Delayed Picture Naming in the First and Second language

Wouter P. J. Broos[1], Alice Bencivenni[2], Wouter Duyck[1], & Robert J. Hartsuiker[1]

[1]Department of Experimental Psychology, Ghent University

[2]Università degli studi di Pavia

Author Note

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

Alice Bencivenni, Università degli studi di Pavia

Wouter Duyck, Department of Experimental Psychology, Ghent University

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

This study 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 R.J. Hartsuiker, Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2 B-9000 Ghent, Belgium, E-mail: robert.hartsuiker@ugent.be, Tel. +32 (0)9 264 64 36.

Abstract

Second language (L2) speakers produce speech more slowly than first language (L1) speakers. This may be due to a delay in lexical retrieval, but it is also possible that the delay is situated at later stages. This study used delayed picture naming to test whether late production stages (leading up to articulation) are slower in L2 than in L1. Dutch-English unbalanced bilinguals performed a regular and a delayed picture naming task in English and Dutch. Monolingual English controls performed these tasks in English. Speakers were slower when naming pictures in L2 during regular picture naming but not in delayed naming. Reaction time costs of using L2 did not vary with phonological complexity, but there was a larger L2 cost in accuracy with more complex words. We conclude that the very last stages prior to articulation are not significantly slower when bilinguals name pictures in their L2.

Keywords: second language delay, picture naming, articulation

Introduction

There are clear second language (L2) disadvantages during speech production. L2 speakers tend to make more mistakes than L1 speakers (Poulisse, 1999), are slower and less accurate at naming pictures (Gollan, Montoya, Cera, & Sandoval, 2008), and report Tip-of-the-Tongue states more frequently (Gollan & Silverberg, 2001). Bilingualism has even been found to have an effect on the native language. Gollan, Montoya, Fennema-Notestine, and Morris (2005), for instance, found that monolingual speakers were faster in naming pictures in their native language and made fewer mistakes than bilinguals who performed the task in their dominant language. Moreover, this effect was still present after the same pictures were repeated three times (also see Ivanova & Costa, 2008).

There are several theories as to why L2 speech production is slower and less accurate (see Runnqvist, Strijkers, Sadat, & Costa, 2011 for a review). A key difference between these theories is that they assume that the locus of the L2 delay is situated at either early (pre-phonological or phonological) or late (post-phonological) stages of speech production (see Figure 1). Figure 1 displays Levelt, Roelofs, and Meyer's (1999) theory of word production. It assumes that word production (for instance in the case of picture naming) involves a series of encoding steps, that activate the semantic, lexical-syntactic, and phonological make-up of the word. For our purposes we will refer to these as *early* processes. Importantly, the output of these early processes is a phonological word, which on the one hand feeds into self-monitoring processes, and on the other hand is the input for phonetic encoding. This process would turn the abstract phonological code into gestural scores, which can be seen as commands for articulation. Finally, the stage of articulation turns the phonetic gestural score into actual movements of the articulators. This stage can be divided into a number of preparatory processes that take place prior to speech onset (retrieving and unpacking the speech motor program) and articulatory processes taking place as speech motor activity begins (execution; see Rastle, Croot, Harrington, & Coltheart, 2005). For our purposes we refer to all processes taking place after encoding of the phonological word as *late processes.*

[Insert Figure 1 around here]

The weaker-links hypothesis (Gollan et al., 2008) assumes that difficulties arise at early production stages (e.g., during lexical access or phonological encoding). It argues that

bilinguals need to divide their language use among their L1 and L2, that they therefore use most words less frequently than monolinguals do, and therefore have weaker lexical representations. This hypothesis is supported, for instance, by the finding of an L1 disadvantage in bilinguals as described above. A further account, the competition for selection hypothesis (Green, 1998; Kroll, Bobb, & Wodniecka, 2006), also argues for a lexical locus of L2 delays. It claims that L1 and L2 representations compete with one another. This happens when a certain task has to be performed in two languages but also when only one language is needed. After all, there seems to be consensus in the literature that there is simultaneous activation of two languages during speech planning (e.g., Colomé, 2001; Monti, Osherson, Martinez, & Parsons, 2007).

However, a number of alternative theories on the L2 disadvantage assume that the slow-down is situated at late stages (e.g., phonetic encoding and motor processes initiating articulation) (Guo & Peng, 2007; Hanulová, Davidson, & Indefrey, 2011; Indefrey & Levelt, 2004). That such later stages can be challenging for L2 speakers is clear from the persistent foreign accent that even very proficient L2 speakers find difficult to shake off. Hanulová, Davidson, and Indefrey (2008) performed an ERP experiment in which Dutch-English unbalanced bilinguals performed a monitoring task and a delayed picture naming task in a go/no-go paradigm. Participants were asked to press a button (or refrain from pressing one) depending on whether a depicted object was manmade or natural or whether it started with a particular phoneme. Hence, both a semantic and phonological N200 (which indicates response inhibition) can be measured in both L1 and L2. The semantic N200 occurred before the phonological one in both languages but there was no time difference between L1 and L2 regarding the time that both N200 components arose. That is to say, there was no language effect on semantic and phonological N200 intervals, which suggests no language difference in early stages.

