Errors were coded in the same way as in experiment 1. Our earlier definition of errors of omission and commission applies in this case as well. The error pattern was somewhat different from that of experiment 1. Table 3 shows the percentage of errors in each condition.

There is no significant difference between the number of omission errors in the two conditions [$t(26)= 0.01$, $p > 0.05$] contrary to experiment 1 in which omissions were far greater in the Word condition [$t(26)=3.62$, $p < 0.01$]. However, there were significantly more commission errors in the Nonword condition [$t(26)=2.03$; $p < 0.01$] which was also found in experiment 1.

The visual lexical decision task

Method

Participants: Five undergraduates of the University of Illinois and twenty paid participants (each receiving \$5), all native speakers of English and all between 19 and 27 years of age, participated in this study.

Materials: The materials were the same as the go/nogo version of the task.

Procedure and data collection: Participants were instructed to respond as quickly and as accurately as possible by pushing “/” or “z” buttons on the keyboard (labeled as “yes” and “no”) in answer to the question “Is this an English word?”. For half of the participants, “yes” corresponded to the dominant hand and for the other half to the non-dominant one. Demonstration of the stimuli and post-performance error checking were exactly the same as the go/nogo version.

Results—Excluding unfit items and participants resulted in the rejection of 12% of word items, 14% of nonword items and no participants. The mean response time for word and nonword items were 612 ms (SE=11 ms) and 656 ms (SE=11 ms) respectively using the participant analysis and 606 ms (SE=10 ms) and 654 ms (SE=8 ms) using the item analysis. The word-nonword difference (44 ms with participant and 48 ms with the item analysis) was significant [$F_1(1,24)=27.56$, 95%CI = ± 16 ms; $F_2(1,27)= 11.9$, 95%CI (27)= ± 28 ms; min F' (1,46) = 8.31, $p<0.05$]. The error rate was 11% for words as well as for nonwords.

The auditory lexical decision task

Method

Participants: Twenty-five native English speakers, none of whom had taken part in any of the previous experiments, participated in the study for \$5. All participants were between 20 and 27 years of age.

Materials: The same list used in the go/nogo and the visual lexical decision tasks was used here as well. Syllables were spoken by the second author and were checked by three native English speakers for their clarity.

Procedure and data collection: Participants received instructions for a lexical decision task, this time with auditory stimuli. A fixation cross was always on the screen and trials followed each other with an inter-trial interval of 2500 ms. Response time recording and error checking was done in the same fashion as the other lexical decision tasks.

Results—Since the length of the soundfile was included in the response latencies it is an important factor to check. The mean durations of lexical and nonlexical soundfile were 636 ms (SE=14 ms) and 623 ms (SE=12 ms), respectively. The difference is negligible [$t(35) = .823, p > 0.05$].

8% of nonwords, 12% of words and no participants were discarded. Mean response time for the remaining words was 887 ms (SE=11 ms), and 938 ms (SE=13 ms) using the participant analysis, with a difference of 51 ms [$F_1(1,23) = 22.47, 95\%CI = \pm 23$ ms], and 881 ms (SE=17 ms) and 935 ms (SE=11 ms) for words and nonwords respectively using the item analysis, with a difference of 54 ms [$F_2(1,30) = 5.71, 95\%CI = \pm 46$ ms; min $F'(1,44) = 4.55, p < 0.05$].

Discussion—The mean reaction times and the small word-nonword differences found in the visual lexical decision task were comparable to those in previous studies (e.g., Ratcliff et al., 2004), suggesting that there is nothing particularly odd about our materials. The reaction times in the auditory version of the task were longer than their counterparts in the visual task because of the unavailability of all the cues at the starting point of latency registration when presentation was auditory. Despite this, the difference between word and nonword reaction times was similar in the visual and auditory lexical decision tasks.

The results of the go/nogo lexical decision task suggest that its go/no-go property is not the main source of cost in experiment 1, because the reaction times in the go/nogo experiment, which used visual stimuli, were quite similar to those obtained in the visual lexical decision task. In fact, there was no significant difference between the two (participant analysis: 612 and 609 ms for words in the canonical visual lexical decision task and the go/nogo version respectively, $F_1(1,37) = 0.02$; 95%CI = ± 32 ms; item analysis: 610 and 606 ms, $F_2(1,31) = 0.025$; 95%CI = ± 33 ms; min $F'(1,68) = 0.01, p > 0.05$). Although the difference between the two experiments was greater for nonwords, it was not significant (participant analysis: 656 and 690 ms for nonwords in the visual and the go/nogo task respectively, $F_1(1,36) = 1.93$; 95%CI = ± 48 ms; item analysis: 654 and 678 ms, $F_2(1,26) = 3.43$; 95%CI = ± 26 ms; min $F'(1,61) = 1.24, p > 0.05$). Moreover, the word-nonword difference did not differ significantly between the two experiments, as determined by the lack of interaction between the lexical status of the outcome and experiment [$F_1(1,73) = 1.39$; $p > 0.05$; $F_2(1,120) = 2.27$; $p > 0.05$; min $F'(1,155) = 0.86, p > 0.05$]. The similar findings from these two experiments might come as a surprise, because it seems that the go/nogo task has an extra step. However, a go/nogo task can be viewed as a forced-choice response task with one of the two options being “no response”. For both tasks a pairing step which binds the result of lexical decision with the appropriate command—either a right or left button push or simply no response—has to be completed (see figure 6).

What is suggested by these results is that the go/nogo nature of the task does not have a huge influence on the duration of this pairing process and responding. Thus, the large word-nonword difference obtained in experiment 1 cannot be attributed to its go/nogo nature.

