Monitoring Speech Production and Comprehension: Where is the Second-Language Delay?

Wouter P. J. Broos, Wouter Duyck, and Robert J. Hartsuiker

Department of Experimental Psychology, Ghent University

Author Note

Wouter P. J. Broos, Department of Experimental Psychology, Ghent University

Wouter Duyck, Department of Experimental Psychology, Ghent University

Robert J. Hartsuiker, Department of Experimental Psychology, Ghent University

This paper received funding from the special research fund of Ghent University (GOA - Concerted Research Action BOF13/GOA/032). Correspondence concerning this article should be addressed to Wouter P. J. Broos, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2 B-9000 Ghent, Belgium, E-mail: wouter.broos@ugent.be, Tel. +32 (0)9 264 64 04.

Abstract

Research on error monitoring suggests that bilingual Dutch-English speakers are slower to correct some speech errors in their second language (L2) as opposed to their first language (L1) (Van Hest, 1996). But which component of self-monitoring is slowed down in L2, error detection or interruption and repair of the error? This study charted the time course of monitoring in monolingual English speakers and bilingual Dutch-English speakers in language production and language comprehension, with the aim of pinpointing the component(s) of monitoring that cause an L2 disadvantage. First, we asked whether phonological errors are interrupted more slowly in L2. An analysis of data from three speech error elicitation experiments indeed showed that Dutch-English bilinguals were slower to stop speaking after an error had been detected in their L2 (English) than in their L1 (Dutch), at least for interrupted errors. A similar L2 disadvantage was found when comparing the L2 of Dutch-English bilinguals to the L1 of English monolinguals. Second, monolingual English speakers and bilingual Dutch-English speakers performed a picture naming task, a production monitoring task, and a comprehension monitoring task. Bilingual English speakers were slower in naming pictures in their L2 than monolingual English speakers. However, the production monitoring task and comprehension monitoring task yielded comparable response latencies between monolinguals in their L1 and bilinguals in their L2, indicating that monitoring processes in L2 are not generally slower. We suggest that interruption and repair are planned concurrently and that the difficulty of repairing in L2 triggers a slow-down in L2 interruption.

Keywords: bilingualism, error detection, self-monitoring

Introduction

There are clear second language (L2) disadvantages in speech processing compared to speech processing in the first language (L1). Such disadvantages have been demonstrated in both L2 speech production (Ivanova & Costa, 2008; Sadat, Martin, Alario, & Costa, 2012) and L2 speech comprehension (Cop, Drieghe, & Duyck, 2015; Lagrou, Hartsuiker, & Duyck, 2011). Here we ask whether there is also a disadvantage in verbal self-monitoring in L2 (see Broos, Duyck, & Hartsuiker, 2016 for a review on L2 verbal self-monitoring). The verbal self-monitoring system is responsible for detecting and correcting speech errors and other problems in speech. Self-monitoring is a crucial aspect of language processing as it ensures that our utterances reflect our communicative intentions and conform to linguistic standards. Self-monitoring involves both error detection and, once an error is detected, processes that are responsible for interrupting speech and resuming with a repair (Hartsuiker & Kolk, 2001). As *error detection* has been argued to directly involve language comprehension (Levelt, 1983), language production (Nozari, Dell, & Schwartz, 2011), or both (Pickering & Garrod, 2013), L2 disadvantages in either modality could slow down detection and hence the moment after the error when speech is interrupted. The process of *repairing* the error most likely involves language production (Hartsuiker & Kolk, 2001). An L2 disadvantage in production might therefore slow down repair onset. It is also possible that production disadvantages will slow down *interruption*, on accounts assuming parallel planning of interruption and repair, with slower repairing delaying interruption onset. This was proposed by Hartsuiker, Catchpole, De Jong, and Pickering (2008) who showed that speakers interrupted a word more slowly when the repair was more difficult to plan. Hence, the current study asks whether there is an L2 disadvantage in self-monitoring and whether any such slow-down originates from error detection, interruption and repair processing, or both.

Are L2 speakers indeed slower in self-interruption or self-repair? Van Hest (1996) compared the time-course of L1 and L2 speech monitoring in bilingual Dutch-English speakers. She elicited several types of speech errors by means of a story-telling task and an interview task. Participants more often repaired their speech in their L2 (English) than in their L1 (Dutch). The types of errors in the L2 were also different: Errors tended to be more grammatical, lexical, and phonological in nature while L1 repairs were mostly appropriateness repairs[1]. Importantly, differences were also found in the speed with which errors were repaired. In particular, she measured the error to cut-off interval (the lag between the error onset and speech interruption) and the cut-off to repair interval (the lag between speech interruption and error repair). She found that the cut-off to repair intervals were longer in L2 but only for appropriateness repairs. The error to cut-off interval and cut-off to repair interval of phonological, lexical, and grammatical errors did not differ between L1 and L2.

It is surprising that Van Hest observed no language effect on the error to cut-off intervals, as processes that are used for error detection and repair (perception and/or production) are slower in L2. However, only very few observations were analysed: there were 33 appropriateness repairs (16 in L1, 17 in L2) and 36 repairs of phonological errors (20 in L1, 16 in L2). The absence of a language effect for most error types may therefore be a power issue. Additionally, Van Hest did not distinguish between interrupted and completed errors (see also Hartsuiker, Corley, & Martensen, 2005, Hartsuiker et al., 2008; Gambi, Cop, & Pickering, 2015; Nooteboom & Quené, 2017). This distinction is important, as interrupted vs. completed errors may reflect different monitoring processes (Nooteboom & Quené, 2017). Thus, an imbalance in the number of interrupted and completed errors might skew the results. Hence, a further test is needed to establish whether there is an L2 disadvantage in error monitoring. Below, we report such a test for the case of phonological errors. Phonological errors are elicited because detection processes (and therefore time intervals) might differ

depending on what type of error is produced. Hence, we chose to focus on phonological speech errors. But first we review the evidence for L2 disadvantages in language production and comprehension.

L2 Disadvantages in Speech Production and Comprehension

L2 speakers are slower (compared to L1) at several basic language processes, such as word recognition and production in the visual and auditory modalities (Cop et al., 2015; De Bot & Schreuder, 1993; Flege, Frieda & Nozawa, 1997; Gollan & Silverberg, 2001; Ivanova & Costa, 2008; Kroll & Stewart, 1994; Lagrou et al., 2011; Sadat et al., 2012; Schreuder & Weltens, 1993). With respect to the auditory modality, Lagrou et al. (2011) tested Dutch-English bilinguals and English monolinguals and asked them to perform an auditory lexical decision task. Bilingual L2 English listeners were slower at the task than monolingual L1 English listeners. This same language effect is seen in reading. In an extensive study that focused on natural reading in the L2, Cop et al. (2015) asked whether Dutch-English bilinguals were slower to read an entire novel in English (L2) than in Dutch (L1). L2 readers took longer to finish a sentence, needed more fixations, and did not skip as many words as L1 readers.

Broos, Duyck, and Hartsuiker (in press) recently investigated the L2 disadvantage that is frequently seen during picture naming. In particular, we addressed the question of whether the L2 slow-down is situated at pre-phonological stages (as argued by Gollan, Montoya, Cera, and Sandoval, 2008) or whether this delay occurs post-phonologically (see Hanulová, Davidson, and Indefrey, 2011). Both a picture naming task and a phoneme monitoring task were used. The picture naming task was included to confirm the L2 delay in picture naming in Dutch-English bilinguals, whose response latencies were compared to monolingual English

speakers. During the phoneme monitoring task, both participant groups were asked to press a button if a particular phoneme was present in the English name of a picture that was presented on the screen. This monitoring task arguably involves lexical retrieval and phonological encoding, but not phonetic encoding or actual articulation (i.e., post-phonological stages of speech production). Phonetic encoding will most likely not occur, as speech does not actually have to be produced in the monitoring task and the phonological code is already sufficient to carry out the task. The monitoring task was combined with a picture-word interference paradigm in order to check that phoneme monitoring taps into regular phonological encoding processes. The distractor words could phonologically overlap with the English picture name (e.g., **b**a**g** – **b**u**g** / ba**g** – fo**g** / **b**ag – **b**et) or not (bag – rod). Phonological overlap between the picture name and distractor word sped up response latencies, just as it does in the regular picture-word interference task; validating the assumption that this task indeed tapped into regular phonological encoding (see also Wheeldon & Levelt, 1995). Dutch-English bilingual speakers were slower to name pictures in their L2 than monolingual speakers. Importantly, there was no L2 disadvantage in the phoneme monitoring task, suggesting that the L2 slow-down in picture naming is situated at post-phonological stages. The main difference between the current study and that of Broos et al. pertains to the speech error elicitation task and the addition of distractor words, which are both added to the current study (see below).

In sum, many studies have revealed L2 disadvantages in several modalities of language processing. L2 speakers are consistently slower at listening (Lagrou et al., 2011), reading (Cop et al., 2015), and speaking (Ivanova & Costa, 2008; Sadat et al., 2012). Our recent study (Broos et al., in press) suggests a late (post-phonological encoding) locus of this disadvantage in word production. Given that self-monitoring arguably involves comprehension and/or production, such L2 delays might slow down certain aspects of self-

monitoring too. We now examine what aspects of monitoring might be affected by such delays.

Self-Monitoring Theories and Potential Delays

Self-monitoring involves a phase of error detection and a subsequent phase of responding to that error, which usually involves interrupting ongoing speech and producing a repair (of course it is also possible that the speaker sometimes decides to ignore a detected error). As we explain below, slower production and/or comprehension can slow down error detection (leading to longer error to cut-off intervals), slower interruption and repair of the error (which increases both the error to cut-off and the cut-off to repair intervals), or both components.

Theories of *error detection* differ in whether they assume error detection uses the comprehension system or the production system. A theory of self-monitoring that assumes error detection uses comprehension is Levelt's (1983) perceptual loop theory, which argues that speech monitoring is based on the comprehension system. This particular theory assumes that there are three loops: the conceptual loop, the inner loop, and the outer loop. The conceptual loop is used to determine whether particular words or expressions are appropriate for a specific context. The inner loop monitors the phonological and phonetic code of an utterance (the "speech plan") before it is pronounced. Finally, the outer loop is based on auditory perception of one's own overt speech. Importantly, the inner loop and the outer loop are both based on the speech comprehension system. All the information from these loops is directed towards a central monitor that decides whether or not a problem has occurred, and this monitor therefore uses comprehension as a basis for error detection. An L2 detection delay could then be explained by arguing that comprehension is slower in L2.

A more recent self-monitoring theory assumes that error detection uses only production-internal mechanisms. The interactive two-step model of Nozari et al. (2011) argues that error detection is performed by comparing activation levels of competing representations. If no speech errors are made, only the lexical representation of the correct word or phoneme will be highly activated (a situation of low conflict). If an error is made, however, both the correct and incorrect lexical representations are activated, leading to competition (a situation of high conflict). Conflict acts as a signal for the self-monitoring system in order to detect errors. An L2 detection delay could then be explained on the assumption that lexical and phonological representations are activated more slowly in L2 than in L1 (Strijkers, Baus, Runnqvist, FitzPatrick, & Costa, 2013; but see Broos et al., in press; Hanulová et al., 2011). Hence, it would also take longer to detect that there is conflict. In sum, theories differ in whether error detection takes place in comprehension or production. Both accounts are compatible with an L2 delay in monitoring, as both comprehension and production are delayed in L2. Any L2 delay in detection would be reflected in a longer error to cut-off interval in L2, as slower error detection postpones the moment at which speech can be interrupted.

Alternatively, an L2 monitoring delay can also reflect a delay in *interruption and repair* of the error. Repairing necessarily involves the language production system, by restarting part of the utterance from scratch (e.g., Hartsuiker & Kolk, 2001), by producing the second-best speech plan available (Nozari et al., 2011), or by editing a stored representation of the utterance (Boland, Hartsuiker, Pickering, & Postma, 2005). Hence, under the assumption that production is slower in L2 but self-interruption is constant, delays to language production should increase the cut-off to repair interval in L2 relative to L1. Additionally, as the repair *itself* might be monitored, slower comprehension of the repair in L2 might further increase this interval assuming that the monitor only admits the repair if it is

adequate. Thus, as speech is produced more slowly in L2, the repair, which is created in the same way as the original utterance, will also take more time to be constructed, resulting in an increased cut-off to repair time (as Van Hest, 1996, indeed found for appropriateness repairs).

It has also been argued that interruption and repair take place concurrently and that they share cognitive resources, so that factors slowing down repair will also slow down interruption (Hartsuiker et al., 2008; also see Gambi et al., 2015; Tydgat, Stevens, Hartsuiker, & Pickering, 2011). For instance, Hartsuiker et al. (2008) presented participants with a to-be-named picture and asked them to occasionally replace it with the name of a new picture while they were in the process of naming the first picture; participants were asked to interrupt their first response and replace it with the name of the replacement picture. This replacement picture could either be intact or visually degraded. The key finding was that if the replacement picture was visually degraded and hence harder to name, it took longer to interrupt the initial picture name. It is possible that speaking in L2 is similarly just harder than speaking in L1 as representations are less detailed in L2 as compared to L1. This results in a slow-down of interruption and hence longer error to cut-off intervals (but not necessarily cut-off to repair intervals).