The theory of a late L2 disadvantage is consistent with results from picture naming studies in Spanish monolinguals and Spanish-Catalan bilinguals (i.e., studies testing the effect of bilingualism on the L1). Sadat, Martin, Alario, and Costa (2012) observed that bilinguals were slower to produce bare nouns and noun phrases than monolinguals, even though both groups were using an L1. Importantly, there was also a bilingual cost in word durations, suggesting a bilingual disadvantage in the late production stage of articulation (although it is possible that problems at earlier stages percolate to these later levels, Runnqvist et al., 2011). Sadat, Martin, Magnuson, Alario, and Costa (2016) conducted a large-scale picture naming study, again with Spanish monolinguals and Spanish-Catalan bilinguals, and observed that

this bilingual disadvantage could not be attributed to lexical frequency. However, phonological similarity between the translation equivalents in the two languages strongly affected the bilingual disadvantage in naming times: with large phonological overlap, the bilingual disadvantage disappeared. The authors interpreted this as evidence that bilinguals' costs "should emerge at rather late stages of language processing" (p. 1929).

Other studies that support an explanation of the L2 advantage in terms of late stages were performed by Broos, Duyck, and Hartsuiker (2018; see also Broos, Duyck, & Hartsuiker, 2019). Broos et al. (2018) used a picture naming task and a phoneme monitoring task combined with a picture-word interference paradigm. The picture naming task was included to verify whether Dutch-English bilinguals were indeed slower than English monolinguals when naming the pictures in English. During the phoneme monitoring task, both participant groups were asked to press a button if a particular phoneme was present in an English picture name. This monitoring task arguably involves lexical retrieval and phonological encoding, but not phonetic encoding or actual articulation. In both tasks, we presented distractor words, which the participants had to ignore (picture-word interference). If the tasks tap into regular speech production, we expect an effect of phonological relatedness between picture name and distractor word. Specifically, the distractor words overlapped phonologically with the English picture name in onset and/or coda (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 shortened response latencies, demonstrating that this task indeed taps into regular word form retrieval (see also Wheeldon & Levelt, 1995). Importantly, bilingual speakers were slower to name pictures in English than monolingual speakers but there was no such L2 disadvantage in the phoneme monitoring task. The L2 delay in picture naming and lack of such a delay in phoneme monitoring were recently replicated by Broos et al. (2019) in a paradigm without distractor words. As early stages of speech production are completed in both picture naming and phoneme monitoring, Broos et al. argued that the L2 slow-down must be situated at later stages. However, it remains unclear whether these stages affect phonetic encoding or even later stages leading up to articulation.

The current study will use a regular and delayed picture naming task to answer the question of whether the slow-down in L2 speech production originates from (preparatory) articulatory processes taking place after phonetic encoding. According to Rastle et al. (2005), the delayed naming task taps into processes that take place after the speech motor plan is compiled (also see Kawamoto, Liu, Mura, and Sanchez, 2008). Following Rastle et al., we will refer to these processes as "post-planning articulatory operations." This study uses the

delayed naming task to focus on the articulation stage. The regular naming task will also be used to verify the L2 slow-down that previous studies have shown as well (see Gollan et al., 2008).

Monolingual English speakers and bilingual Dutch-English speakers performed the regular and delayed picture naming task in their L1 and L2. If participants are slower in naming pictures in English than in Dutch in the delayed task, then slower post-planning articulatory operations of picture names in their L2 is the only explanation for the L2 disadvantage. This would support an account that assumes the locus of the slow-down is situated at the very latest stages of speech production. However, if there is no difference between the delayed task in L1 and L2, then such post-planning articulatory operations cannot be responsible for the slow-down. Taking the findings of Broos et al. (2018) into consideration, the latter finding would suggest that the delay is still situated at a late stage, but not at the very last one (i.e., post-planning articulatory operations). It would rather suggest that the stage of phonetic planning would be responsible for the L2 delay. An additional goal was to see whether phonological complexity of the onset and coda of the picture names would influence the response latencies in either the regular or the delayed picture naming task (phonologically simple: '**l**eg' vs. phonologically complex: '**st**ool'). We suspected that L2 costs at late processing stages would be limited in the case of phonologically simple syllables, whereas more considerable costs would surface with phonologically complex syllables, as the latter type of syllables might more heavily tax motor planning and post-planning articulatory operations.

Method

Participants

Forty monolingual English speakers (10 male and 30 female) and 43 (7 male and 36 female) bilingual Dutch-English speakers participated in the experiment and were recruited at the University of Leeds and Ghent University, respectively. Participants all reported to have normal hearing, normal to corrected-to-normal sight, and not to have dyslexia. All participants performed an adapted version of the MINT test (Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2012), a picture naming task that measures English proficiency. This version avoided English stimuli that overlapped in form and meaning (i.e., were cognates) with their Dutch or French translation equivalents. We only used an English version of the MINT. There was no

overlap in pictures between the MINT test and the stimuli used in the experiment. The monolingual English speakers scored a total mean average of 48.85/52 (= 93.9%) whereas the bilingual Dutch-English speakers obtained a total mean score of 30.65/52 (= 58.9%). The difference between the scores of the mono- and bilingual speakers was significant ($t(44.03) = 8.82$, $p < .001$).

Flemish university students, like the Dutch-English speakers tested here, are typically clearly dominant in Dutch and know English and French as second languages. They typically follow formal instruction in English in secondary school from age 12-13 until graduation (French instruction already starts at age 10-11 in primary school). However, most students are exposed to a considerable amount of English on a daily basis, starting already in childhood: they pick up English from television and films (which are usually subtitled in English), social media, games, the internet, and so on. As students, they are of course still exposed to these media, and will additionally use English materials for study. Despite their earlier start with French, most students reach a much higher proficiency in English. Our sample corresponded well with this profile: the Flemish students had a mean age of 19 years (range 17 - 28) and started learning English at age 10.8 years on average (range 1 - 14). Twenty-three participants reported that they first started learning English at school and 20 participants first started learning English elsewhere (at home, internet, television). On a Likert scale from 1-7, they rated Dutch speaking skill as 6.35 (SD=.72) and overall Dutch proficiency as 6.21 (SD=.56); English speaking skill was rated as 4.93 (SD=.96) and overall English proficiency as 5.04 (SD = .87). The difference between Dutch and English self-rated speaking skill was significant ($t(40) = 10.14$, $p < .001$) and so was the difference between overall Dutch and English self-rated proficiency ($t(40) = 9.88$, $p < .001$).