As for the error pattern in the go/nogo version, we replicated, in part, the findings of experiment 1 that the commission errors are much more prevalent in the Nonword condition compared to the Word condition. A unique feature of the current pattern, though, was the lack of a statistically significant difference in the number of omission errors between the two conditions. Nonetheless, the results clearly indicate that it is much more difficult to halt words than nonwords, even when instructions emphasize a different strategy. This finding seems to be a robust one that is independent of response modality; that is, it was present in both experiment 1 and the go/nogo lexical decision task.

In order to compare the absolute monitoring times in experiment 1 with those in the lexical decision tasks we need to dissect the reaction times in the lexical decision experiments so that:

$$\text{MRT} = \text{MDT} + \text{MNDT},$$

where MRT, MDT and MNDT stand for Mean Response Time, Mean Decision Time and Mean Non-decision Time respectively. Figure 5 provides a schematic representation of the steps taken in the canonical lexical decision task. The visual processing and responding boxes are labeled the same as the ones for experiment 1. This does not indicate that we suggest these are the same processes, it merely indicates the specific steps for each task based on the materials and requirements of that task. As will be seen shortly, the method we use for subtracting the differences is insensitive to the qualitative differences between these stages.

Two components of the above equation are also depicted in the figure. It should be obvious that the MNDT is the sum of visual processing and responding times and that we are interested in the MDT as a measure of the monitoring cost to be compared with experiment 1. To obtain this measure we used a mathematical model, the EZ-diffusion model (Wagenmakers et al., 2007), to calculate MNDT and extract MDT out of the MRT's. We first briefly introduce the EZ-diffusion model and its source model and then explain how our data were fit to the model.

The EZ-diffusion model: As implied by its name, the EZ-diffusion model is a simplified version of a sophisticated cognitive model for two-choice reaction time tasks developed by Ratcliff (1978), the diffusion model. The main objective of the diffusion model was to address the trade-off between speed and accuracy which has long bedeviled the interpretation of data from such tasks. The model estimates a number of parameters (see appendix 2 for an extended discussion of the model), one of which, the mean of the non-decision component of making a forced choice response (T_{er}), is equivalent to MNDT in our equation. For the reasons discussed in the appendix we have decided to use the EZ version of the diffusion model which was more suitable for our data.

Fitting the data to the EZ-diffusion model: Although the EZ-diffusion model provides a feasible method to fit the small data sets, it puts some constraints on the nature of the data because of the strong assumptions that the model makes. The model will recover parameters no matter what the properties of the data set are, however, those parameters will be less accurate if the data set violates the assumptions of the model. As a remedy, Wagenmakers et al. (2007) suggest three "EZ" checks to be performed on each data set before feeding it to the model. These check methods, along with the results of our own data when checks were applied to them, are discussed in appendix 2. The core message is that our data did not violate any of the assumptions of the EZ-diffusion model and was, therefore, appropriate for such a model.

Three input parameters, MRT (mean response time for the correct responses), VRT (variance of the correct response times) and Pc (the percentage of correct responses) were calculated for each participant in the visual and auditory lexical decision tasks and were fed to the model which, in turn, generated three parameters for each participant. Since we are only concerned with the non-decision time parameter we do not discuss the other outputs. According to the formula $MRT = MDT + MNDT$ the non-decision time for each participant (parameter T_{er} recovered by the model) was subtracted from participant's mean reaction times for words and nonwords to calculate the MDT for that participant. Averaging MDT's over participants returned a decision (monitoring) time of 143 ms for words and 187 ms for nonwords in the visual lexical decision task and 158 ms for words and 209 ms for nonwords in the auditory lexical decision task. These results indicate that the monitoring times for the two versions of the lexical decision task were reasonably close to each other by the overlap of the 95% CIs of these estimates (124–162 ms for words in the visual vs. 143–169 ms in the auditory version; and 166–209 ms for nonwords in the visual vs. 187–231 ms in the auditory version) and both were obviously very different from the reaction time cost from monitoring estimated in experiment 1 (304 ms monitoring time for words with a 95% CI of 277–341 ms; and 426 ms for nonwords with a 95% CI of 385–467 ms).

Furthermore, the difference between the monitoring times for words and nonwords shows the same pattern. For the two versions of the lexical decision task the 95% CI of this difference overlaps (44 ms difference in the visual version, 95% CI = 28–61 ms; 51 ms difference in the auditory version, 95% CI = 31–76 ms). However, these two do not overlap with the boundaries of the difference in experiment 1 (122 ms difference, 95% CI = 84–159 ms).

Discussion of the model's results: By using the subtraction from the baseline in experiment 1 and simulations with the EZ diffusion model, we extracted the costs of monitoring for lexical status in production and in lexical decision tasks respectively. There is a dramatic difference between the monitoring costs for words in experiment 1 (304 ms) and those of the visual and auditory lexical decision tasks (143 ms and 158 ms respectively). The same holds true for nonwords (426 ms in experiment 1 vs. 187 ms and 209 ms in the canonical lexical decision tasks). We present two possible accounts for such differences, either of which, or both, can contribute to the longer monitoring times in experiment 1 compared to the lexical decision tasks. We call the first account the “impoverished inner speech hypothesis” and the second one the “interruption hypothesis”.