Finally, it is possible that both error detection and interruption/repair are slower in L2. Specifically, the slower rate of speech production and comprehension in L2 itself might account for longer time intervals during repairs of certain types of errors. Indeed, Oomen and Postma (2001) demonstrated that error to cut-off and cut-off to repair intervals became longer with slower speech rates. Hartsuiker and Kolk's (2001) computational model of self-monitoring simulated these data, on the assumption that in slower speech, *all* production and self-comprehension processes become slower. An error will therefore be detected and repaired later in slower speech, leading to a longer error to cut-off and cut-off to repair interval.

The Current Studies

The first study we performed contained three experiments (Experiment 1, 2, and 3) that tested whether there is an L2 disadvantage during monitoring for phonological errors. This study tested whether there is indeed a phonological L2 monitoring delay and if so, it will help us delineate which monitoring components (error detection, interruption and repair, or both) are responsible for this delay. We decided to measure the time course of error interruptions and repairs from three error-elicitation experiments (i.e., the error to cut-off and cut-off to repair intervals). This approach has the advantage that the errors were collected under controlled circumstances and all concerned the same linguistic representational level (phonology).

The second study also contained three experiments (Experiment 4, 5, and 6), all with the same subjects and stimuli: a picture naming task, a phoneme monitoring in production task, and a phoneme monitoring in comprehension task. We asked bilingual Dutch-English and monolingual English participants to monitor for particular phonemes in multiple modalities in English, with the aim of determining the origin of any L2 slow-down during error monitoring. During the production monitoring task, the speaker produces speech internally, inspects an internal speech code, and then compares it to a target. An L2 disadvantage in this task would suggest that an L2 slow-down of monitoring could either be situated at the early, lexical stages of production or at comprehension processes. However, if an L2 disadvantage is not found (as in Broos et al., in press), it would suggest that later, post-phonological stages of speech production are responsible for the L2 slow-down. If an L2

disadvantage would be found in the comprehension monitoring task, this would suggest that the comprehension processes are responsible for slower monitoring. In this task, speech is merely perceived and production processes are not performed. The picture naming task taps into both early and late processes of speech production. Based on previous findings of L2 speech production studies (including Broos et al., in press, testing very similar groups of subjects), we hypothesise that bilingual Dutch-English speakers will make more errors and will be slower in naming pictures in English than monolingual English speakers. If the slow-down is *only* observed in this task, then slower production and/or repair is responsible for the L2 disadvantage.

Our studies compare L1 English monolingual speakers with L2 English bilingual speakers. This was the only comparison that could be performed while still being able to use the same stimuli across all experiments. For example, the translation of the English target picture name ‘broom’ is ‘bezem’ in Dutch. Not only would there be a difference in the number of syllables, the syllable ‘be’ in ‘bezem’ does not have a coda. This means that we would have needed to use different stimuli for the L1 Dutch experiment, which in turn would have lead to a comparison of experimental results where stimuli are not matched.

The reason why the phoneme monitoring tasks are able to elucidate processes of error monitoring is because several important processes that are needed for both phoneme and error monitoring are shared. Specifically, the production monitoring task requires lexical access, phonological encoding, and inspection of the internal speech code. We argue that internal error monitoring at the phonological level requires the same processes. This is true both on a comprehension monitoring account (e.g., Levelt et al., 1999, Özdemir, Roelofs, & Levelt, 2007) according to which an internal phonological representation is inspected by the speech comprehension system and by a production monitoring account (e.g., Nozari et al., 2011), according to which a monitoring device inspects the pattern of activation in a layer of

phonological units. Furthermore, the comprehension monitoring task necessarily requires basic auditory processing and speech perception and these processes are also involved in error monitoring via the auditory monitoring channel. We acknowledge, of course, that error monitoring and phoneme monitoring differ in their criteria (i.e., is this an error? vs. is this the target phoneme?) and in whether the process occurs consciously or not. However, the phoneme monitoring task is still arguably the best proxy to investigate which processes of error monitoring are delayed as the inspection of internal and external speech codes occurs in both tasks (either explicitly or not).

Study 1: Analysis of Speech Error Data

Experiment 1, 2, and 3

Below we ask whether language (L1 vs. L2 within bilinguals) and Language group (L1 monolinguals vs. L2 bilinguals) affect the time course of speech interruption and repair. We analysed results from three experiments that used the Spoonerisms of Laboratory-Induced Predisposition technique (also known as the SLIP task). This task was first used by Baars, Motley, and Mackay (1975) to elicit phonological speech errors (sometimes called Spoonerisms) where the first consonant of two words are switched (e.g., when 'pig – bill' becomes 'big – pill'). During this task, people are presented with a series of word or non-word pairs and are asked to silently read these pairs. When they hear a beep, they must pronounce the pair they see on the screen as quickly as possible. The pair that has to be pronounced, the target pair, is always preceded by several *biasing pairs* with the reverse phonological construction (i.e., with the initial consonants of the two items swapped). Thus, if the target pair would be 'pig – 'bill', then an example of a biasing pair would be 'bind – pipe'.

Phonological priming by the biasing pairs increases the number of speech errors. It is typically found that errors are produced more often if they result in a word pair rather than a non-word pair (the lexical bias effect, see Baars et al., 1975, Hartsuiker et al., 2005, Nooteboom and Quené, 2008, and many others).

For our purposes, the types of errors are not relevant; rather we focus on the time intervals of error to cut-off and cut-off to repair both within bilinguals (bilingual L1 and L2; Experiments 1 and 2) and between monolingual L1 English speakers (Experiment 3) and the bilingual L2 English speakers from Experiments 1 and 2. Note that the task elicits phonological errors. This has the advantage that it is the same linguistic level on which our subsequent (phoneme monitoring) experiments will focus. Stimuli for all three experiments are placed in Appendix B, C, and D. SLIP Experiments 1 and 2 are reported in full in a preprint published on Open Science Framework (https://osf.io/egr93/). SLIP Experiment 3 is not reported in the preprint but had the exact same procedure (only the stimuli differed, as can be seen in Appendix D).

Method

We tested 171 speakers in three experiments: 48 non-balanced bilingual Dutch-English speakers participated in SLIP Experiment 1, 48 participants of the same participant pool participated in SLIP Experiment 2, and 75 monolingual English speakers participated in SLIP Experiment 3. Participants were monetarily compensated and recruited at Ghent University (Experiment 1 and 2) and at the University of Leeds as well as at the University of Edinburgh (Experiment 3). All bilingual speakers received formal education in English starting from the age of 12 in secondary school, receiving three to four hours of English lessons a week. Next to formal instruction, Belgian students are confronted with English video games, books,

television series, and other media (also before age 12). All participants reported to have normal hearing and normal or corrected-to-normal sight. None of the participants were diagnosed with dyslexia. The LexTALE was used as a post-test to assess English proficiency (Lemhöfer & Broersma, 2012). This test is a lexical decision task that has been argued to provide a reliable and valid measure of English proficiency. An overview of all LexTALE scores for all tasks is provided in Table 1 below.

[Insert Table 1 here]

Language group comparisons were performed across studies in order to show that there were no significant differences in English proficiency between the two L1 monolingual English groups or the two Dutch-English bilingual groups. Two-sample t-tests were used to compare both the L1 groups in Study 1 (SLIP task) and Study 2 (Naming/Monitoring task) and the L2 groups across studies. There were no significant differences between L1 groups ($t(97.75) = -.02$, $p = .50$) or L2 groups ($t(113.7) = -1.05$, $p = .15$).

In SLIP Experiment 1, bilingual participants were asked to silently read word and non-word pairs in L1 and L2 and to produce some non-word pairs in four blocks that differed in their composition. Each block consisted of 400 trials of which 80 trials were to be pronounced (there were thus 1600 trials per participant of which 320 were to be pronounced). The blocks could contain English non-word pairs, Dutch non-word pairs, English word and non-word pairs, or Dutch word and non-word pairs. Hence, language and lexical context were manipulated. The Dutch and English non-word pairs were created based on phonological characteristics of either language. For instance, the bigram /sh/ can occur at the beginning and end of English words, but not of Dutch ones (e.g., ‘show’ or ‘push’) meaning that the non-word pair ‘shik – mish’ could be considered an English non-word pair. Every target pair was

non-lexical and could either result in word or non-word pairs after switching the first two consonants of the individual words (a word pair after switching would be ‘hust – dunt’ instead of ‘dust – hunt’ while a non-word pair after switching would be ‘fais – raig’ instead of ‘rais – faig’). Control pairs were inserted in order to obscure the purpose of the experiment. We ensured that none of the word pairs used in the experiment consisted of Dutch-English cognates or false friends. In Experiment 2, further Dutch-English target pairs were presented with similar blocks as those described in Experiment 1. But now, every block was a mixture of word and non-word pairs (i.e., mixed context). Moreover, target pairs were not only made up of non-words but also contained words. Experiment 3, with English monolinguals, consisted of three blocks of English stimuli; two blocks were made up non-word target pairs and one block of word target pairs and all blocks had a mixed context.

During the experiments, participants were seated in front of a computer screen in a quiet room. They were asked to wear headphones that played back white noise of 70 decibels, following the procedure of Baars et al. (1975) and Hartsuiker et al. (2005). The participants were instructed to silently read the word pairs that were presented on the screen. However, if they heard a beep over the headphones, they were asked to name the last word pair they saw on the screen as quickly as possible. Participants only heard a beep if the word pair was a target pair or control pair. They were asked to pronounce the word or non-word pair as quickly as possible and to make sure that they finished speaking before they heard the second beep (where the time between the first and second beep amounted to 1000 ms). The next trial was presented immediately after the second beep. The inter trial interval (ITI) was identical in L1 and L2. Responses were annotated in Praat (Boersma & Weenink, 2018) after the experiment ended and errors were categorised into interrupted errors (e.g., ‘d…hust-dunt’) and completed errors (e.g., ‘dust-hun…hust-dunt’). This categorisation was made since Hartsuiker et al. (2005, 2008) and Gambi et al. (2015) also considered these two types of

interruptions separately. Error to cut-off intervals (the time lag between the error and speech interruption) and cut-off to repair intervals (the time lag between speech interruption and repair) of both error categories were measured in milliseconds.

Results

The three experiments combined resulted in 286 repairs, allowing us to measure the error to cut-off and cut-off to repair intervals. The total number of missed trials in SLIP Experiment 1 and 2 amounted to 29/3840 (.76%) for L1 blocks and 32/3840 for L2 blocks (.83%). SLIP Experiment 3 contained a total of 66/4500 (1.47%) missed trials. Separate linear mixed effects models were created for the error to cut-off and cut-off to repair intervals. The only fixed factor that was included in each model was Language or Language group (depending on whether the comparison was within or between participants), while taking subject and item variability into account. No random slopes were added, because the models did not converge if these were included.

[Insert Table 2 here]

Table 2 clearly shows that bilingual Dutch-English speakers were much slower to stop speaking after making an error in their L2 than in their L1, at least for interruptions where the first word was not completely pronounced. The same holds for the comparison between L1 monolingual English speakers and L2 bilingual Dutch-English speakers, where L2 English was slower. The cut-off to repair intervals did not significantly differ.

Discussion

Contrary to the findings of Van Hest (1996), we did find an L2 delay in phonological errors in the error to cut-off interval. The delay was approximately 115 ms in both the estimated and observed reaction times when comparing L1 and L2 within participants. There was also a delay of almost 70 ms in the between participant comparison (i.e., L1 monolingual English vs. L2 bilingual English). These findings are compatible with an account according to which phonological error detection takes place more slowly in L2 than L1. It is also possible that these delays result from slower interruption/repair processes in L2, so that any difficulty in resuming in L2 is reflected in postponed interruption. The data are less compatible with accounts assuming a delay only in repairing (with a constant interruption time) or assuming an L2 delay across the board (in detection and repair) as such accounts predict an L2 delay in cut-off to repair intervals as well.

Note that the L2 delay in error to cut-off times was only found for errors that were interrupted and not for completed errors. However, the number of completed errors was so small that it would be inadvisable to draw strong conclusions about this category. Moreover, the cut-off to repair intervals were short, not even 200 ms in either interrupted or completed errors, supporting the notion that speech is interrupted when the repair is (almost) ready to be produced and vice versa (see Hartsuiker et al., 2008 for further discussion on this topic).

Study 2: Picture Naming and Phoneme Monitoring Experiments

Experiments 4, 5, and 6 described below aim at teasing apart the remaining accounts: the L2 delay on interruptions is either a result of delayed error detection or of postponed interruption triggered by slower repair. If the former account is right, the detection delay could either be a result of delayed comprehension or production. We test these accounts in three experiments

that ask subjects to monitor for phonemes in language production, monitor for phonemes in language comprehension, and to name pictures. We focus on bilingual Dutch-English speakers and monolingual English speakers who performed the tasks in English due to stimuli constraints (see above). We present the three tasks as separate experiments for expository reasons. All experiments were completely counterbalanced, contained the exact same pictures, were completed by the same participants, and were all performed in English. In order to be able to counterbalance task order (six possible orders) and stimuli (three stimuli lists for three tasks), we created 18 different versions of the experiment. It was impossible to conduct these experiments with the same pictures in Dutch as this would lead to constraint violations of the stimuli (see below).