The student population in Leeds is linguistically diverse. We imposed the restriction however that our Leeds participants were monolingual speakers of English. We did not present a demographic questionnaire to our sample in Leeds.

Materials

Twenty-five target pictures with monosyllabic names were included in the stimulus list. They were taken from the set of pictures normed for Dutch by Severens, Van Lommel, Ratinckx, and Hartsuiker (2005). The pictures were black-and-white line drawings of simple objects. The translation equivalents matched in phonological complexity and all target picture names

were monosyllabic (e.g., 'ca**st** – gi**ps**')[1]. Of these 25 target picture names, 13 picture names had a simple phonological construction without consonant clusters (e.g., 'leg') while twelve picture names had a complex construction at either the onset (e.g., '**st**ool') or coda (e.g., 'ca**st**'). The target picture names were relatively high frequent (4.26 out of 7 on the Zipf-scale, see Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Frequency values were taken from the Dutch and English SUBTLEX database (Van Heuven et al. (2014) and Keuleers, Brysbaert, & New (2010)). Other lexical covariates that were measured were length of the target picture names in number of characters (which always varied between three and five), visual complexity of the picture (based on jpeg file size), and mean orthographic Levenshtein distance. Nineteen fillers were added to the stimulus list so that a total of 44 pictures were presented. Monolingual English speakers saw all 44 pictures twice (once in a regular naming block and once in a delayed naming block) whereas bilinguals saw the pictures four times as they also performed both tasks in Dutch. All blocks were counterbalanced, leading to a total number of two versions of the experiment for the monolingual English group and 24 versions for the bilingual group. A list of target stimuli and their corresponding Zipf and mean orthographic Levenshtein values are presented in Appendix A.

Procedure

Participants were seated in front of a computer screen in a quiet room. Before the experiment began, participants were asked to perform the MINT test to measure their English proficiency. Next, the picture naming and delayed picture naming tasks were explained. During the regular picture naming task, a fixation cross was presented on the screen for 700 ms after which the picture appeared for 3000 ms. After the picture disappeared from the screen, a blank screen was presented for 500 ms after which the next trial began. Participants were asked to name the picture as fast and accurately as possible as soon as it appeared on the screen. The delayed picture naming task was almost identical, except for the cue that appeared 1250 ms after the picture was presented on the screen. This time the picture remained on the screen for 2000 ms after the cue was presented. Now, participants were asked to name the picture as soon as they saw the exclamation mark on the screen (see Figure 2). The experiment started with a two-blocked practice phase where five regular picture naming trials and five delayed trials were presented. The pictures used in the practice trials did not overlap with the ones used in the rest of the experiment. Monolingual speakers were presented with a two-blocked experimental

phase (regular and delayed) whereas bilingual speakers saw four blocks (regular and delayed in each language). All blocks were counterbalanced.

[Insert Figure 2 around here]

Data analysis

Two data sets were created: one data set that combined the data of L1 speakers and L2 trials of L2 speakers (between-subjects data set) and another data set that combined L1 and L2 trials of L2 speakers (within-subjects data set). Before the data sets were analysed, incorrect trials were removed first (1584/4150 trials for between subjects and 1937/4300 for within subjects; see Appendix B for a split-out per condition). There are several reasons for this high amount of data loss but the most influential one is the lack of a familiarisation phase, which was not included to avoid facilitation effects. Incorrect trials were considered trials where the wrong picture name was used, an article was put in front of the picture name, or when the trial was not fluently pronounced. The monolingual English speakers answered 430 trials incorrectly whereas bilinguals answered 1154 trials incorrectly in their L2 (English) and made 783 errors in their L1 (Dutch). Additionally, trials where the response was uttered too early (before the cue appeared in the delayed task, which almost never occurred) were deleted as well. A second exclusion criterion was put into place, namely that the corresponding trials of target pictures that were answered less than 30% correctly were removed from the data set. Mean accuracy per language group, per task was used to determine which target pictures fell below the accuracy threshold. In the current study, the total number of deleted target pictures amounted to three out of 25 target pictures for the between-subjects data set and five out of 25 target pictures for the within-subjects data set. Most of these trials were already removed by the first removal procedure, but the remaining correctly answered trials of the < 30% accuracy target pictures were removed as well (92/2566 trials for the between-subjects data set and 143/2363 for the within-subjects data set). Finally, extremely fast trials (< 100 ms) were also removed from the data sets. The number of deleted trials according to this third exclusion criteria were 27/2474 trials for the between-subjects data set and 10/2220 for the within-subjects data set. Despite the high number of removed trials, we argue that we have more than enough data to make claims about response latencies and language differences. The L2 delay

during picture naming is a very large effect as several studies have shown language differences of around 100 ms (Kroll & Stewart, 1994; Broos et al., 2018; 2019).

Response latencies were manually determined with the computer program Praat (Boersma & Weenink, 2017). In case of regular picture naming trials, response latencies were measured from picture onset. For delayed picture naming trials, however, response latencies were measured from the onset of the naming cue (i.e., the exclamation mark). For both kind of trials, a recording started at the moment the critical visual stimulus (picture or exclamation mark) was presented (also see Broos et al., 2019).