Impoverished inner speech hypothesis: Our first account has to do with the nature of the material being monitored. The difference between experiment 1 and the two lexical decision experiments is a difference between monitoring *inner speech* and monitoring an available stimulus. Although a visual and an auditory stimulus are present in experiment 1, neither of those, but rather a new syllable that is assembled in participant's mind, is being monitored. This is in contrast with the lexical decision tasks, in which monitoring should be performed on a syllable that is made available externally. Furthermore, in experiment 1, a decision to articulate a syllable is contingent on the result of the monitoring process. Therefore, the syllable should remain in the inner speech, until the monitor signals the system to start the process of articulation. According to the impoverished inner speech hypothesis, it is harder to monitor inner speech compared to an available stimulus simply because inner-speech representations are abstract and lack surface details. This is in line with some empirical evidence on the nature of inner speech. Using subjectively reported error data on monitored inner speech, Oppenheim and Dell (2008) hypothesized that inner speech lacks some phonetic features. An important corollary to this conclusion is the claim that inner-speech monitoring may take place at the level of phonological, rather than a phonetic, representation (Wheeldon & Levelt, 1995). With a directly available stimulus, more features (visual or auditory) are present and these cues

should speed the decision. To put it in perspective with the diffusion model, such available cues will increase the drift rate.

Interruption hypothesis: Our second account is focused on the actual procedure of transmission from monitoring to responding phase, and is a specific instance of the cost associated with inserting one process within another.

Figure 6 shows that after a decision is made on the lexical status of a letter string it has to be paired with the correct command of execution (e.g., if “word” then press “/” or if “word” then “do nothing”). For the lexical decision task, this decision/response-pairing phase is immediately followed by response execution, while for the speech production task a disrupted process has to be resumed. In the latter, while monitoring is taking place, the phonological representation of the to-be-uttered word has to be kept in working memory and when monitoring is complete the phonological representation needs to be retrieved in order to build a phonetic plan for articulation. Thus, making a lexical decision on planned speech, where the decision controls subsequent speech, is particularly difficult in comparison to the lexical decision task.

Our consideration of the impoverished inner-speech hypothesis and the interruption hypothesis recalls a secondary goal of our study, to compare the information processing of overt and internally generated stimuli. Both of these hypotheses can be offered as general accounts of the costs of decisions based on internal elements. The elements being processed may be incompletely specified because of their internal nature (impoverished inner speech), and/or there may be processing costs associated with the need to hold and manipulate internal stimuli in working memory, and deal with other concurrent tasks (interruption hypothesis). In our case, the experiments do not definitively point to one, the other or both. At this point, our goal is to raise the issue, specifically in the case of lexical decisions to planned speech, and generally as well.

Comparing monitoring times for words and nonwords: The second issue concerns the differences in response latencies for words and nonwords in our experiments. The difference in the reaction times to words and nonwords is a robust finding in the visual lexical decision literature and our canonical lexical decision tasks replicated this difference. In testing the assumptions of the EZ-model we showed that it is plausible to assume that the non-decision time for words and nonwords is the same. Therefore, the difference in the overall reaction times in these experiments can be reliably attributed to the difference in the monitoring times for words and nonwords. This 44–51ms difference found in the canonical lexical decision tasks, however, is in sharp contrast to the 122 ms difference of experiment 1. The most interesting possible explanation of this difference is that the cognitive system has a fundamental bias toward letting words rather than nonwords to be spoken. Consequently, *when the basic task is production rather than button pushing*, the word-nonword difference is exaggerated.

GENERAL DISCUSSION

Although this paper is based upon the classic finding of lexical bias, it is unique in focusing its investigation on the properties of a possible explanation for this effect, that is, a flexible, attention-demanding system that monitors planned output for lexical status. To determine the nature of the hypothetical lexical editor we addressed two questions: first, if there is a lexical editor, is it fast enough to monitor every word in our daily speech? And second, how flexible is this lexical editor? That is, can it easily be configured to edit out words instead of nonwords? In an attempt to answer these questions, we used a production paradigm to actively engage participants in monitoring their planned speech for its lexical status with “go if a word” in one condition and “go if a nonword” in the other condition. We also added a baseline condition to

estimate the time needed to complete the task without monitoring demands. By subtracting the reaction times for items in the baseline from their counterparts in the critical conditions we estimated the absolute monitoring costs for words and nonwords in our experimental paradigm. The long duration of monitoring (around 304 ms for nonwords and 426 ms for words) does not support the quickness of the process. Notice that we only used mono-syllabic words and nonwords in this study, so it would be logical to assume that monitoring times for longer items could be even longer than these estimates. Moreover, the word-nonword difference of 122 ms with many more commission errors in the Nonword condition compared to the Word condition (16% vs. 3%) suggests that this kind of explicit editing has a strong bias to filter out nonwords, rather than words. It is *possible* to filter out words, as proposed by Hartsuiker et al. (2005), but it is not the preferred criterion.

It is important to emphasize that we only looked at monitoring for *lexical status*. Therefore, our interpretation of this specific subtype of speech monitoring is, by no means, generalizable to other subtypes such as monitoring for social appropriateness of the utterance or monitoring whether a planned utterance actually matches the intended utterance. However, we do believe that lexical editing, if part of the general speech monitoring system, should follow some of the general rules of such system. For instance, the nature of the inner speech that is monitored should not differ from one monitoring subtype to another. Overall, although we believe in a general monitoring system *whose primary goal is to determine whether a planned or uttered string is the intended string*, the existence of such a system does not automatically support the existence of a lexical editor that explains lexical bias. Simply because the speech monitoring system performs efficiently on some functions does not imply that it is also apt for other functions, unless proven otherwise. Following this assertion, we believed it to be highly worthwhile to test specifically for the efficiency of the specific subtype of monitoring which is assumed to perform the lexical editing. To address the issue of general efficiency of lexical editing, we will discuss the absolute monitoring times obtained in our experiment 1. To consider questions concerning the flexibility of such an editor, we combine the results of error analyses with the difference between the duration of monitoring for words and nonwords.