Experiment 4: Picture Naming

Method

Participants

We tested 108 participants, namely 54 non-balanced Dutch-English bilinguals (male = 14, mean age = 23) and 54 English monolinguals (male = 10, mean age = 30). Bilingual participants were tested at Ghent University whereas monolingual participants were tested at the University of Leeds. The same equipment was used to test both participant groups. All participants were monetarily compensated for their participation. All L2 speakers received formal education in English starting from the age of 12 in secondary school, receiving three to four hours of English lessons a week. Next to formal instruction, Belgian students are confronted with English video games, books, television series, and other media (also before

age 12). All participants reported to have normal hearing and normal or corrected-to-normal sight. None of the participants were diagnosed with dyslexia. The LexTALE was used as a post-test to assess English proficiency (Lemhöfer & Broersma, 2012). The difference in LexTALE-scores between L1 and L2 speakers (Table 1) was significant ($t(94.89) = 7.62$, $p < .001$).

Materials

Stimuli. Twenty-five black-and-white pictures (see Appendix A for overview of target picture names) were selected from the Severens, Lommel, Ratinckx, and Hartsuiker (2005) database. In addition to the 25 target pictures per list, we selected 25 filler pictures, which were used in every stimulus list. Hence, every participant was asked to name 50 pictures. Exactly two-third of all target pictures was monosyllabic while the remaining one-third consisted of disyllabic nouns. The reason to include this factor stems from the availability of the useable stimuli in the monitoring tasks; the picture database did not contain sufficient monosyllabic picture names that fit the conditions of the monitoring experiments.

Procedure

Participants were seated in a quiet room and were positioned in front of a computer screen. Before the experimental phase started (Figure 1), participants were presented with *one* familiarization phase at the beginning of the experiment in which they saw all the pictures used in this task on the screen with their corresponding names written underneath. During the experimental phase, participants saw these same pictures again (in a different order) without the corresponding names and they were asked to pronounce the English picture name as fast and accurately as possible. A fixation cross was presented for 250 ms after which a blank

screen was displayed for 250 ms. Subsequently, the picture was presented for 3000 ms followed by another blank screen of 250 ms before the next trial began. Reaction latencies were measured from the moment the picture was displayed on the screen by means of a recording that was started by E-prime 2.0. Every trial was recorded separately and annotated in the computer program Praat.

[Insert Figure 1 here]

Data Analysis

The total number of target trials amounted to 2700 (108 participants times 25 trials). Due to technical difficulties, 60 trials were not recorded. In total, 5.77% (77/1334) of the trials was answered incorrectly by L1 speakers while 10.41% (136/1306) was answered incorrectly by L2 speakers. A trial was considered an outlier when the response latency for that trial was 2.5 standard deviations away from the group mean. The total number of outliers in the picture naming task was 45 out of 2427 trials (1.85%). Outliers and trials that were answered incorrectly were removed from the data set before the data were analysed.

The cleaned data sets were analysed by means of linear mixed effects models with the lme4 (version 1.1-15), car (2.1-5), lsmeans (2.27-2) and lmerTest (version 2.0-33) package of R (3.4.1) (R Core Team, 2013). By applying this analysis, both subject and item variability can be taken into account (Baayen, Davidson, & Bates, 2008). Sum coding was used for all analyses where the mean of all factors amounted to zero. Type II Wald Chi square tests were used to test both main effects and interaction effects. The function ‘lsmeans’ was used to determine significant differences between all different contrasts and to assess the importance of Language group as a factor in the model. Random slopes were included based on the ‘maximal random effects structure’ approach, as suggested by Barr, Levy, Scheepers, and

Tily (2013). As we needed to use both monosyllabic and disyllabic target nouns (for practical reasons), we also included the factor Number of syllables in the models. Models were validated by plotting the residuals of the linear mixed effects model while inspecting random distribution of the residuals. The R-scripts and data sets for the analyses of the current experiments can be found on Open Science Framework (https://osf.io/xwp98/).

Results

Reaction Times

We first tested for normality by using the Shapiro-Wilk test ($W = .83$, $p < .001$). As the data set was not normally distributed, we transformed response latencies to $logRT^2$. The final model for the picture naming task included the fixed factors Language group, Number of syllables and their interaction. The maximal random effects structure of the final model contained Language group as random slope to item (Picture) and Number of syllables to subject (Subject). The reason why both fixed factors can only be added as random slope to one random intercept is that Language group is a between-subject variable whereas Number of syllables is a between-item variable. The factor Number of syllables consisted of the two levels monosyllabic and disyllabic picture names while Language group consisted of L1 English of monolinguals and L2 English of Dutch-English bilinguals.

[Insert Figure 2 here]

Figure 2 shows that the bilingual Dutch-English speakers were slower in naming the pictures in English than the monolingual English speakers (Effect of Language group: $\chi^2(1) = 29.64$, *p*

< .001). There was no effect of Number of syllables (χ^2 (1) = .90, p = .34) and no interaction of Language group and Number of syllables (χ^2 (1) = .008, p = .93).

A comparison between a model with Language group and one without Language group would not only test for Language group itself but also for all of its interactions. We therefore used lsmeans to assess the importance of the fixed factor Language group in the model, which is a more appropriate way of testing for the importance of factors. Language group significantly contributed to the model (β = -.15, SE = .03, t = -5.33, p < .001). It is not evident to determine an effect size (e.g., Cohen's d) in an analysis using mixed effects models. As an indication of effect size, we estimated the proportion of variance explained by the factor Language group by using the package 'r2glmm' (version 0.1.2) (see Jaeger, Edwards, Das, and Sen, 2017). In particular, we determined the proportion of variance explained (R^2) for the model, the predictors, and the interactions of predictors. The model explained a total variance (R^2) of .232 of which Language group could explain .178. Thus, Language group accounted for 76.7% of variance in the model. The variance of the entire model is not exceptionally high, which is probably due to the lack of lexical variables in the model. Yet, this analysis does show that Language group accounts for a large proportion of the explained variance and therefore corresponds to a large effect size.

Accuracy

The types of errors that were included in the current analyses were trials that were unanswered and trials where a different picture name than the target picture name was used. Figure 3 shows the accuracy scores in percentages by Language group and Number of syllables.

[Insert Figure 3 here]

A generalized mixed effects model that was created with a logit link function was run to determine whether L2 speakers were less accurate than L1 speakers. The fixed factor Language group and Number of syllables were included and an interaction of these factors was added. Number of syllables was added as random slope to subject (Subject) while Language group was included as random slope to item (Picture). There was a main effect of Language group (χ^2 (1) = 6.15, p = .01) indicating that bilingual Dutch-English speakers were less accurate in their L2 than monolingual English speakers in their L1. The factor Number of syllables was not significant (χ^2 (1) = .06, p = .81) nor was there an interaction of Language group and Number of syllables (χ^2 (1) = .14, *p* = .70).

Discussion

Experiment 4 clearly shows that English monolingual speakers are faster and more accurate when naming pictures in their L1 when compared to Dutch-English bilingual speakers. The advantage in naming latency is more than 100 ms. The control variable Number of syllables of the target word did not affect the speed or accuracy on picture naming. In sum, there is a clear L2 disadvantage in picture naming.

Experiment 5: Production Monitoring Task

Method

Participants

The same participants who performed Experiment 4 also participated in Experiment 5.

Materials

***Design*.** The same design was used as in Experiment 4.

***Stimuli*.** We used the same 75 black and white line drawings as in Experiment 4 in the three stimulus lists, with each list containing 25 target pictures. The target phoneme could be situated at either the onset or the coda of the picture name. In case of a disyllabic picture name, the final consonant of the first syllable was considered the coda. In one half of the trials, the target phoneme was present (target trials) while it was absent in the other half (filler trials). All target phonemes were consonants (i.e., /m, l, k, s, t, f, d, p, r, w, n, b, z, g, h/); they were presented to the participants as their corresponding letters. The total number of target trials in this task was 50, twice as much as in the picture naming task because there were now two trials per target picture: one trial for the onset phoneme and one for the coda phoneme. The total number of filler trials also amounted to 50 as an equal number of filler trials were inserted for these same target pictures. So, every participant saw each target picture four times and completed 100 trials as the variable 'position' (onset vs. coda) was nested under the absent/present manipulation condition. Picture names were mono- and disyllabic nouns and the mapping between orthography and phonology was regular for all picture names. There were several restrictions pertaining to the presentation of the stimuli: 1. No more than three trials with the same correct answer were presented in a row (yes or no) / 2. No more than three successive trials were presented where the target phoneme was presented at either the beginning or end of the word (onset vs. coda) / 3. A maximum of two trials with identical target phonemes were presented in a row.

Procedure

Participants were seated in a quiet room and were positioned in front of a computer screen. Participants first saw a letter on the screen after which they saw a picture (Figure 4). They

were asked to press the green button (right) if the letter was present in the picture name and the blue button (left) if it was absent. In order to avoid unnecessary variation in reaction times, participants were asked to keep their hands near the buttons when responding and to be as fast and accurate as possible. A fixation cross was presented for 250 ms after which a blank screen followed that also lasted for 250 ms. The target letter was displayed on the screen for 1000 ms after which another blank screen followed for 250 ms. A fixation cross and blank screen were shown respectively (both displayed for 250 ms) after which the picture was presented. The experiment continued only if the participant responded to the trial. A final blank screen was presented on the screen for 250 ms before the next trial began.

[Insert Figure 4 here]

Data Analysis

A total of 10800 trials were performed (108 participants times 100 trials). The trials where the target phoneme was absent (filler trials) were not included in the final analyses, leaving a total of 5400 target trials. This means that every Language group has 2700 target trials. We excluded 322 trials because of problems with the stimuli that were discovered after the experiment had been run[3 4]. L1 speakers made errors on 13.52% of the trials (353/2610) whereas L2 speakers made errors on 13.79% of the trials (360/2610). We excluded 1.24% of the trials as outliers (56/4507). Outliers were identified in the same way as in the picture naming task.

Results

Reaction Times

Data was transformed to logRT as the Shapiro-Wilk test revealed that the data was not normally distributed ($W = .81$, $p < .001$). The final model contained the fixed factors Language group, Place, and Number of syllables. Interactions of these fixed factors were included in the model as well. Place and Number of syllables were added as random slopes to subject (Subject) and Place and Language group were added to item (Picture). Models were validated by plotting the residuals of the linear mixed effects model.

[Insert Figure 5 here]

As shown in Figure 5, there was an effect of the factor Place (χ^2 (1) = 118.01, $p < .001$), with faster responses when the target phoneme was positioned in the onset of the picture name. The factors Language group (χ^2 (1) = .22, $p = .64$) and Number of syllables (χ^2 (1) = 2.00, $p = .16$) were not significant. The interaction of Place and Language group was significant (χ^2 (1) = 19.67, $p < .001$), indicating that the Place effect was larger in L1 than in L2. No other interactions were significant (all p-values > .1).

Further analyses of Language group within the factor Place were performed by means of contrast comparisons in order to observe the effect of language per position. The package lsmeans was used to obtain all of the contrast comparisons of Language group and Place. In the onset, the difference between monolinguals and bilinguals was not significant ($\beta = -.02$, $SE = .04$, $t = -.47$, $p = .64$), but it did reach significance in coda position ($\beta = .08$, $SE = .04$, $t = 2.08$, $p = .04$). Lsmeans was also used to test for the importance of Language group in the model. Language group did not turn out to significantly contribute to the model ($\beta = .03$, $SE = .03$, $t = .87$, $p = .39$). We also assessed whether the effect size of the factor Language group

was high or low. The linear mixed effects model explained a total variance of .559 of which Language group could only explain .007 (1.3%). Hence, the presence of Language group in the model does not help in explaining the total variance; Language group is therefore considered to have a small effect size.

Accuracy

Figure 6 below shows the distribution of accuracy scores as a function of Number of syllables, Place, and Language group in percentages.

[Insert Figure 6 here]

A generalized linear mixed effects model with a logit link function was created for accuracy. The fixed factors in the final model were Language group, Place, and Number of syllables. Interactions of these fixed factors were included in the model as well. Place was added as random intercept to both subject (Subject) and item (Picture), Language group was added to item (Picture), and Number of syllables was added to subject (Subject). Most importantly, no significant difference was found between monolinguals and bilinguals (χ^2 (1) = .10, p = .76). The only significant main effect was that of Place (χ^2 (1) = 48.55, p < .001) with higher accuracy for target phonemes in onset position. The interaction between Language group and Place also reached significance (χ^2 (1) = 15.25, p < .001). No other main effects and interactions were significant (all p-values > .1). Since an interaction between Language group and Place was found, Language group contrasts within onset and coda were compared. In the onset, the difference between L1 and L2 was significant (β = .72, *SE* = .24, *z* = 3.07, *p* = .002)

where L1 speakers were more accurate than L2 speakers. This difference did not reach conventional levels of significance in the coda (β = -0.23, *SE* = 0.13, *t* = -1.83, *p* = .07).