The data set was analysed by means of linear mixed effects models with the lme4 (version 1.1-14), car (version 2.1-5), lsmeans (version 2.27-2), and lmerTest (version 2.0-33) package of R (version 3.4.1) (R Core Team, 2013). This allowed for inclusion of both subject and item as random factors (Baayen, Davidson, & Bates, 2008). The first step of the analysis was to create a linear mixed effects model with a maximal random effects structure (see below) that did not include lexical covariates (i.e., lexical frequency, mean orthographic Levenshtein distance, character length, and visual complexity of the picture). Next, the lexical covariates were standardized as these were all presented on different scales. Potential multi-collinearity was tested for by calculating the VIF (Variance Inflation Factor) where a value exceeding 10 is indicative of multi-collinearity (O'Brien, 2007). Finally, the lexical covariates were added to the model after which interactions with each fixed factor was tested for. Interactions were tested by means of model comparisons where we compared a model without interactions between a fixed factor and lexical covariates and a model that did interact with a fixed factor. Note that fixed factors were interacted with the lexical covariates one at a time in separate models. Likelihood ration tests were run on the optimal model in order to determine the main effects and interaction effects. All R-scripts and CSV files that were used to analyse the data can be found on Open Science Framework (https://osf.io/r7tep/).

Results

Between-subjects analysis

Reaction times

The fixed factors that were included in the model for the between-subjects data set were Language Group (L1 speakers vs. L2 speakers of English), Task (Delayed vs. Regular picture

naming), and Complexity (Simple vs. Complex consonant clusters). Interactions of all fixed factors were included in the model. Trial and Proficiency were added to the model as co-variates. The lexical covariates lexical frequency, mean orthographic Levenshtein distance, character length, and visual complexity of the picture were included as well. No interactions between lexical covariates and fixed factors were added. Random slopes were determined based on the maximal random effects approach (Barr, Levy, Scheepers, and Tily (2013)) meaning that all fixed factors and their interactions were added as random slopes. Random intercepts for both subject (sbjID) and item (ItemID) were included. Language Group (and its interaction with Task) could only be added to the item random intercept as this was a between-subject variable whereas Complexity could only be added to the subject random intercept since this was a between-item variable. Hence, the random slope for subject was the interaction of Language Group and Task whereas the random slopes for item was the interaction of Task and Complexity. The VIF values of all factors and interactions fell below 5, which indicates that no multi-collinearity issues arose.

[Insert Figure 3 around here]

Figure 3 shows that there is a clear interaction between Language Group and Task (χ^2 (1) = 26.04, $p < .001$): descriptively, there was an L2 cost in the regular naming task but not in the delayed naming task. The main effect of Language Group also reached significance (χ^2 (1) = 17.20, $p < .001$): overall, L2 speakers of English were slower than L1 speakers of English. There was also a main effect of Task (χ^2 (1) = 437.44, $p < .001$) indicating that the delayed naming trials were reacted to faster than regular naming trials. This might seem somewhat counterintuitive but recall that response latencies were logged when the cue appeared on the screen in the delayed task (when participants already retrieved the lexical representation) whereas reaction times were measured as soon as the picture appeared on the screen in the regular task. Complexity also reached significance (χ^2 (1) = 5.50, $p = .02$): phonologically simple words were reacted to faster than complex ones. There was also an interaction between Complexity and Task (χ^2 (1) = 4.34, $p = .04$): the complexity effect was larger in the delayed task than the regular naming task. The only lexical covariate that reached significance was visual complexity of the picture (χ^2 (1) = 4.09, $p = .004$): more complex pictures were reacted to more slowly. No other interactions or lexical co-variates were significant (all p-values > .1).

The package lsmeans was used to determine which contrasts were significant and which ones were not. The contrast L1 Regular Task vs. L2 Regular Task was significant (β = -0.23, *SE* = 0.05, t = -4.61, $p < .001$) where L2 was slower. However, the contrast L1 Delayed Task vs. L2 Delayed Task did not reach significance (β = 0.06, *SE* = 0.07, t = 0.85, p = .40).

Accuracy

The model without lexical covariates did not converge when the maximal random effects structure was inserted. We therefore followed the backward fitting procedure where we first ran the model without random correlations. If that model did not converge, we recursively removed the random slopes that explained the least variance. The final model contained the fixed factors Language Group, Task, and Complexity. Interactions of all fixed factors were included in the model. Complexity was added as random slope for subjects and Language Group for items. The co-variates Trial and Proficiency were added to the model as well. No lexical co-variates were added as this prevented the model from converging. All VIF values fell below 7 and therefore did not exceed the threshold of multi-collinearity.

[Insert Figure 4 around here]

Figure 4 above reveals that L2 speakers are less accurate in both the regular and delayed picture naming task compared to L1 speakers. This is confirmed by the main effect of Language Group (χ^2 (1) = 14.38, p < .001). A main effect of Task was also observed (χ^2 (1) = 16.38, p < .001) where participants were more accurate in the delayed task than the regular task. Proficiency also reached significance (χ^2 (1) = 37.92, p < .001) indicating that proficient speakers made fewer errors than less proficient ones. Finally, there was an interaction between Language Group and Complexity (χ^2 (1) = 9.73, p = .002) indicating that the difference between L1 and L2 was larger for picture names with consonant clusters.

Contrasts showed that L2 was more error prone in both naming tasks (L1 Regular Task vs. L2 Regular Task: β = 1.66, *SE* = 0.39, z = 4.30, $p < .001$; L1 Delayed Task vs. L2 Delayed Task: β = 1.44, *SE* = 0.39, z = 3.70, $p < .001$).