Absolute monitoring times

We found that editing out monosyllabic nonwords imposes a cost of 304 ms to production time of the monosyllabic words that are spoken. When words are required to be removed, the mean production time of the spoken nonwords was 426 ms slower than baseline. Given the fact that we produce 2–3 words per second in fluent speech, each word takes between 300–500 ms to be produced (Levelt, Roelofs, & Meyer, 1999). If routine monitoring for lexical status is in action for every word we speak this would mean that there is little time left for the production system to execute the job of producing a word. Of course, there are some important assumptions behind this conclusion. First, is the assumption that monitoring is ongoing for *all* to-be-spoken words. One could argue that some parts of our speech, for example function words, do not undergo monitoring, saving the system some time for monitoring other words. How this cherry picking is performed, if in fact it is the case, is outside the scope of this paper. Here, we are simply arguing that performing the monitoring function on each and every word seems too time-consuming for fluent speech.

A second assumption is that the monitoring costs that we have identified do in fact subtract from our abilities to do the other components of production -conceptualizing, linguistic encoding, and articulating. This assumption is consistent with claims from Levelt (1983) that, perceptual monitoring is a “central corrective action” and that this very property of the monitor “makes it subject to the usual limitations of working memory”. He backs up his claim with empirical evidence for higher rates of error detection towards the end of phrases, taking that as evidence that error detection competes with production processes, whose demand for central

resources tends to become reduced toward the end of clauses (Levelt, 1983). There is also evidence in favor of monitoring being a strategy-driven process (Postma et al., 1990; Postma & Kolk, 1990, 1992, 1993). For example, Postma and Kolk (1990) showed that the ratio of disfluencies to speech errors increases if accuracy is emphasized, an implication of the process being strategic and attention-controlled. In summary, there is reason to expect monitoring to compete with production for resources. Thus, the large temporal cost that we claim comes with monitoring for lexical status would be expected to severely diminish performance in other components of the production system (see, e.g., Ferreira & Pashler, 2002 for analysis of which components of production may be subject to “central bottlenecks”). We emphasize that what matters for the logic of this study is that adding monitoring to production is costly. Such cost could be incurred because monitoring cannot occur at the same time as production processes, that is, they occur serially, or because a parallel engagement of production and monitoring slows down production. Thus, although our figures have conceptualized monitoring as a “stage” in the production process, we do not commit to an assumption that it cannot be carried out at the same time as production processes on other parts of the utterance. We only commit to the cost of doing so.

A third assumption concerns the possibility that monitoring may be feasible during a stage of production that is less subject to central bottlenecks, namely articulation. Levelt (1989) and Hartsuiker and Kolk (2001) suggest that words that have already been phonologically encoded may be stored in a buffer prior to their articulation. If capacity demanding aspects of production such as phonological encoding occur at a rate that is faster than the actual time it takes to articulate, there will be some “slack time” available for monitoring when the production system buffers several words. If a long phrase has completed its phonological encoding and is stored in the buffer, while articulation is catching up, the editor could examine the contents of the buffer. Although we emphasize that the time might be better spent editing for correctness than for lexical status, it is possible that enough time would build up in long phrases such that lexical editing is possible, even given our estimates of how slowly this process proceeds. However, there is reason to question whether a great deal of buffering occurs during production. First of all, Levelt (1989) examined whether such buffering was needed for the purpose of generating prosodic and phonological forms, and concluded that the system need only to “look ahead” to the next word in each utterance to determine the form of a given word. A second piece of evidence for limited look-ahead comes from the analysis of phonological speech errors. Garrett (1975) showed that sound exchanges are most common in adjacent words or words separated by only one word (typically a function word), with a sharp drop in phonological exchanges in words separated by two or more words. This finding suggests that buffering of the phonological forms must be limited to no more than one or two words at a time.

A final consideration is the accuracy of the editor. In experiment 1, in which speed and accuracy were both emphasized, our subjects in the lexical editing (Word) condition made 12% omission errors and 3% commission errors. The monitoring costs that we identified likely depend on this level of accuracy. If we were to allow lexical editing to be less accurate, say, detecting only half of the nonwords, it could be done more quickly. Thus, faster, but less accurate, lexical editing might be feasible. But again, we would question the value of inserting a cognitively demanding editorial process that misses all word errors and catches only a fraction of the nonword errors. Is it worth it?

We are indeed not the first to emphasize that editing needs to be rapid. Hartsuiker (2006) suggested that rapid monitoring does occur and pointed to the quick detection in a study by Motley, Camden, and Baars, (1982). That study measured the efficiency of monitoring for social appropriateness (taboo words) not for lexical status and the mean response delay was 800 ms (from cue to the onset of speech) or 1800 ms (from stimulus presentation to the onset of speech) which is not exactly quick for the routine speech, given the assumptions discussed