Discussion

Most importantly, Experiment 5 did not reveal a main effect of Language group concerning response latencies in production phoneme monitoring. Moreover, Language group did not help in explaining the total variance in the model. This study replicates the findings of the study of Broos et al. (in press) who also did not find a main effect of Language group while testing 97 subjects. Important to note is that every subject in Broos et al. also performed a picture naming task and a phoneme monitoring task in production. The only differences in methodology of Broos et al. was the inclusion of phonologically related or unrelated distractor words in the picture naming task and phoneme monitoring in production task, rendering it a picture-word interference paradigm.

There was one significant interaction between Language group and Place and when analysing contrasts in more detail, the effect was shown to be driven by the coda trials. Note however that instead of an L2 delay, there seemed to be an effect in the other direction where bilinguals were somewhat *faster* in the coda condition than monolinguals (see below for a more elaborate discussion). As no such interaction was found in Broos et al. (in press), with very similar procedures and stimuli, we suspect that this effect is a false positive. Place of the target phoneme greatly influenced the speed and accuracy with which the phoneme was monitored. Phonemes were monitored more quickly and more accurately when these were positioned at the onset of the target picture name, consistent with findings from Wheeldon and Levelt (1995). Number of syllables did not show an effect meaning that participants did not react differently to disyllabic picture names as opposed to monosyllabic ones. Analyses on the accuracy data mostly replicated the patterns of results found in response latency analyses.

In short, it seems that the L2 delay in error to cut-off times cannot be easily attributed to lexical selection, phonological encoding and/or processes of inspecting an internal phonological code because no main Language group effects were found on reaction times. We next turn to the comprehension monitoring task, which taps into language comprehension processes.

Experiment 6: Comprehension Monitoring Task

Method

Participants

The same participants who performed Experiment 4 and 5 also participated in Experiment 6.

Materials

Design. The same design was used as in Experiment 4 and 5.

Stimuli. The criteria and number of stimuli used in this task were identical to that of the production monitoring task (Experiment 5). The only difference here was that participants were asked to monitor for a phoneme in the incoming speech stimulus that was auditorily presented. Stimuli were recorded by means of a USB-microphone (SE electronics, USB 1000a Plug and Play USB microphone). A female native English speaker pronounced the stimuli in standard British English.

Procedure

The procedure of the comprehension monitoring task (Figure 7) was identical to that of the production monitoring task with the exception that a recording of the English picture name was presented through headphones instead of the picture being shown on the screen.

[Insert Figure 7 here]

Data Analysis

A total of 10800 trials were performed (108 participants times 100 trials). The trials where the target phoneme was absent (filler trials) were not included in the final analyses, leaving a total of 5400 target trials. L1 speakers responded incorrectly on 7.98% of the trials (210/2631) whereas L2 speakers responded incorrectly on 8.52% of the trials (224/2628). The total percentage of outliers for this task was 1.51% (73/4825). Outliers were identified in the same way as in the picture naming task.

Results

Reaction Times

The same transformation and fitting procedure were used as for the previous tasks ($W = .90$, $p < .001$). The final model consisted of the fixed factors Language group, Place, and Number of syllables. Interactions of these fixed factors were included in the model as well. Place and Language group were added as random slopes to item (Sound) while Place and Number of syllables were added as random slopes to subject (Subject). Models were again validated by plotting the residuals of the linear mixed effects model.

[Insert Figure 8 here]

Figure 8 shows that there was a large difference between onset and coda. This difference was significant (χ^2 (1) = 209.14, p < .001) where target phonemes placed in onset position of the auditorily presented word were reacted to faster than those in coda position. Language group (χ^2 (1) = .01, p = .92) and Number of syllables (χ^2 (1) = .03, p = .86)) were not significant. There was also no interaction between Language group and Place (χ^2 (1) = .29, p = .59). The other interactions did not reach significance either (all p-values > .1). Lsmeans was once again used to assess the importance of Language group, but no significance was reached (β = .003, *SE* = .03, *t* = .12, *p* = 91). Effect size was again estimated by means of R^2. The model explained a substantial proportion of total variance (R^2 =.626) of which Language group, however, could not explain any variance at all (R^2 = .000; 0%).

Accuracy

Figure 9 below shows the total number of incorrect responses subdivided by Language group, Place, and Number of syllables in percentages.

[Insert Figure 9 here]

A generalized linear mixed effects model with a logit function was created for the data of both monolinguals and bilinguals. The fixed factors that were included in the model were Language group, Place, and Number of syllables. Interactions of these fixed factors were included in the model as well. Place and Language group were added as random slopes to

item (Sound) whereas Place and Number of syllables were added to subject (Subject). The factor Place was highly significant (χ^2 (1) = 14.46, p < .001) in that there was higher accuracy in onset than in coda position. Number of syllables was not significant (χ^2 (1) = 1.64, p = .20). There was also no effect of Language group (χ^2 (1) = 1.22, p = .27). Finally, the interaction between Number of syllables and Place did reach significance (χ^2 (1) = 4.99, p = .03) where the difference in accuracy between monosyllabic and disyllabic picture names is larger in the onset than the coda. This pattern is confirmed by contrast comparisons between Place and Number of syllables. The difference between mono- and disyllabic picture names was significantly different in the onset (β = -2.16, *SE* = .82, *t* = -2.62, *p* = .009) but not in the coda (β = -.16, *SE* = .30, *t* = -.53, *p* = .60).

Discussion

Experiment 6 has shown that Language group did not affect phoneme monitoring in comprehension in either response latencies or accuracy scores. Place of the phoneme in the target pictures name was again highly influential in comprehension; response latencies were faster and more accurate if the phoneme was positioned in the onset. This effect has been shown to be robust as it arises in both production and comprehension. Participants also made significantly fewer mistakes in trials with a disyllabic target picture name, but only in onset trials. Do keep in mind though that only one-third of the data was made up of disyllabic words, which means that accuracy scores are based on a lower number of observations. In sum, the delay in L2 error to cut-off times cannot be easily attributed to a delay in comprehension-based monitoring.

General Discussion

The main aim of the current studies was to test whether there is an L2 disadvantage in self-monitoring for phonological errors, and if so, which component(s) of speech monitoring cause this L2 monitoring delay and whether this delay reflects a disadvantage in production or comprehension processes. Analyses of three speech-error elicitation experiments (Experiment 1, 2, and 3) provided evidence for an L2 disadvantage in phonological error monitoring. Error to cut-off intervals were longer in the L2 of Dutch-English bilinguals than in their L1, at least for interruptions within the error word. The L2 disadvantage was more than 100 ms. The same pattern of results was found when comparing L1 monolingual English speakers and L2 bilingual English speakers where the difference was around 70 ms. One account for these data patterns might be that phonological error detection happens faster in L1. Yet, these results are also in line with accounts assuming that interruption/repair processes are slower in L2, so that difficulties in resuming L2 speech cause later interruption. Three further experiments (Experiment 4, 5, and 6) aimed to disentangle these two possibilities. Experiment 4 revealed that bilingual Dutch-English speakers were slower and less accurate in naming pictures in English than monolingual English speakers; the disadvantage was more than 100 ms. Thus, there is a clear L2 disadvantage in word production. However, no main effect of Language group was found in the speed with which phoneme monitoring was performed, either in production (Experiment 5) or comprehension (Experiment 6). These findings are in line with those of Broos et al. (in press) who also found no effect of Language group during phoneme monitoring, but did find a substantial L2 delay during picture naming. Taking the results from both studies together, we argue that the L2 delay in error monitoring is caused by difficulties in planning the L2 repair.

The finding that the error to cut-off interval was longer (for the word-internal interruptions) are not in line with those of Van Hest (1996) who did not find any L2 delay for

phonological errors (she only found an L2 delay for the cut-off to repair interval in appropriateness repairs). But as mentioned, there are some important differences between the study of Van Hest and the current one. One such difference concerns the number of observations that were analysed. We had almost four times as many observations as Van Hest when calculating the error to cut-off and cut-off to repair intervals (i.e., 36 in Van Hest's study compared to 121 within bilinguals and 168 between Language groups in our study). It is likely therefore that we had a larger power to detect an effect than Van Hest. Furthermore, we distinguished between errors that were interrupted and those that were completed (see also Hartsuiker et al., 2005, 2008; Gambi et al., 2015). This distinction was not made in the analyses of Van Hest. A final difference relates to the nature of the task that was used. Van Hest used a more naturalistic tasks (i.e., story-telling task and interview task) whereas our tasks were more controlled (and hence more artificial).

Recall that Hartsuiker et al. (2008) argue that the interruption and repair of errors take place in parallel. In their study, participants were asked to name a picture that was occasionally replaced with another one. This replacement took place while participants were still naming the previous picture. In one experiment, participants were asked to name the picture that replaced the previous picture whereas participants simply stopped naming the picture in the other experiment. The picture could be either visually degraded or intact. It was found that the time between beginning naming the first picture and to stop naming it was increased when the target picture was visually degraded than when it was intact. Hartsuiker et al. (2008) therefore argued that interruption and repair are planned in parallel (see also Gambi et al. (2015) who replicated this finding in dialogue). Moreover, they claimed that some cognitive resources are shared between repair and interruption. Given these assumptions and findings, our explanation of the L2 slow-down observed in the error to cut-off interval (but

not in the cut-off to repair interval) is that interruption is postponed when difficulties arise, which leads to a longer error to cut-off time.

The effect of Language group was evident in the picture naming task whereas no Language group effect was seen in the monitoring tasks regarding response latencies. It is important to note here that the picture naming task (where L2 speakers were slower) and the production monitoring task (where L2 speakers were not slower) presumably share the same processes of lexical retrieval and phonological encoding; in both tasks, participants need to retrieve a word form from the mental lexicon and encode it phonologically as well. Up until this moment in time, the retrieval process is identical. The phonological representation is monitored internally and compared to a standard representation. What differs after this stage is the task that has to be performed (either to name the picture or monitor for a particular phoneme). When naming the picture, the speaker also has to perform phonetic encoding, articulatory planning, and actual articulation; during phoneme monitoring this is replaced by response selection, planning, and executing a button press. Comprehension also plays a role during picture naming as the pronounced picture name can be monitored for errors auditorily. Since no differences were found between monitoring tasks but reaction times between L1 and L2 speakers did differ for the picture naming task, the slow-down during picture naming in L2 might originate from phonetic or articulatory planning and/or articulation (see also Broos et al. (in press) and Hanulová et al. (2011)). The L2 delay in the post-phonological stages is not in line with an explanation which assumes that all monitoring processes are slowed down. Note that there are also studies that argue that the L2 disadvantage during picture naming lies at earlier stages of phonological processing (e.g., lexical retrieval) (Runnqvist, Strijkers, Sadat, & Costa, 2011; Strijkers et al., 2013). Yet, the lack of response latency differences between L1 and L2 speakers during the monitoring tasks cannot be explained by assuming that lexical access is responsible for the L2 disadvantage.

It could be argued that task difficulty might also have affected response latencies in both the naming and phoneme monitoring tasks. In particular, the trials in the naming task did not contain visual cues (i.e., overlapping target phonemes) that are sometimes present in the monitoring tasks, which may have made the phoneme task relatively easy, at least for the trials where yes-answers are correct. If there were a larger difference between the L1 and L2 speakers in the (presumably difficult) no-answers than in the (presumably easy) yes-answers, this would be compatible with the idea that the lack of an L2 disadvantage in phoneme monitoring is related to task difficulty. In order to test this, we conducted an additional analysis on the yes-answers and on the no-answers in Experiment 5, where most of the picture name has to be retrieved by the speaker. The interaction between Language group and Answer type (yes vs. no) was not significant while effect size of this interaction was low (see Appendix E). Hence, there is no support for the task difficulty hypothesis[6].

The effect of Place of the target phoneme in the picture name or auditorily presented word did play a vital role when considering monitoring speed. If the target phoneme was placed in onset position, both L1 and L2 speakers responded faster than when it was positioned in the coda, which is in line with the findings of Wheeldon and Levelt (1995). This indicates a regular time course of phonological encoding during the production monitoring tasks. These patterns indicate that the participants were indeed monitoring for the target phoneme. It also suggests that monitoring is a sequential process in that initial phonemes are monitored first.