Within-subjects analysis

Reaction times

The model without lexical covariates did not converge when the maximal random effects structure of the random slopes was inserted. Following the backward fitting procedure, the final model turns out to be very similar to the model of the between-subjects analysis. The only differences are that density is residualized (as VIF for density was 10.99), that an interaction of Complexity and the lexical co-variates is added, and that the random slopes for subject are Task*Language instead of Task*Complexity.

[Insert Figure 5 around here]

Figure 5 indicates that the interaction between Language and Task did not reach significance within bilinguals (χ^2 (1) = 1.95, p = .16), in contrast to the clear interaction observed in the between-subjects analysis. It seems that particularly, the L2 cost in the regular naming task is smaller in the within- than the between-subjects analysis. Yet, there was a main effect of Language (χ^2 (1) = 8.16, p = .004), demonstrating an L2 cost overall. Task also reached significance (χ^2 (1) = 489.00, p < .001): participants were slower in the regular naming task. The factors Trial and Proficiency did not reach significance, nor did any of the lexical covariates (all p-values > .1). There were, however, significant interactions between Complexity and frequency (χ^2 (1) = 9.63, p = .002) as well as Complexity and visual picture complexity (χ^2 (1) = 6.30, p = .01). The former interaction denotes that response latencies for complex pictures go up if frequency goes up as well whereas the latter implies that the difference in reaction times between complex and simple pictures is smaller in more visually complex pictures.

Contrast comparisons were performed for completeness sake and because numerically there is a larger L1 and L2 difference in the regular naming task. The difference between Dutch and English was significant in the regular picture naming task (β = -0.14, *SE* = 0.04, z = -3.31, p = .003) but not in the delayed task (β = -0.08, *SE* = 0.05, z = -1.62, p = .11).

Accuracy

The generalized linear mixed effects model without lexical covariates did not converge when the maximal random effects structure was applied. We therefore followed the backward fitting procedure (also see previous analysis for accuracy). Lexical covariates could not be included in the model because of convergence errors. The final model contained the fixed factors

Language, Task, and Complexity. Interactions of all fixed factors were included in the model. Trial and Proficiency were added as co-variates. Language was added as random slope to both subject (sbjID) and item (itemID). All VIF values fell below 6 meaning that no multicollinearity was observed.

[Insert Figure 6 around here]

Figure 6 shows that Dutch trials were answered more accurately than English trials (χ^2 (1) = 10.17, p = .001). Task was also significant (χ^2 (1) = 32.42, p < .001) where fewer mistakes were made in the delayed picture naming task than the regular task. There was also a main effect of Proficiency (χ^2 (1) = 7.13, p = .008) where accuracy was higher in participants with a higher proficiency score. No other main effects or interaction effects were significant (all p-values >.1). Contrast comparisons confirm that the Language effect is significant in both the regular and delayed picture naming task (delayed: β = 1.24, *SE* = 0.39, z = 3.20, p = .001; regular: β = 1.22, *SE* = 0.39, z = 3.16, p = .002).

Discussion

The current study aimed to answer the question of whether the L2 disadvantage in picture naming was caused by a delay in the very last stages of picture naming. Analyses compared L1 English of monolingual speakers and L2 English of bilingual speakers (between-subject analysis) as well as L1 Dutch and L2 English of bilingual speakers (within-subject analysis). Contrast comparisons in both types of analyses demonstrated that there was an L2 disadvantage in the regular picture naming task. However, no significant differences between L1 and L2 were found in response latencies in the delayed picture naming task. That is to say, post-planning articulatory operations do not appear to be slower in L2 compared to L1. Task was always significant because speech can be fully planned in the delayed naming task whereas this is not the case in the regular naming task. When the cue is detected, all that has to be done is to initiate and execute a pre-planned motor program for the picture name. Most importantly, the interaction between Language (Group) and Task reached significance in the between-subjects analyses which further confirms that L2 disadvantages are only found in the regular picture naming task when comparing L1 and L2 speakers. Accuracy scores revealed that L2 trials were also reacted to less accurately than L1 trials and this was true in both the regular and delayed naming task. More proficient speakers made fewer mistakes. Delayed

picture naming trials, however, were answered correctly significantly more often than regular picture naming trials. The reason for this effect is most likely due to the prolonged period of time that participants have to think about the picture name in the delayed task. Finally, phonological complexity interacted with language group: there was a larger L2 cost with complex words. However, this interaction was not observed in the within-subjects analysis.

The L2 reaction time disadvantage in the regular picture naming task is consistent with the findings of many studies (e.g., Gollan et al., 2008) that found L2 disadvantages during picture naming tasks. The lack of such an L2 disadvantage in the delayed picture naming task suggests that the delay is not situated during post-planning articulatory operations. Recall that the studies of Broos et al. (2018; 2019) suggested that early processes of speech production are not slowed down in L2. As the slow-down is not seen in earlier stages nor in the very late stage of speech production that delayed named taps into, it seems likely that the L2 disadvantage is situated at the phonetic encoding stage of speech production. This is compatible with a number of findings in the literature, including the lack of a difference in time course between the semantic and phonological N200s in L1 and L2 (Hanulová et al., 2011) and the finding that phonological similarity nullifies the bilingual naming cost in picture naming (Sadat et al., 2016). At the same time, it is important to acknowledge that there is empirical support for alternative accounts such as the weaker links and competition account as well. For instance, ERP studies by Strijkers, Costa, & Thierry (2010) and Strijkers, Baus, Runnqvist, Fitzpatrick, and Costa (2013) demonstrated that wave forms elicited in picture naming in L1 and L2 diverge already after about 200 ms after picture onset (a time window in which the ERPs also pick up frequency and cognate effects). Both the early timing of the effect and the fact that it coincides with lexical effects is support for an early account of the L2 cost. It should further be noted that these accounts are not mutually exclusive. It is possible, for instance, that "weaker links" are not restricted to lexical or phonological representations only, but also affect representations downstream such as syllables and motor programs (Runnqvist et al., 2011).