above. Leaving aside the question of whether this study's monitoring for tabooeness is quick or not, the evidence for such monitoring in the task is nonetheless strong. Thus, while the process of lexical editing seems to take so long, other forms of monitoring may be more efficient. As we noted previously, Hartsuiker and Kolk (2001) estimated the time needed for error detection in the speech to be no more than 200 ms. It is plausible that monitoring in service of error detection and interception, and error repair would be rapid and effective, given its obvious utility for communication. More importantly, though, monitoring for error implies a comparison between a *specific* intended representation (e.g., of a word) and a representation that was constructed for production. Monitoring for lexical status does not involve such a specific comparison. The speaker is just attempting to tell whether the planned string is *any* word. In this regard it is useful to note that lexical decision reaction times are around 600 ms, while word-monitoring reaction times (does this stimulus match a designated target word?) are consistently under 400 ms (e.g., Foss & Swinney, 1973). Thus, monitoring for whether something matches a specific intention is likely much easier, and definitely more useful, than monitoring for lexical status. Recently, Nooteboom and Quené (2008) confirmed this finding, by showing that most speech errors are detected based on a comparison with the intended words, rather than their lexical status. It is also plausible that monitoring for taboo status (Motley et al., 1982) or the emotional valence of planned words (Slevc & Ferriera, 2006) is relatively rapid as well. Tabooeness or emotional intensity may be detected by "trip wire" mechanisms. For example, if the activation of a word node for a taboo word exceeds a threshold, the wire is tripped and some action is required, such as halting production. Nonwords, by definition, lack specific representations, so there can be no trip wire. In fact, this may be why determining the lexical status of a string in standard lexical decision is more than 200 ms slower than determining that the string is a particular word.

Furthermore, the relatively shorter monitoring times in the lexical decision tasks in our experiments (143 and 158 ms for words and 187 and 209 ms for nonwords in visual and auditory lexical decision tasks respectively), show that we are capable of reasonably rapid lexical monitoring on available stimuli using a button pushing technique. Clearly, this procedure differs noticeably from monitoring for lexical status in our planned speech. We proposed two accounts to explain the difference between the monitoring costs observed in experiment 1 and the decision times in the lexical decision tasks. In our "impoverished inner speech hypothesis", in keeping with findings of Wheeldon and Levelt (1995), we argued that speech monitoring may take place at the phonological level because of the hypothesized impoverished nature of inner speech with regard to phonetic features (Oppenheim & Dell, 2008). The lack of some features (or lexicality clues) in inner speech, as opposed to an actual stimulus, might be responsible for longer reaction times in monitoring planned speech (experiment 1) compared to monitoring available stimuli (the canonical lexical decision tasks). As a second possibility, we explained in our "interruption hypothesis" how we think of monitoring as a disruption in the normal processing of the potential utterance, after which retrieving the phonological representation to encode into the phonetic plan would take some time. This is not an issue for the lexical decision tasks, since the monitoring in this case does not co-occur with (or interrupt) a second process like language production.

The word/nonword difference, editor flexibility, and the SLIP task

A difference of 122 ms, in favor of faster production when nonwords, as opposed to words, are to be filtered out, was observed in experiment 1. Moreover, in both experiment 1 and the go/nogo lexical decision task, there was a bias to produce words over nonwords in commission errors (experiment 1: 16% words produced in the Nonword condition vs. 3% nonwords produced in the Word condition; go/nogo lexical decision task: 12% vs. 5%, respectively). These data are relevant to the question of editor flexibility. On the one hand, we note that it is indeed possible for participants to follow instructions to produce nonwords and filter out words.

The cognitive system is flexible to this extent. However, our experimental lexical editor clearly functions faster when the criterion is “Let the words pass,” and when it erroneously lets something through, that something is more often a word. Notice that any inherent differences in the speed of articulating and constructing words and nonwords were already captured in the baseline condition, and so the 122 ms word-nonword difference concerns the monitoring process alone. This finding suggests that if a lexical editor is to be used, it clearly prefers to filter out nonwords, which is indeed the generally accepted purpose of the lexical editor.

As mentioned before, the flexibility of the lexical editor was supported by studies using the SLIP paradigm in which the proportion of word and nonword targets was varied (Baars et al., 1975; Hartsuiker et al., 2005). In these studies, the lexical bias effect was altered in response to context. Similarly, Dell (1986; 1990) showed that lexical bias was eliminated when the deadline in the SLIP task was very fast. Although Dell interpreted the dependence of lexical bias on the deadline in terms of the time it takes for activation to spread between a phonemic and a lexical level, one could also attribute these results to a flexible lexical editor. When there is no time for editing, perhaps the editor is shut down. Notice that there are two separable aspects to flexibility, the ability to turn off or turn on the editor, and the ability to alter the criterion from filtering out, say, nonwords, to filtering out words. Both aspects of flexibility are highlighted in Hartsuiker et al. (2005). Using the SLIP procedure, Hartsuiker et al. found no lexical bias effect in the dominantly nonword context as opposed to a significant one in the mixed context. They ascribed this finding to the flexibility of the editor, which uses a criterion of non-lexicality in the nonword context (filters out words) but drops lexical editing completely for the mixed context, where the lexical status of the context is not informative. Thus, the domination of the lexical errors at the formulation stage (attributed to interactive feedback) is suppressed in the nonword context but remains intact in the mixed context.

Our results are not at odds with these findings. In fact, our study resembles Hartsuiker et al.’s “mixed-context” condition, in which they did find more lexical errors. While we did not reproduce their nonword-context condition (in which no significant difference between lexical and non-lexical errors were found) we believe the conclusion would not be different. Although the lack of the lexical bias effect in the pure non-word context in Hartsuiker et al.’s experiment suggests some flexibility of the editor, the degree of flexibility is not high enough to overturn the results and produce a “non-lexical bias effect”. So, as far as the flexibility of the editor is concerned, we view our results, as well as those of Hartsuiker et al.’s, as endorsing limited flexibility if a lexical editor is to be used in speech production.