One might ask whether L1 and L2 speakers monitor the picture names in the same way. In our stimuli, the target phonemes (e.g., /b/) always consistently corresponded with a letter (<b>)[5], so that, in theory, speakers could have solved the monitoring tasks by internally inspecting an orthographic code rather than a phonological code. Put differently, the participants could have detected the target by using spelling and orthographic matching rather

than phonological encoding and phonological matching. Two main hypotheses exist that relate to how spelling is conducted. On the one hand, there is the orthographic autonomy hypothesis, which assumes that spelling can be performed without phonological mediation (Rapp & Caramazza, 1997). That is to say, semantic information can be used directly to create an orthographic representation suggesting that monitoring these representations can be performed faster. Note that these representations are still likely to be monitored from onset to coda in a sequential manner. On the other hand, the obligatory phonological mediation hypothesis argues that phonological mediation must be applied in order to spell words (Geschwind, 2009; Luria, 1970). The monitoring process might therefore take longer because an extra step (phonological mediation) must be executed, which is not necessary when monitoring the orthographic representation.

The paradoxical effect in the production monitoring task (where L2 speakers tend to be faster than L1 speakers in coda position) might partially be explained by assuming that L2 speakers directly monitor orthography via semantics while L1 speakers also need to create the phonological code before orthography is monitored. But even if one assumes that L1 speakers monitor differently than L2 speakers and are therefore slower, then the L2 speakers should also be faster when the target phoneme is placed in the onset position, which is not the case. Moreover, both the direct and indirect hypothesis assume that many of the same speech production stages need to be performed (the exception being phonological encoding). It therefore seems very unlikely that L1 and L2 monitor picture names differently. Still, this leaves the interaction of Language group and Place unexplained. It was an unexpected finding, especially because no interaction effects regarding language were found in Broos et al. (in press) (where similar experiments were conducted with pictures taken from the same database). We are therefore hesitant to ascribe theoretical importance to the effect itself. What

we can claim is that L2 speakers are not *slower* than L1 speakers when it comes to phoneme monitoring.

A possible limitation to this study concerns the L2 proficiency of the participants, which was rather high, but certainly not native-like. Additionally, the mean English LexTALE score of Dutch-English bilinguals in our study is similar to the scores that Lemhöfer and Broersma (2012) found, who also tested the same type of bilinguals. Different patterns of results might have been obtained if other L2 groups were to be tested (i.e., participants who are substantially less proficient in their L2 or simultaneous bilinguals whose proficiency is even better than that of the current group). It is conceivable that less proficient bilinguals would show an L2 cost in phoneme monitoring. Furthermore, the two languages that our bilingual speakers are both Germanic. Hence, L2 speakers with an L2 from a different language family might therefore show a different pattern of results as well. That being said, there was no significant difference in proficiency between the L2 speakers that performed Study 1 and those that completed Study 2. The question of whether different patterns are seen with speakers who have different proficiency levels or have an L2 from a different language family therefore remain topics for future research.

To conclude, we have seen an L2 disadvantage during error monitoring within Dutch-English bilinguals when comparing their L1 and L2 but also between L1 English speakers and L2 English speakers. Moreover, we found an L2 slow-down in the picture naming task for L2 English speakers when compared to English monolinguals. However, this same L2 disadvantage was *not* found in either of the monitoring tasks. The effects of Language group on picture naming and on error-to-cut-off times for phonological errors on the one hand dissociate from those of monitoring for a target phoneme in production or comprehension on the other hand. Assuming that phoneme monitoring shares a number of important processes with monitoring for phonemic errors, and based on Hartsuiker et al.'s theory that self-

interruption is postponed when repair is more difficult, we propose that the L2 disadvantage in speech interruption results from difficulty in L2 repair planning.

References

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *14*(4), 382-391. doi: https://doi.org/10.1016/S0022-5371(75)80017-X

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.40, retrieved 11 May 2018 from http://www.praat.org/

Boland, H. T., Hartsuiker, R. J., Pickering, M. J., & Postma, A. (2005). Repairing inappropriately specified utterances: revision or restart? *Psychonomic Bulletin & Review*, *12*, 472-477. doi: 10.3758/BF03193790

Broos, W. P., Duyck, W., & Hartsuiker, R. J. (2016). Verbal Self-Monitoring in the Second Language. *Language Learning*, *66*(S2), 132-154. doi: https://doi.org/10.1111/lang.12189

Broos, W. P., Duyck, W., & Hartsuiker, R. J. (in press). Are higher-level processes delayed in second language word production? Evidence from picture naming and phoneme monitoring. *Language, Cognition and Neuroscience*, 1-16. doi: https://doi.org/10.1080/23273798.2018.1457168

Cop, U., Drieghe, D., & Duyck, W. (2015). Eye Movement Patterns in Natural Reading: A Comparison of Monolingual and Bilingual Reading of a Novel. *PLOS ONE*, *10*(8), e0134008. https://doi.org/10.1371/journal.pone.0134008

De Bot, K., & Schreuder, R. (1993). Word Production and the Bilingual Lexicon. In R. Schreuder & B. Weltens (Eds.), *Studies in Bilingualism* (Vol. 6, p. 191). Amsterdam: John Benjamins Publishing Company. Retrieved from https://benjamins.com/catalog/sibil.6.10bot

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, *25*(2), 169–186. https://doi.org/10.1006/jpho.1996.0040

Gambi, C., Cop, U., & Pickering, M. J. (2015). How do speakers coordinate? Evidence for prediction in a joint word-replacement task. *Cortex*, *68*, 111-128. doi: http://dx.doi.org/10.1016/j.cortex.2014.09.009

Geschwind, N. (2009). Problems in the anatomical understanding of the aphasias. *Brain and Behavior: Research in Clinical Neuropsychology*, 107-128.

Gollan, T. H., & Silverberg, N. B. (2001). Tip-of-the-tongue states in Hebrew–English bilinguals. *Bilingualism: Language and Cognition*, *4*(01). https://doi.org/10.1017/S136672890100013X

Gollan, T. H., Montoya, R. I., Cera, C., & Sandoval, T. C. (2008). More use almost always

means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language, 58*(3), 787-814. doi: https://doi.org/10.1016/j.jml.2007.07.001

Hanulová, J., Davidson, D. J., & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Psycholinguistic and neurocognitive evidence on second language word production. *Language and Cognitive Processes, 26*(7), 902-934. doi: http://dx.doi.org/10.1080/01690965.2010.509946

Hartsuiker, R. J., Catchpole, C. M., de Jong, N. H., & Pickering, M. J. (2008). Concurrent processing of words and their replacements during speech. *Cognition, 108*(3), 601-607. doi: http://dx.doi.org/10.1016/j.cognition.2008.04.005

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to Baars et al.(1975). *Journal of Memory and Language*, *52*(1), 58-70.

Hartsuiker, R. J., & Kolk, H. H. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology*, *42*(2), 113-157. doi: http://dx.doi.org/10.1006/cogp.2000.0744

Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production?. *Acta Psychologica, 127*(2), 277-288. doi: http://dx.doi.org/10.1016/j.actpsy.2007.06.003

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R 2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics, 44*(6), 1086-1105. doi: https://doi.org/10.1080/02664763.2016.1193725

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language, 33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Lagrou, E., Hartsuiker, R. J., & Duyck, W. (2011). Knowledge of a second language influences auditory word recognition in the native language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 952.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38. doi: 10.1017/s0140525x99001776

Luria, A. R. (1970). *Traumatic aphasia: Its syndromes, psychology and treatment* (Vol. 5). Walter de Gruyter, The Hague, doi: 10.1515/9783110816297

Nooteboom, S., & Quené, H. (2008). Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language*, *58*(3), 837-861.

Nooteboom, S. G., & Quené, H. (2017). Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language, 95*, 19-35. doi: https://doi.org/10.1016/j.jml.2017.01.007

Nozari, N., Dell, G. S., & Schwartz, M. F. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology, 63*(1), 1-33. doi: 10.1016/j.cogpsych.2011.05.001

Oomen, C. C., & Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, *30*(2), 163-184. doi: 10.1023/A:1010377828778

Özdemir, R., Roelofs, A., & Levelt, W. J. (2007). Perceptual uniqueness point effects in

monitoring internal speech. *Cognition*, *105*(2), 457-465. doi: https://doi.org/10.1016/j.cognition.2006.10.006

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329-347. doi: https://doi.org/10.1017/S0140525X12001495

R Core Team. (2013). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rapp, B., & Caramazza, A. (1997). From graphemes to abstract letter shapes: levels of representation in written spelling. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(4), 1130. doi: 10.1037/0096-1523.23.4.1130

Runnqvist, E., Strijkers, K., Sadat, J., & Costa, A. (2011). On the temporal and functional origin of L2 disadvantages in speech production: A critical review. *Frontiers in Psychology*, *2*, 379. doi: https://doi.org/10.3389/fpsyg.2011.00379

Sadat, J., Martin, C. D., Alario, F. X., & Costa, A. (2012). Characterizing the Bilingual Disadvantage in Noun Phrase Production. *Journal of Psycholinguistic Research*, *41*(3), 159–179. https://doi.org/10.1007/s10936-011-9183-1

Schreuder, R., & Weltens, B. (Eds.). (1993). *The Bilingual Lexicon* (Vol. 6). Amsterdam: John Benjamins Publishing Company. Retrieved from http://www.jbe-platform.com/content/books/9789027282859

Severens, E., Lommel, S. V., Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica*, *119*(2), 159–187. https://doi.org/10.1016/j.actpsy.2005.01.002

Strijkers, K., Baus, C., Runnqvist, E., FitzPatrick, I., & Costa, A. (2013). The temporal dynamics of first versus second language production. *Brain and*

Language, 127(1), 6-11. doi: https://doi.org/10.1016/j.bandl.2013.07.008

Tydgat, I., Stevens, M., Hartsuiker, R. J., & Pickering, M. J. (2011). Deciding where to stop speaking. *Journal of Memory and Language, 64*(4), 359-380. doi: http://dx.doi.org/10.1016/j.jml.2011.02.002

Van Hest, E. (1996). *Self-repair in L1 and L2 production*. Tilburg: Tilburg University Press.

Wheeldon, L. R., & Levelt, W. J. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language, 34*(3), 311.

Figures and Tables

Table 1. Overview of LexTALE scores for every participant group for both the SLIP Experiments and the Naming/Monitoring Experiments.

Task	**Language**	**Mean LexTALE Score**
SLIP Exp 1 (Study 1)	L2 English	75.26 (SD = 9.89)
SLIP Exp 2 (Study 1)	L2 English	74.30 (SD = 11.99)
SLIP Exp 3 (Study 1)	L1 English	91.62 (SD = 6.95)
Naming/Monitoring Exp 4/5/6 (Study 2)	L2 English	76.71 (SD = 10.53)
Naming/Monitoring Exp 4/5/6 (Study 2)	L1 English	91.64 (SD = 8.71)

Table 2. Estimate reaction times of error to cut-off and cut-off to repair intervals (Standard Error) as a function of error type (interrupted (e.g., b…veam-beal) vs. completed (beam-vea…veam-beal)) and Language group (L1 monolingual English vs. L2 bilingual English and L1 bilingual Dutch vs. L2 bilingual English).

Interval (error type)	**Reaction Time (SD)**	**N**	**t**	**p**
Error to	L1 Eng: 282 (14) – L2 Eng: 346 (22)	133	2.60	.009**
cut-off (interrupted)	L1 Du: 231 (30) – L2 Eng: 346 (22)	97	3.87	.0003***
Error to	L1 Eng: 809 (78) – L2 Eng: 751 (76)	35	-.54	.59
cut-off (completed)	L1 Du: 797 (105) – L2 Eng: 751 (76)	24	-.44	.67
Cut-off to repair	L1 Eng: 112 (14) – L2 Eng: 124 (21)	133	.71	.48
(interrupted)	L1 Du: 144 (28) – L2 Eng: 124 (21)	97	-.73	.47
Cut-off to repair	L1 Eng: 136 (30) – L2 Eng: 185 (54)	35	.89	.37
(completed)	L1 Du: 181 (73) – L2 Eng: 185 (54)	24	.06	.95

Figure 1. Procedure of the picture naming task

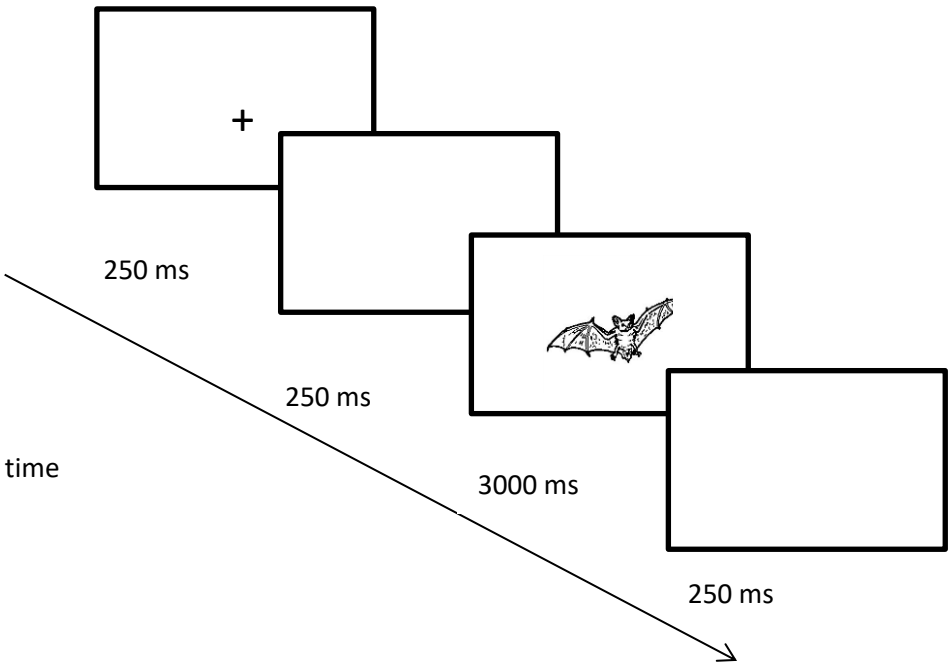


Figure 2. Observed response latencies for the picture naming task as a function of Language group (L1 English monolinguals vs. L2 English bilinguals) and Number of syllables (monosyllabic vs. disyllabic). Error bars denote standard error away from the mean (SEM).