Response latencies of the current study show that the difference between regular and delayed picture naming amounts to 200 ms in the English monolingual group. Note that the response latencies of the regular naming task are measured from the earliest possible stage of speech production whereas these are measured just before articulation in the delayed task. This suggests that the speech production stages up until articulation are completed within 200 ms and that the largest part of the response latencies of regular picture naming must be made up of post-phonological stages. Yet, the bilinguals show a difference of 400 ms between the

regular and delayed naming task. This larger difference might reflect the effect of bilingualism itself on response latencies of regular picture naming (see also Ivanova and Costa (2008) who showed that bilinguals name pictures slower in their L1 than monolinguals). That being said, the L1 of the monolinguals and bilinguals is, of course, not the same language, meaning that language itself might also be responsible for this difference.

It must be noted that the interaction between Language (Group) and Task was significant in the between-subjects analyses but not in the within-subjects analyses, whereas contrast comparisons revealed an L2 delay in the regular naming in both types of analyses. A possible explanation for this observation is that the languages that are compared are different when considering the within-subjects analysis (Dutch vs. English) but is identical in the between-subjects analysis (English). This also means that the variability in the stimuli for the within-subjects analysis increases due to translation equivalents. These differences might have caused the lack of an interaction effect in the within-subjects analysis.

One might ask whether phonetic encoding is indeed completed before the cue appeared on the screen. If it is true that these processes are slower in the L2, then some participants might not have had the chance to form their phonetic plan or to set their articulators in the appropriate position. Kawamoto et al. (2008) used the delayed picture naming task and varied the delay period (150, 300, 450, 600, and 750 ms). One of the aspects that was tested and was shown to affect preparation time was the type of consonant that was placed at the onset of a word. Specifically, they examined the acoustic latencies of plosives and non-plosives across the different delay periods. The difference between the two consonant types was significant at 150 ms but non-existent at 750 ms. This indicates that different phonemes have different preparation times but that these differences disappear after a certain amount of time. Potential differences between L1 and L2 might therefore also dissolve, keeping in mind that both native and non-native phonemes are produced (under the assumption that the delay period is sufficiently long). The delay period of the current experiment was 1250 ms, which suggests that non-native phonemes will most likely be fully retrieved as well. Furthermore, response latencies of a regular picture naming task (Broos et al., 2018) were measured prior to constructing the current experiment. Only two participants out of 54 showed a mean response latency that surpassed 1250 ms when naming pictures in their L2. Therefore, it is safe to assume that participants finished phonetic encoding in both L1 and L2 before the end of the delay period.

The bilinguals in this study used two languages (Dutch and English) that are related (they are Germanic languages, they have many cognates). However, there are also many

phonetic and phonological differences between these two languages: speaking English is clearly a challenge for native Dutch speakers' speech motor systems. It is conceivable that any late L2 cost is much reduced for speakers of language pairs that are more similar at the sound level, but enhanced for speakers of more distant language pairs. On the other hand, stronger similarity may also hinder production. For instance, Acheson, Ganushchak, Christoffels, and Hagoort (2012) concluded that form overlap in the names for pictures in bilinguals' two language led to stronger response conflict, as indexed by an event-related potential. This is clearly an issue that further research needs to address.

The current study focused on the production of single words in a picture naming paradigm, a task that, arguably, has little to do with language production under more ecologically valid circumstances. However, many L2 speakers are confronted with the need to produce language in L2 on a daily basis: this is the case, for instance, for many students in higher education. Establishing that there are L2 costs and pinpointing the locus or loci of these costs in the cognitive system is important, for instance in order to assess whether any training or support for these students should focus more on early or late production processes.

To conclude, we observed an L2 delay during regular picture naming whereas this disadvantage disappeared in the delayed naming task and was significant in both between- and within-subjects analyses. The current results suggest that post-planning articulatory operations are not slower in L2 compared to L1, whereas earlier planning operations were slower in L2. In tandem with the earlier results of Broos et al. (2018; 2019) which suggested no L2 delay in the processes up to and including phonological encoding, it seems that the best candidate locus for the L2 delay is phonetic planning. However, this is not conclusive evidence, as we have no task that can isolate phonetic planning. Follow-up experiments are therefore needed to determine the origin of the L2 delay in the picture naming task. One potential avenue might be to search for temporal markers of early and late processes in on-line neural data, such as the electro-encephalogram, while speakers are preparing to name pictures. Laganaro (2017) argued that during such tasks, there are distinct "microstates" with a global stable electrical field, and that such microstates might be linked to distinct cognitive processes taking place during the task. If it were possible to isolate a microstate corresponding to phonetic encoding, then the present account predicts a delay in the duration of this state in L2 speakers.