The findings of Hartsuiker et al. (2005) and those of another recent study by Nootboom and Quené (2008) represent the strongest support for the claim that at least a component of the lexical bias effect is due to a flexible lexical editor. Nootboom and Quené, again using the SLIP task, found differences in the lexical bias effects in completed and interrupted phoneme exchange errors (e.g., keep dot → deep cot, versus keep dot → dee...keep dot, respectively). The complete exchange errors showed the normal lexical bias effect, but the interrupted errors showed a reverse lexical-bias effect for stimuli in which the interacting consonants were dissimilar. Given that interrupting an error is likely the result of a monitoring process, differences in the influence of the word versus nonword outcome manipulation on interrupted and completed errors suggests that monitoring for lexical status is occurring. Nootboom and Quené presented a specific multinomial model of this process and, like Hartsuiker et al., concluded that monitoring for lexical status and automatic feedback both contribute to the lexical bias effect. It is important to note that their finding of differences in the effect of lexical status of error outcomes between interrupted and completed errors was also present in a study of natural slips (Nootboom, 2004) and hence data supportive of lexical editing is not limited to the SLIP task (although only the experimental data provide good control over materials).

Nooteboom and Quené also consider speed stress to be important in determining the nature of the editing that may occur in the SLIP task. Specifically, they argue that there is some evidence for monitoring for lexical status in the differences between completed and interrupted errors under the speed stress. Here the speed stress is assumed to *encourage* rather than discourage prearticulatory editing by forcing quick corrections. We believe that Nooteboom and Quené's interaction between lexical status and whether an error is interrupted or not may be due to the tendency for more familiar motor patterns (e.g. words) to be less interruptible. That is, saying a word is more ballistic than saying a nonword. A ballistic process is one that once launched, must proceed to completion. Consequently, completed errors will tend to be words, while interrupted ones will be more likely to be nonwords. Notice that this explanation does not assume monitoring for lexical status, only that there is some kind of monitoring and a subsequent interruption process that is more successful when the utterance is a nonword. Moreover, Nooteboom and Quené's finding that this pattern was particularly in evidence under speed stress is consistent with our explanation. Logan and Cowan (1984) show how the efficiency of stopping performance on a primary task is affected by the finishing time of the task. Given that the stop signal might be sent to the articulator as soon as an error is detected, the probability of interruption increases with the rightward shift of the response time distribution (i.e., longer response times). The shorter the response delays are (responding under time pressure), the less successful the process of inhibition will become. Given the above discussion, a word outcome and time stress both favor a ballistic process of articulation, therefore the expectation would be a relatively stronger lexical bias effect in completed rather than interrupted errors.

The possibility of lexical editing occurring in the studies of Hartsuiker et al. (2005) and Nooteboom and Quené (2008) raises the general question about the conditions under which lexical editing is still possible, given our results. Clearly, if speech rate is very slow there should be time to edit speech for lexical status. However, if there is time to edit for lexical status, we would claim that there is also time to do a more useful kind of editing, to edit for whether the planned utterance is in fact the target. As we mentioned before, studies show that monitoring a verbal stimulus for a pre-specified word target is much faster than making a lexical decision to that stimulus. If so, why edit for lexical status, when one can edit for the real thing -the target?

Concerns about our model of lexical editing

Psycholinguists have been talking about a "lexical editor" for decades, but to date nobody has proposed a clear mechanism for how such editor might operate. As the first step in clarifying the nature of the lexical editor, we created a model of lexical editing that works by making a lexical decision to mentally-constructed stimuli. We provided evidence that this model of lexical editing is probably too slow and too inefficient to be routinely used in speech production. Of course, there may be some other way that lexical editing occurs and this other way may be quite fast and efficient.

We are not claiming that our model necessarily corresponds to others' conceptions of the lexical editor, but we do claim that the paradigm used in experiment 1 was a good fit for our model. We acknowledge that this paradigm does not reflect all aspects of normal speech production. Speakers prepare word forms from meaning, not from an auditory syllable and a visual stimulus. Nonetheless, we claim that once the prepared syllable is assembled, there is no reason to believe that the production process should be any different. Indeed, many production paradigms investigating phonological encoding (e.g., SLIP procedure) do not include the retrieval of words from meaning, either implicitly or explicitly making the assumption that common phonological/phonetic/articulatory production stages are called upon when the impetus to speak is from meaning, and when it is from phonological/orthographic input. This

sharing of word-form encoding processes across tasks is supported by studies of production deficits in aphasia, both deficits in producing from meaning (e.g., picture naming) and producing by shadowing auditory input (e.g., Dell, Martin & Schwartz, 2007).

Furthermore, breaking syllables into parts is considered to be a natural step in phonological encoding (Dell, 1986; Shattuck-Hufnagel, 1987) and these units are most likely onset and rhyme constituents (Meyer, 1991). This claim has especially strong support for the word-onset constituents, since many speech errors involve these portions (Shattuck-Hufnagel, 1987). Our paradigm follows this natural division.

Although we cannot generalize the results produced by this paradigm to routine speech with certainty, we still take it to be a better choice than others like reading (orthographic input processing) paradigms (see Bock, 1996). Our data clearly indicate that a reading paradigm with a monitoring component (i.e., the go/nogo lexical decision task) is quite different from a speech production paradigm and therefore its results should not be taken as representing the actual phenomena taking place in speech production.