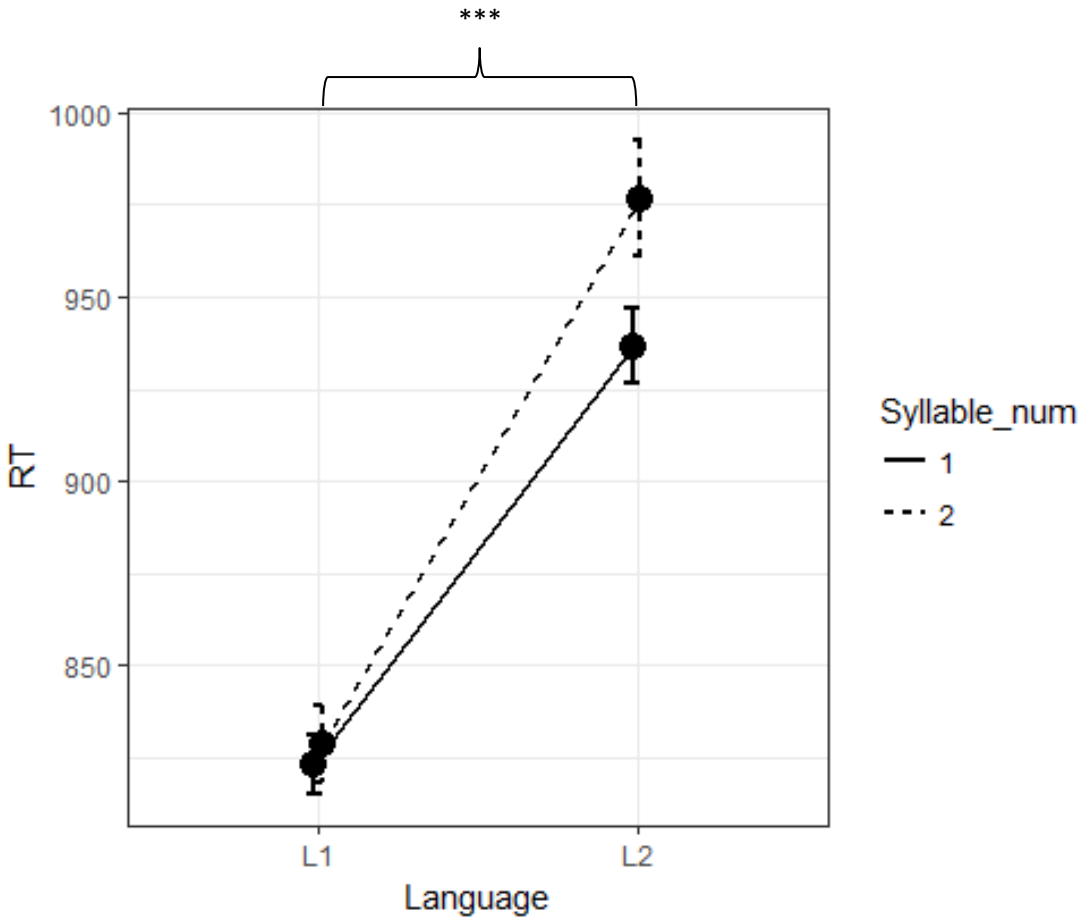


Figure 3. Accuracy as a function of Language group (L1 English monolinguals vs. L2 English bilinguals) and Number of syllables (monosyllabic vs. disyllabic) for the picture naming task. Error bars denote standard error away from the mean (SEM). Accuracy ranges from 0.00 (no correct answers = 0%) to 1.00 (all answers correct = 100%).

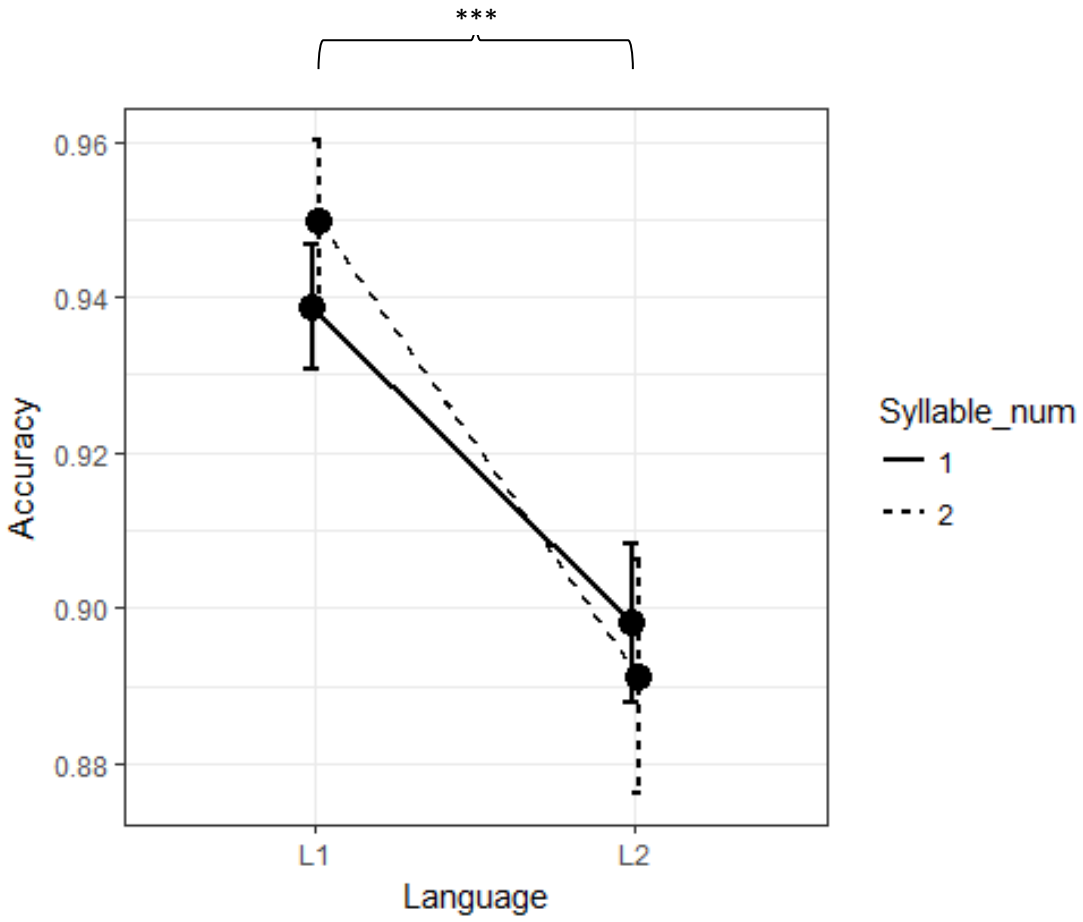


Figure 4. Procedure of the production monitoring task

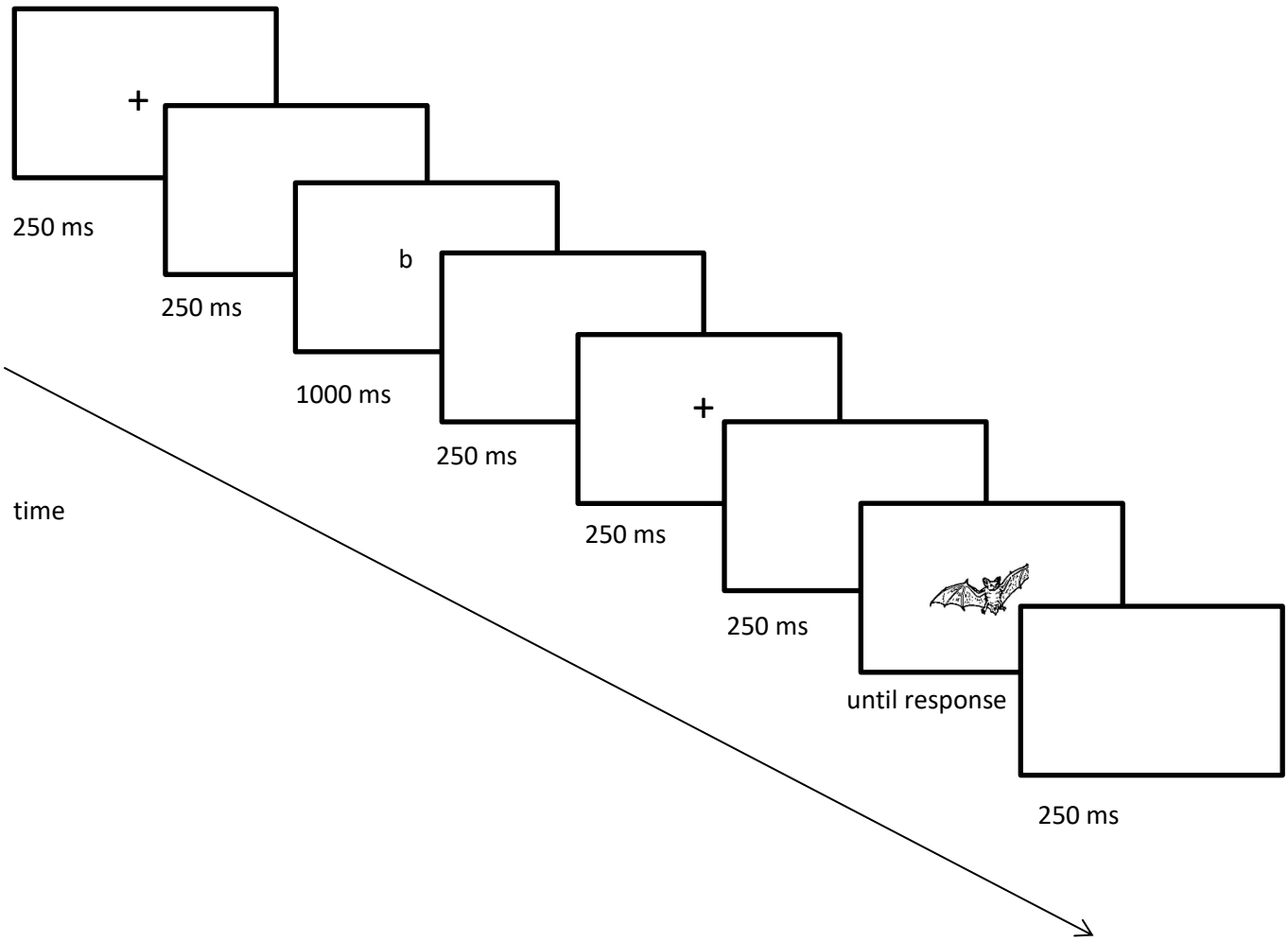


Figure 5. Observed reaction times for the production monitoring task as a function of Language group (L1 English monolinguals vs. L2 English bilinguals), Place (onset vs. coda), and Number of syllables (monosyllabic vs. disyllabic). Error bars denote standard error away from the mean (SEM).