References

Acheson, D. J., Ganushchak, L. Y., Christoffels, I. K., & Hagoort, P. (2012). Conflict monitoring in speech production: Physiological evidence from bilingual picture naming. *Brain and Language, 123(2)*, 131-136. doi: https://doi.org/10.1016/j.bandl.2012.08.008

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. doi: 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. doi: https://doi.org/10.1016/j.jml.2012.11.001

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.36, retrieved 11 November 2017 from http://www.praat.org/

Broos, W. P., Duyck, W., & Hartsuiker, R. J. (2018). Are higher-level processes delayed in second language word production? Evidence from picture naming and phoneme monitoring. *Language, Cognition, and Neuroscience, 33*(10), 1219-1234. doi: https://doi.org/10.1080/23273798.2018.1457168

Broos, W. P., Duyck, W., Hartsuiker, R. J. (2019). Monitoring speech production and comprehension: Where is the second-language delay? *Quarterly Journal of Experimental Psychology*, 72, 1601-1619. doi: https://doi.org/10.1177/1747021818807447

Colomé, À. (2001). Lexical activation in bilinguals' speech production: Language-specific or language-independent?. *Journal of Memory and Language*, *45*(4), 721-736. doi: https://doi.org/10.1006/jmla.2001.2793

Gollan, T. H., & Silverberg, N. B. (2001). Tip-of-the-tongue states in Hebrew–English bilinguals. *Bilingualism: Language and Cognition*, *4*(1), 63-83. doi: 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

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., & Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Memory & Cognition*, *33*(7), 1220-1234. doi: https://doi.org/10.3758/BF03193224

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*(3), 594-615. doi: https://doi.org/10.1017/S1366728911000332

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67-81. doi: https://doi.org/10.1017/S1366728998000133

Guo, T. M., & Peng, D. L. (2007). Speaking words in the second language: From semantics to phonology in 170 ms. *Neuroscience Research*, 57(3), 387-392. doi: https://doi.org/10.1016/j.neures.2006.11.010

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

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1), 101-144. doi: https://doi.org/10.1016/j.cognition.2002.06.001

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

Kawamoto, A. H., Liu, Q., Mura, K., & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. *Journal of Memory and Language*, *58*(2), 347-365. doi: https://doi.org/10.1016/j.jml.2007.06.002

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism: Language and Cognition, 9*(2), 119-135. doi: https://doi.org/10.1017/S1366728906002483

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. doi: https://doi.org/10.1006/jmla.1994.1008

Laganaro, M. (2017). Inter-study and inter-Individual Consistency and Variability of EEG/ERP Microstate Sequences in Referential Word Production. *Brain Topography, 30*(6), 785-796. doi: https://doi/org/ 10.1007/s10548-017-0580-0

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), 952-965. http://dx.doi.org/10.1037/a0023217

Levelt, W. J. M., Roelofs, A., & Meyers, A. S. (1999). A theory of lexical access in speech production. Behavioral and Brain Sciences, 22(1), 1-38.

Monti, M. M., Osherson, D. N., Martinez, M. J., & Parsons, L. M. (2007). Functional neuroanatomy of deductive inference: a language-independent distributed network. *Neuroimage*, *37*(3), 1005-1016. doi: https://doi.org/10.1016/j.neuroimage.2007.04.069

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673-690. doi: https://doi.org 10.1007/s11135-006-9018-6

Poulisse, N. (1999). *Slips of the tongue: Speech errors in first and second language production* (Vol. 20). John Benjamins Publishing.

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

Rastle, K., Croot, K. P., Harrington, J. M., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: consonantal effects on delayed naming latency and onset duration. *Journal of Experimental Psychology: Human Perception and Performance*, *31*(5), 1083-1095. doi: http://dx.doi.org/10.1037/0096-1523.31.5.1083

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/doi: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*, 159-179. doi: https:// doi.org /10.1007/s10936-011-9183-1

Sadat, J., Martin, C. D., Magnuson, J.S., Alario, F. X., & Costa, A. (2016). Breaking down the bilingual cost in speech production. *Cognitive Science, 40,* 1911-1940. doi: https:// doi.org/10.1111/cogs.12315

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

Strijkers, K., Costa, A., & Thierry, G. (2010). Tracking lexical access in speech production: Electrophysiological correlates of word frequency and cognate effects. Cerbral Cortex, 20, 4, 912-928. doi: https://doi.org/10.1093/cercor/bhp153

Strijkers, K., Baus, C., Runnqvist, E., FizPatrick, 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

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190. doi: 10.1080/17470218.2013.850521

Footnotes

1: Translations of three picture names did not match the exact place of phonological complexity, due to constraints on material selection.

Appendix A: Target Picture Names, Zipf values, and mean orthographic Levenshtein values

Picture Names (English - Dutch)	Zipf Values (English - Dutch)	Mean Orthographic Levenshtein Distance
cast – gips	3.73 – 3.37	1.10 – 1.55
plug – stop	3.75 – 4.81	1.75 – 1.10
road – weg	5.05 – 5.29	1.55 – 1.00
horse – paard	4.96 – 4.92	1.55 – 1.00
doll – pop	4.39 – 4.31	1.35 – 1.00
witch – heks	4.44 – 4.41	1.50 – 1.55
heel – hak	3.78 – 3.44	1.50 – 1.00
dress – kleed	4.81 – 4.75	1.75 – 1.00
raft – vlot	3.67 – 3.21	1.40 – 1.55
coat – jas	4.61 – 4.68	1.35 – 1.00
ghost – spook	4.56 – 4.05	1.90 – 1.55
leg – been	4.75 – 4.73	1.35 – 1.00
plate – bord	4.41 – 4.44	1.50 – 1.20
snail – slak	3.24 – 3.36	1.80 – 1.10
stool – kruk	3.55 – 3.40	1.80 – 1.45
sock – kous	3.82 – 3.34	1.15 – 1.35
knife – mes	4.67 – 4.67	2.15 – 1.00
tire – wiel	1.29 – 3.83	1.25 – 1.00
shirt – hemd	4.67 – 4.08	1.70 – 1.80
cloud – wolk	4.03 – 3.72	1.70 – 1.30
wall – muur	4.82 – 4.82	1.20 – 1.40
roof – dak	4.55 – 4.74	1.55 – 1.00
bag – zak	4.96 – 4.98	1.00 – 1.00
shed – hut	3.71 – 5.91	1.50 – 1.05
hat – pet	4.80 – 4.55	1.00 – 1.20
	Mean English (SD): 4.20 (0.78) Mean Dutch (SD): 4.31 (0.69)	Mean English (SD): 1.49 (0.28) Mean Dutch (SD): 1.25 (0.25)

Appendix B. Accuracy as function of speakers' first language, response language, condition, and phonological complexity. N is the number of trials in the respective condition.