Conclusions: Homage to Ockham's razor

The principal of parsimony, Ockham's razor, is an important consideration for theories of language production (Levelt et al., 1999). Is the lexical editor parsimonious? It is, in two respects. First, it can be assumed to operate using the comprehension system, just as proposed in the perceptual-loop theory of monitoring, so there is no need to duplicate the mechanisms that retrieve words from speech (or potential speech) input. Second, there is no doubt that we monitor our speech and our planned speech and, hence, the general notion of monitoring what we say is well motivated. We have argued here, however, that these parsimonious aspects of lexical editing only go so far. The function of the lexical-comprehension system is not to do lexical decision; it is to recognize words. Although this comprehension machinery can support an editor that responds to lexical status, one still must *assume* the existence of a lexical-decision process that normally operates during production to explain lexical bias, and that assumption is therefore not parsimonious. In addition, the monitoring functions that seem clear and well motivated by decades of research do not concern monitoring for lexical status. Rather, they involve monitoring for correctness, tabooeness, and other aspects of social appropriateness. So again, although the idea of monitoring is well motivated, monitoring for lexical status is not. Given these considerations, we do not find the lexical editor account of lexical bias to be particularly parsimonious.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by the National Institute of Deafness and Other Communication Disorders (191100). We would like to thank Robert Hartsuiker, Sieb Nooteboom and an anonymous reviewer for their valuable comments.

References

- Baars BJ, Motley MT, MacKay DG. Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* 1975;14:382–391.
- Berg T. Monitoring via feedback in language production: Evidence from cut-offs. 1983 Unpublished manuscript.
- Blackmer ER, Mitton JL. Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition* 1991;39:173–194. [PubMed: 1841032]

- Bock JK. Language production: Methods and methodologies. *Psychonomic Bulletin & Review* 1996;3:395–421.
- Bock JK. Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review* 1982;89:1–47.
- Del Viso S, Igoa JM, Garci'a-Albea JE. Autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research* 1991;20:161–185.
- Dell GS, Reich PA. Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior* 1981;20:611–629.
- Dell GS. Representation of serial order in speech: evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory and Cognition* 1984;10(2):222–233.
- Dell GS. A spreading-activation theory of retrieval in sentence production. *Psychological Review* 1986;93:283–321. [PubMed: 3749399]
- Dell GS. The retrieval of phonological forms in production: Test of predictions from a connectionist model. *Journal of Memory and Language* 1988;27:124–142.
- Dell GS. Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes* 1990;5:313–349.
- Dell GS, Martin N, Schwartz MF. A case-series test of the interactive two-step model of lexical access: Predicting word repetition from picture naming. *Journal of Memory and Language* 2007;56:490–520.
- Ferreira V, Pashler H. Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Human Learning and Memory* 2002;28:1187–1199.
- Foss D, Swinney D. On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior* 1973;12:246–257.
- Garrett, MF. The analysis of sentence production. In: Bower, GH., editor. *The psychology of learning and motivation*. Vol. 9. New York: Academic Press; 1975.
- Gordon JK, Baum SR. Rhyme priming in aphasia: The role of phonology in lexical access. *Brain and Language* 1994;47:661–683. [PubMed: 7859058]
- Harley TA. A critique of top-down independent levels models of speech production: Evidence from non-plan-internal speech errors. *Cognitive Science* 1984;8:191–219.
- Hartsuiker RJ. Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes* 2006;21:856–891.
- Hartsuiker RJ, Anton-Mendez I, Roelstraete B, Costa A. Spoonish Spanerisms: A lexical bias effect in Spanish. *Journal of Experimental Psychology: Learning, Memory, & Cognition* 2006;32:949–953.
- Hartsuiker RJ, Corley M, Martensen H. The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related reply to Baars, Motley, and MacKay (1975). *Journal of Memory and Language* 2005;52:58–70.
- Hartsuiker RJ, Kolk HHJ. A computational test of the perceptual loop theory. *Cognitive Psychology* 2001;24:113–157. [PubMed: 11259106]
- Humphreys, KR. Lexical bias in speech errors. Unpublished doctoral dissertation. University of Illinois; Urbana-Champaign: 2002.
- Laver, JDM. Monitoring systems in the neurolinguistic control of speech production. In: Fromkin, VA., editor. *Errors in Linguistic Performance*. New York: Academic Press; 1980.
- Levelt WJM. Monitoring and self-repair in speech. *Cognition* 1983;14:41–104. [PubMed: 6685011]
- Levelt, WJM. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press; 1989.
- Levelt WJM, Roelofs A, Meyer AS. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 1999;22:1–75. [PubMed: 11301520]
- Logan GD, Cowan WB. On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review* 1984;91:295–327.
- MacKay DG. Problems with Popper: The initial goal is to develop viable theories, not disconfirm them. *Consciousness and Cognition* 1992;1:231–240.
- Martin N, Gagnon DA, Schwartz MF, Dell GS, Saffran EM. Phonological facilitation of semantic errors in normal and aphasic speakers. *Language and Cognitive Processes* 1996;11:257–282.