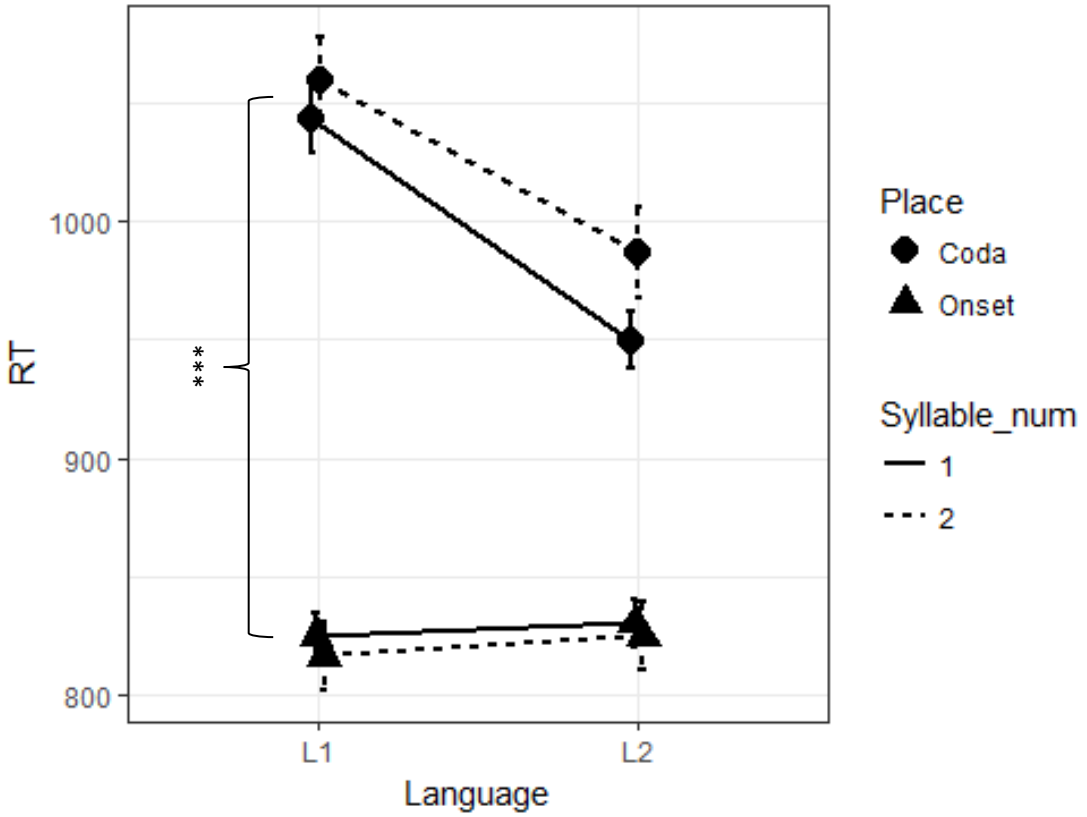


Figure 6. Accuracy as a function of Language group (L1 English monolinguals vs. L2 English bilinguals), Place (onset vs. coda), and Number of syllables (monosyllabic vs. disyllabic) for the production monitoring task. Error bars denote standard error away from the mean (SEM). Accuracy ranges from 0.00 (no correct answers = 0%) to 1.00 (all answers correct = 100%).

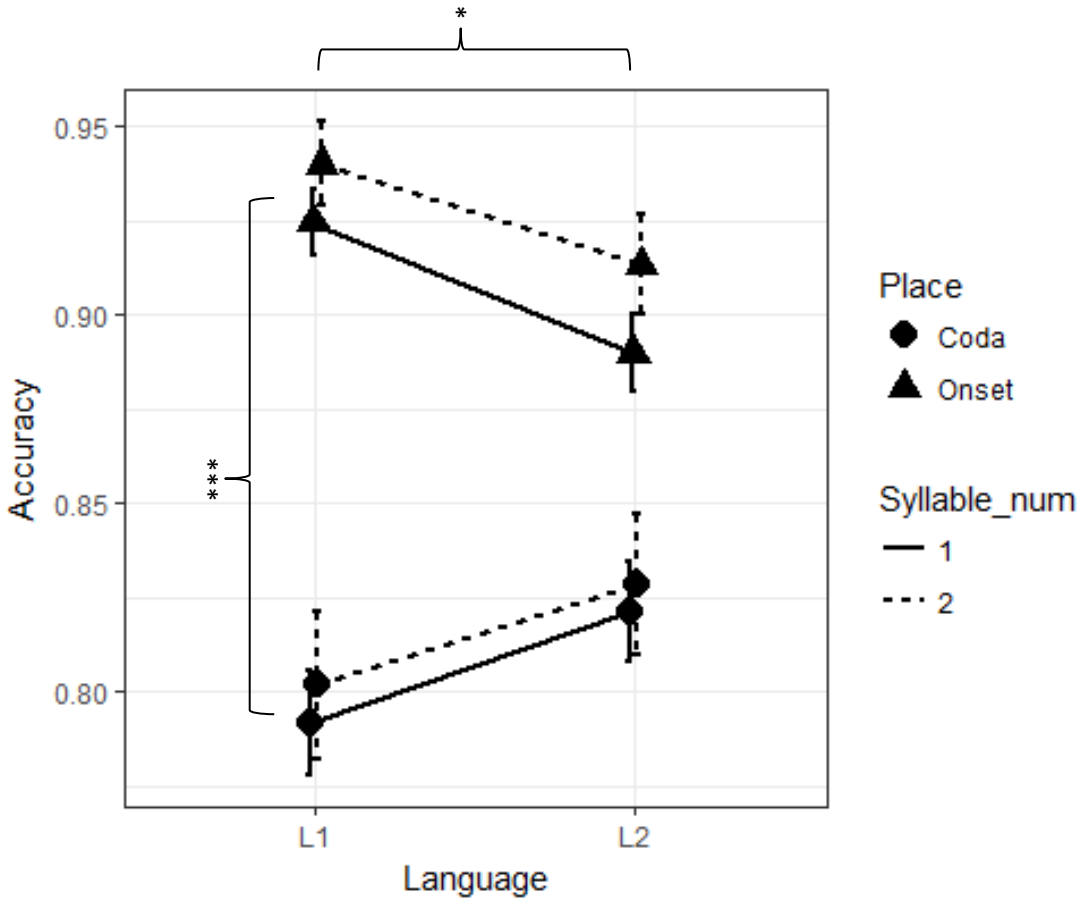


Figure 7. Procedure of the comprehension monitoring task

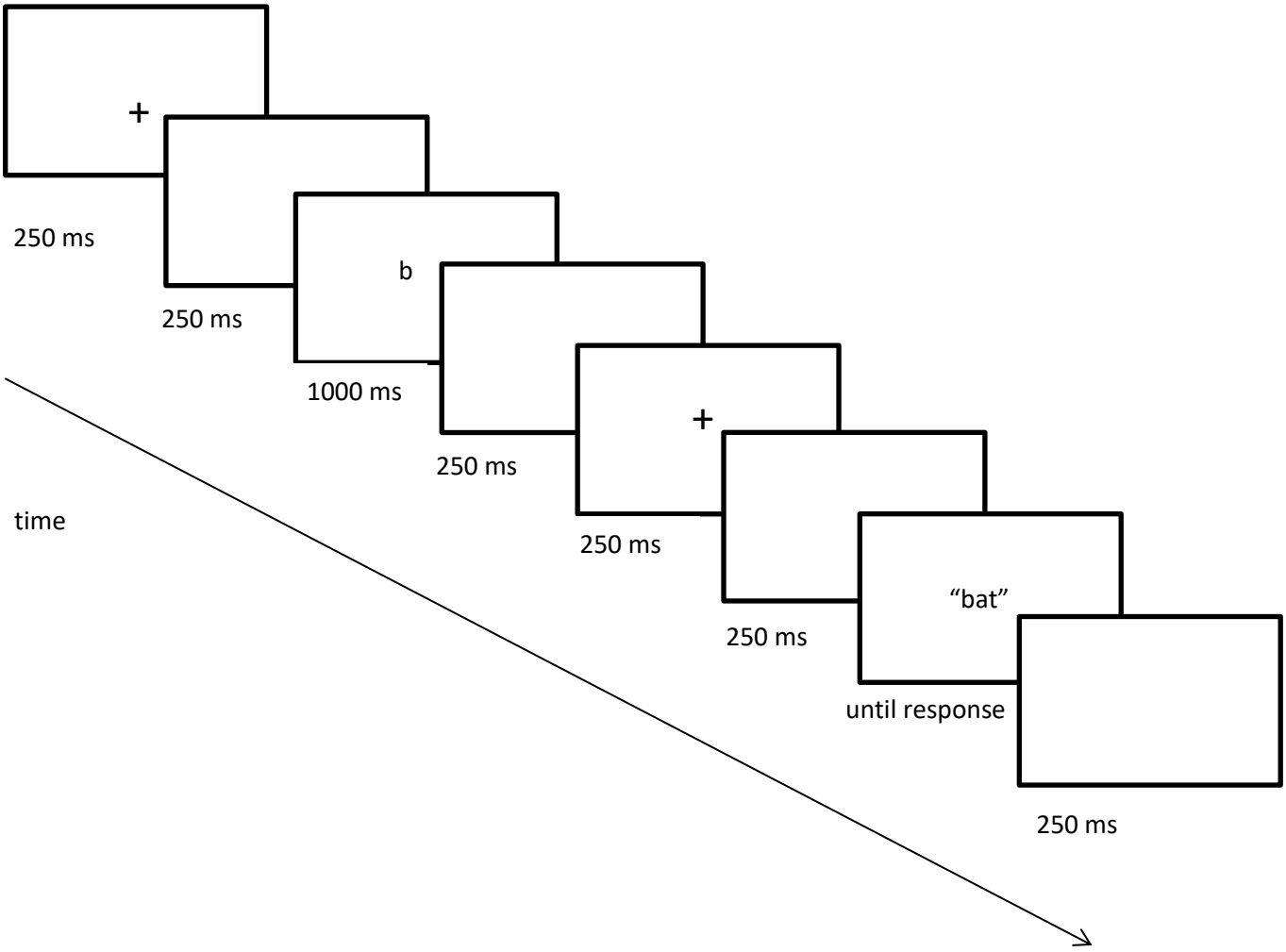


Figure 8. Observed reaction times for the comprehension monitoring task as a function of Language group (L1 English monolinguals vs. L2 English bilinguals), Place (onset vs. coda), and Number of syllables (monosyllabic vs. disyllabic). Error bars denote standard error away from the mean (SEM).

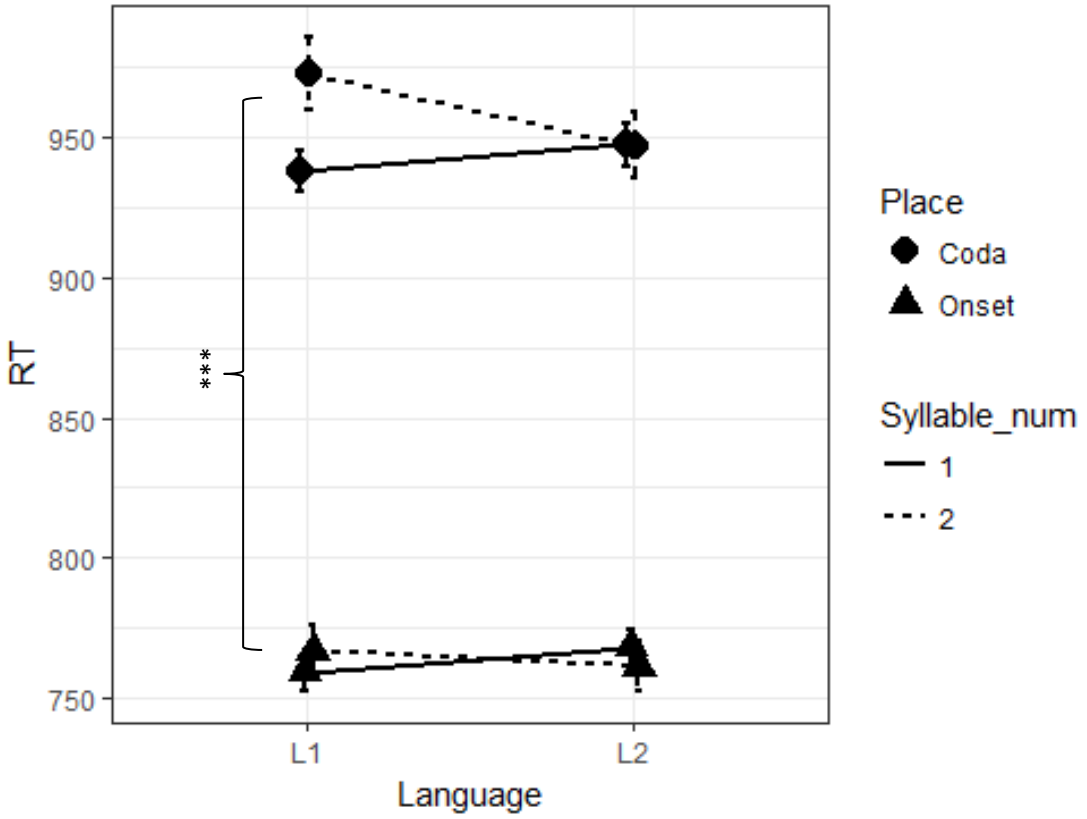


Figure 9. Accuracy as a function of Language group (L1 English monolinguals vs. L2 English bilinguals), Place (onset vs. coda), and Number of syllables (monosyllabic vs. disyllabic) for the comprehension monitoring task. Error bars denote standard error away from the mean (SEM). Accuracy ranges from 0.00 (no correct answers = 0%) to 1.00 (all answers correct = 100%).

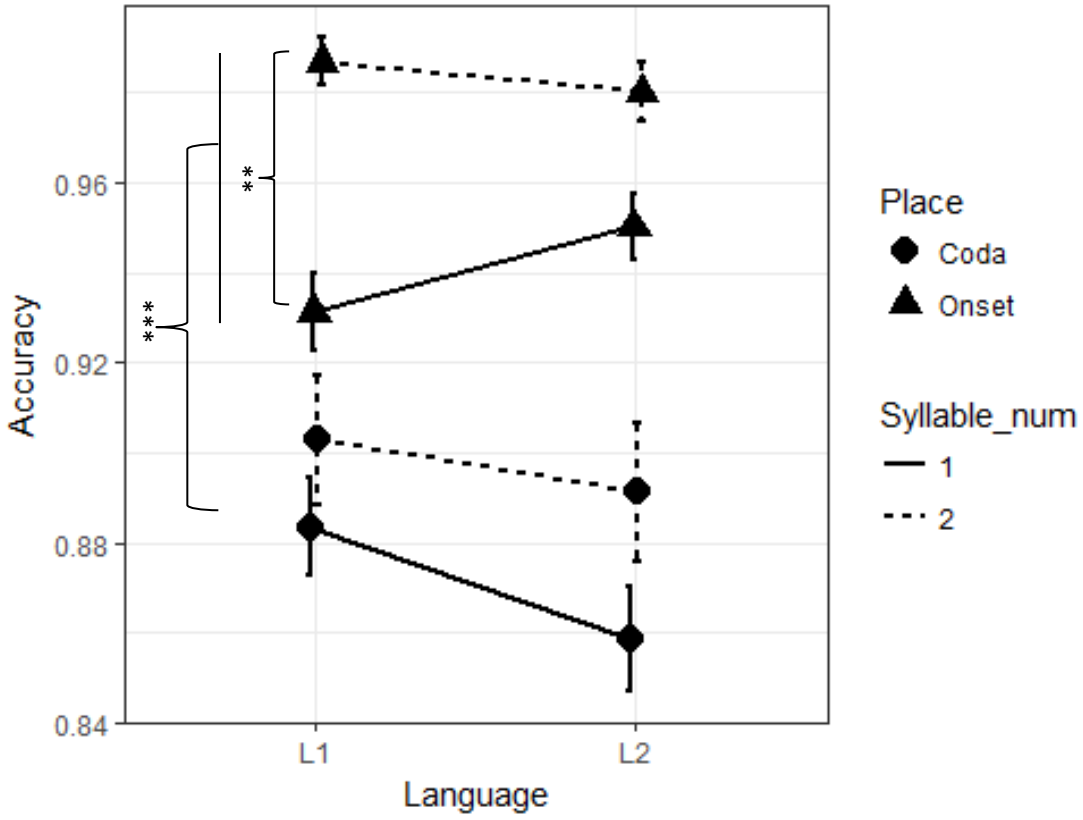
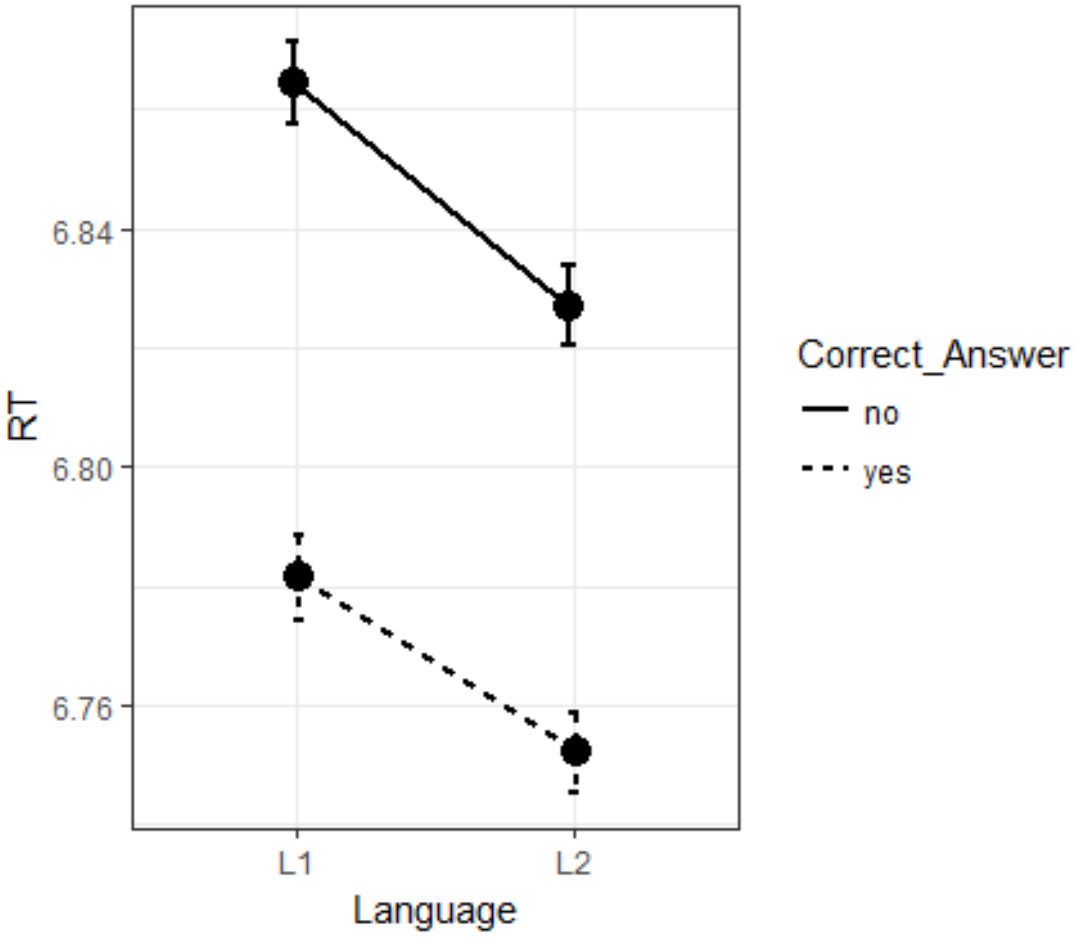


Figure 10. Observed reaction times for the production monitoring task as a function of Language group (L1 English monolinguals vs. L2 English bilinguals) and Correct Answer (yes vs. no). Error bars denote standard error away from the mean (SEM).



Footnotes

[1] These errors typically replace one utterance with a more appropriate one (e.g., 'the table…uh…the red table').

[2] Analyses that used raw RT showed almost identical patterns of results

[3] Faulty stimuli were trials where the phoneme /k/ was shown for the silent /k/ in knife, where the /p/ in pipe (both onset and coda) is present twice, and where the phonemes /t/ in rabbit and /r/ in zipper were placed at the end of the second syllable (instead of the first as in /b/ and /p/). Every faulty stimulus amounted to 18 deleted trials. Multiplied by 5, this amounts to 90 trials. This number must be doubled as they appear in both L1 and L2 data, leading to 180 out of 5400 deleted trials (≈ 3.33%). 142 out of 5400 trials (≈2.61%) were deleted for the comprehension monitoring task because the phoneme /k/ in knife was not present in the presented audio file.

[4] In the L1 data, one subject was eventually deleted since not all data was written to a file by E-prime. Because of faulty stimuli, 47 trials were analysed (out of 50). An additional subject was run but he received a different version than the subject who was deleted. Therefore, there is a difference in three trials between the L1 and L2 data.

[5] We even presented the target as a letter

[6] Additional analyses were also performed regarding repetition effects and task order. In particular, the data sets of Experiments 4, 5, and 6 were split in three smaller data sets, where each smaller data set represented the order with which the tasks were performed (e.g., the data set of Experiment 4 was split into three sets where the naming tasking was performed as first, second, or third task). However, these task order effect analyses did not yield informative results.

Appendix A

Target Items

backpack
basket
belt
bone
bowl
broom
deck
dentist
desk
dime
doll
dress
duck
dustpan
farm
file
frog
girl
glasses
gun
hammock
horse
hose
kite
knife
knight
knot
lettuce
lighthouse
lock
mirror
mitt
moose
mountain
napkin
necklace
nurse
paint
paintbrush
parrot
pencil
pillow
pipe
plate
purse
rabbit
rock
roof
rope
safe
salt
scale
scarf
sink
skirt
smoke
snail
snake
spade
spoon
suit
suitcase

tape
turkey
turtle
wall
wallet
well
wheat
wheelchair
whip
whistle
wig
window
zipper

Appendix B Stimuli SLIP Experiment 1

List 1 English	List 2 English	List 3 Dutch	List 4 Dutch
mift - gitt	hust - dunt	dalf - karm	hers - kesp
veag - beax	dift - rish	weps - venp	rors - nomp
gail - tain	duts - nuck	zits - mins	malf - zals
fath - mang	yalt - sawn	kals - hast	huts - muls
simp - rirg	coad - roat	herl - weln	gerf - vesp
lelt - beft	sich - rilk	zoch - norg	herp - weks
yant - salm	barm - fald	huks - murn	fuid - zuif
dilm - rilf	dalk - wark	rong - nolk	neug - zeut
yelt - mell	kest - jept	dont - boch	keuk - beur
sump - bung	veam - beal	beus - reul	welg - venk
hulf - dufk	mirg - gilp	hemp - kelf	darg - kafk
foft - sont	lirs - wilk	vorf - korm	voem - hoen
dufs - nush	gaif - taip	marg - zamk	ziln - mirk
nesk - dext	farl - mamc	meft - herg	kams - harn
coag - roan	yamp - marb	gelm - verp	gork - molp
yark - mard	lerf - belp	gond - mort	zols - nonf
bilf - firp	nelm - derk	fuir - zuin	mern - helg
dalp - wamf	folp - sosh	neuf - zeup	vokt - komp
lerg - jesp	yemb - merf	keut - beug	dofs - bolf
ling - wimb	surk - bulm	voek - hoeg	beuf - reup

Appendix C Stimuli SLIP Experiment 2

English non-words	English words	Dutch non-words	Dutch words
hust - dunt	duck – lump	dalf - karm	mest - berg
surk - bulm	lean - real	weps - venp	kers - hesp
duts - nuck	must – dusk	zits - mins	maf - gat
yalt - sawn	push – bull	kals - hast	zoen - doek
lerf - belp	tail – gain	herl - weln	zalf - mals
sich - rilk	tell – sent	zoch - norg	dorp - wolk
farl - mamc	math – fang	huks - murn	muts - huls
lirs - wilk	felt – left	rong - nolk	werp - heks
kest - jept	lash – back	dont - boch	duim - ruik
veam - beal	bug - mud	beus - reul	ruit - buik
mirg - gilp	seem – reef	hemp - kelf	boog - kool
dalk - wark	bag – lad	vorf - korm	zaag - haal
gaif - taip	bump – sung	marg - zamk	velg - wenk
barm - fald	pig - bill	meft - herg	kaal - maas
yamp - marb	burn – hurt	gelm - verp	raam - taal
coad - roat	wish – dig	gond - mort	veeg - leen
nelm - derk	bark – yard	fuir - zuin	hert - merk
folp - sosh	wing – limb	neuf - zeup	verf - gesp
yemb - merf	leaf – meat	keut - beug	nors - romp
dift - rish	tall – walk	voek - hoeg	deeg - ween

Appendix D Stimuli SLIP Experiment 3

List 1 English	List 2 English	List 3 English
hust - dunt	duck - lump	mift - gitt
surk - bulm	lean - real	veag - beax
duts - nuck	must - dusk	gail - tain
yalt - sawn	push - bull	fath - mang
lerf - belp	tail - gain	simp - rirg
sich - rilk	tell - sent	lelt - beft
farl - mamc	math - fang	yant - salm
lirs - wilk	felt - left	dilm - rilf
kest - jept	lash - back	yelt - mell
veam - beal	bug - mud	sump - bung
mirg - gilp	seem - reef	hulf - dufk
dalk - wark	bag - lad	foft - sont
gaif - taip	bump - sung	dufs - nush
barm - fald	pig - bill	nesk - dext
yamp - marb	burn - hurt	coag - roan
coad - roat	wish - dig	yark - mard
nelm - derk	bark - yard	bilf - firp
folp - sosh	wing - limb	dalp - wamf
yemb - merf	leaf - meat	kerg - jesp
dift - rish	tall - walk	ling - wimb

Appendix E

The linear mixed effects model that was created for the analysis on yes vs. no answers contained the fixed factors Language group and the new factor Correct Answer (yes vs. no). The interaction between Language group and Correct Answer was also added to the model.

[Insert Figure 10 around here]

Figure 10 shows that there was no interaction effect between Language group and Correct Answer (χ^2 (1) = 3.00, p = .08), but there was main effect of Language group (χ^2 (1) = 11.40, p < .001) and a main effect of Correct Answer (χ^2 (1) = 79.45, p < .001). The reason as to why Language group is now significant is that there are twice as many trials that were analyzed (not only for the yes answers, but also for the no answers). Note however that the L2 speakers were faster overall, contrary to the hypothesis of an L2 disadvantage. The lack of an interaction effect with an analysis of 9500 trials is a clear indication that the interaction is not present. Furthermore, we estimated the effect

size of the factor Language group by means of variance. The linear mixed effects model explained a total variance of .285 of which Language group could only explain .014. Hence, the presence of Language group in the model does not help in explaining the total variance; Language group is therefore considered to have a small effect size.