- Meyer AS. The time course of phonological encoding in language production: phonological encoding inside a syllable. *Journal of Memory and Language* 1991;30:69–89.
- Motley MT, Baars BJ. Semantic bias effects on the outcomes of verbal slips. *Cognition* 1976;4:177–187.
- Motley MT, Camden CT, Baars BJ. Covert formulation and editing of anomalies in speech production: evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior* 1982;21:578–594.
- Nooteboom, SG. Listening to oneself: Monitoring speech production. In: Hartsuiker, RJ.; Bastiaanse, R.; Postma, A.; Wijnen, F., editors. *Phonological encoding and monitoring in normal and pathological speech*. Hove, England: Psychology Press; 2005. p. 167–186.
- Nooteboom SG, Quené H. Self-monitoring and feedback: a new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language* 2008;58:837–861.
- Nooteboom, SG. Self-monitoring of inner speech and overt speech. In: Slifka, J.; Manuel, S.; Matthies, M., editors. *From Sound to Sense: 50+ years of discoveries in speech communication; Proceedings of a symposium; MIT. 11/06/2004 – 13/06/2006; 2004. p. C-187-C-190*. URL: <http://www.rle.mit.edu/soundtosense/>
- Oppenheim GM, Dell GS. Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition* 2008;106:528–537. [PubMed: 17407776]
- Postma A. Detection of errors during speech production: A review of speech monitoring models. *Cognition* 2000;77:97–131. [PubMed: 10986364]
- Postma A, Kolk HHJ. The effects of noise masking and required accuracy on speech errors, disfluencies, and selfrepairs. *Journal of Speech and Hearing Research* 1992;35:537–544. [PubMed: 1608244]
- Postma A, Kolk HHJ. The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech and Hearing Research* 1993;36:472–487. [PubMed: 8331905]
- Postma A, Kolk HHJ, Povel DJ. On the relation among speech errors, disfluencies, and self-repairs. *Language and Speech* 1990;33(1):19–29. [PubMed: 2283918]
- Postma A, Kolk HHJ. Speech errors, disfluencies, and self-repairs in stutterers in two accuracy conditions. *Journal of Fluency Disorders* 1990;15:291–303.
- Ratcliff R, Gomez P, McKoon G. A diffusion model account of the lexical decision task. *Psychological Review* 2004;111:159–182. [PubMed: 14756592]
- Ratcliff R. A theory of memory retrieval. *Psychological Review* 1978;85:59–108.
- Roelofs A. Error biases in spoken word planning and monitoring by aphasic and nonaphasic speakers: Comment on Rapp and Goldrick (2000). *Psychological Review* 2004;111:561–572. [PubMed: 15065924]
- Shattuck-Hufnagel, S. The role of word onset consonants in speech production planning: New evidence from speech error patterns. In: Keller, E.; Gopnik, M., editors. *Motor and sensory processes of language*. Hillsdale, NJ: Erlbaum; 1987. p. 17–51.
- Slevc LR, Ferreira VS. Halting in single word production: a test of the perceptual loop theory of speech monitoring. *Journal of Memory and Language* 2006;54:515–540. [PubMed: 17917694]
- Stemberger, JP. Lexical bias in errors in language production: Interactive components, editors, and perceptual biases. Carnegie-Mellon University; Pittsburgh: 1984. Unpublished manuscript
- Wagenmakers EJ, Van Der Maas HLJ, Grasman RPPP. An EZ-diffusion model for response time and accuracy. *Psychonomic bulletin & review* 2007;14:3–22. [PubMed: 17546727]
- Wheeldon LR, Levelt WJM. Monitoring the time course of phonological encoding. *Journal of Memory and Language* 1995;34:311–34.

Figure 1.

Illustration of lexical bias as explained by the three different accounts: (a) feedback account, (b) monitoring account and (c) mixed account. Dark bars represent formulated errors and light bars the observed ones. The difference between each two adjacent bars represents the degree of editing by the monitor. Error counts are hypothetical and do not reflect actual data.

Figure 2.
The nature of interaction in experiment 1. Word stimuli provide facilitation for speaking nonwords but not words.

Figure 3.

Stages of processing in the assembling task of experiment 1. Processing of the auditory syllable is not included in this scheme because measuring reaction times start with the appearance of the visual stimulus, when auditory presentation is finished.

Figure 4.

Response latencies in experiment 1. First and second clusters show reaction times to words and nonwords respectively. Dark bars represent Word condition in the first and Nonword condition in the second cluster. Light bars are response latencies in the baseline groups paired with their relevant conditions.

Figure 5. Stages of processing in the lexical decision task in experiment 2. MDT = Mean Decision Time, MRT = Mean Response Time.

Figure 6.

Comparison between the steps of monitoring and response execution for experiments 1 and 2. Dark boxes are covered in the baseline and non-decision time durations of experiment 1 and 2 respectively. Before that stage, an additional step, retrieving the phonological representation, has to be completed when the response modality is speech production.

Table 1

An example set used in experiment 1. The syllables printed in *italic* are the ones that participants should assemble. SWOW = Stimulus Word/Outcome Word; SNOW = Stimulus Nonword/Outcome Word; SWON = Stimulus Word/Outcome Nonword; SNON = Stimulus Nonword/Outcome Nonword.

trial number & type	auditory syllable	visual stimulus	outcome	trial number & type	auditory syllable	visual stimulus	outcome
1. SWOW	/ʌn/	gap	<i>gun</i>	5. SWOW	/ʌl/	miss	<i>much</i>
2. SNOW	/ʌn/	gack	<i>gun</i>	6. SNOW	/ʌl/	mez	<i>much</i>
3. SWON	/ʌl/	gap	<i>guch</i>	7. SWON	/ʌn/	miss	<i>mun</i>
4. SNON	/ʌl/	gack	<i>guch</i>	8. SNON	/ʌn/	mez	<i>mun</i>

Table 2

Error rates in experiment 1. Two types of errors are reported for each condition. See text for definitions.

Error Types		
Condition	Omissions	Commissions
Word	12%	3%
Nonword	4%	16%

Table 3

Error rates in experiment 4. See text for definitions.

Error Types		
Condition	Omissions	Commissions
Word	3%	5%
Nonword	5%	12%