L1	response language	Condition	Complexity	N	Correct	%
English	English	delayed	complex	480	400	83%
English	English	delayed	simple	520	396	76%
English	English	regular	complex	480	382	80%
English	English	regular	simple	520	392	75%
Dutch	English	delayed	complex	516	221	43%
Dutch	English	delayed	simple	559	311	56%
Dutch	English	regular	complex	516	193	37%
Dutch	English	regular	simple	559	271	48%
Dutch	Dutch	delayed	complex	516	346	67%
Dutch	Dutch	delayed	simple	559	373	68%
Dutch	Dutch	delayed	complex	516	308	60%
Dutch	Dutch	delayed	simple	559	340	61%

Table note. Number of observations differ per condition as there were more participants with L1 Dutch than L1 English, and slightly more phonologically simple than complex items.

Figure Captions

Figure 1. A sketch of Levelt, Roelofs, and Meyer's (1999) language production model. The figure is an adaptation of Levelt et al. (1999; Figure 1) and used with permission (License number: 4253680871328).

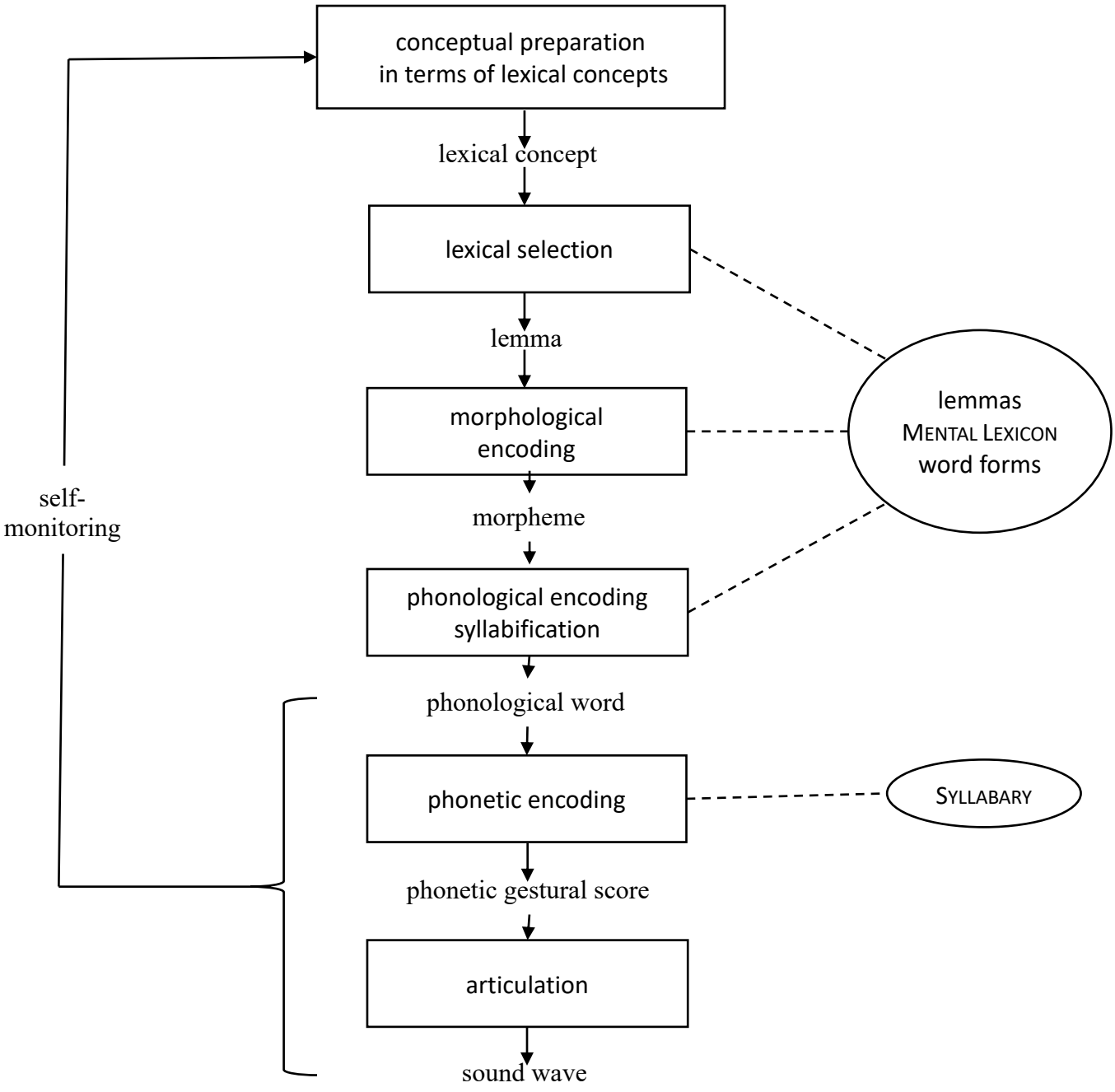
Figure 2. Graphical depiction of a delayed picture naming trial and a regular picture naming trial.

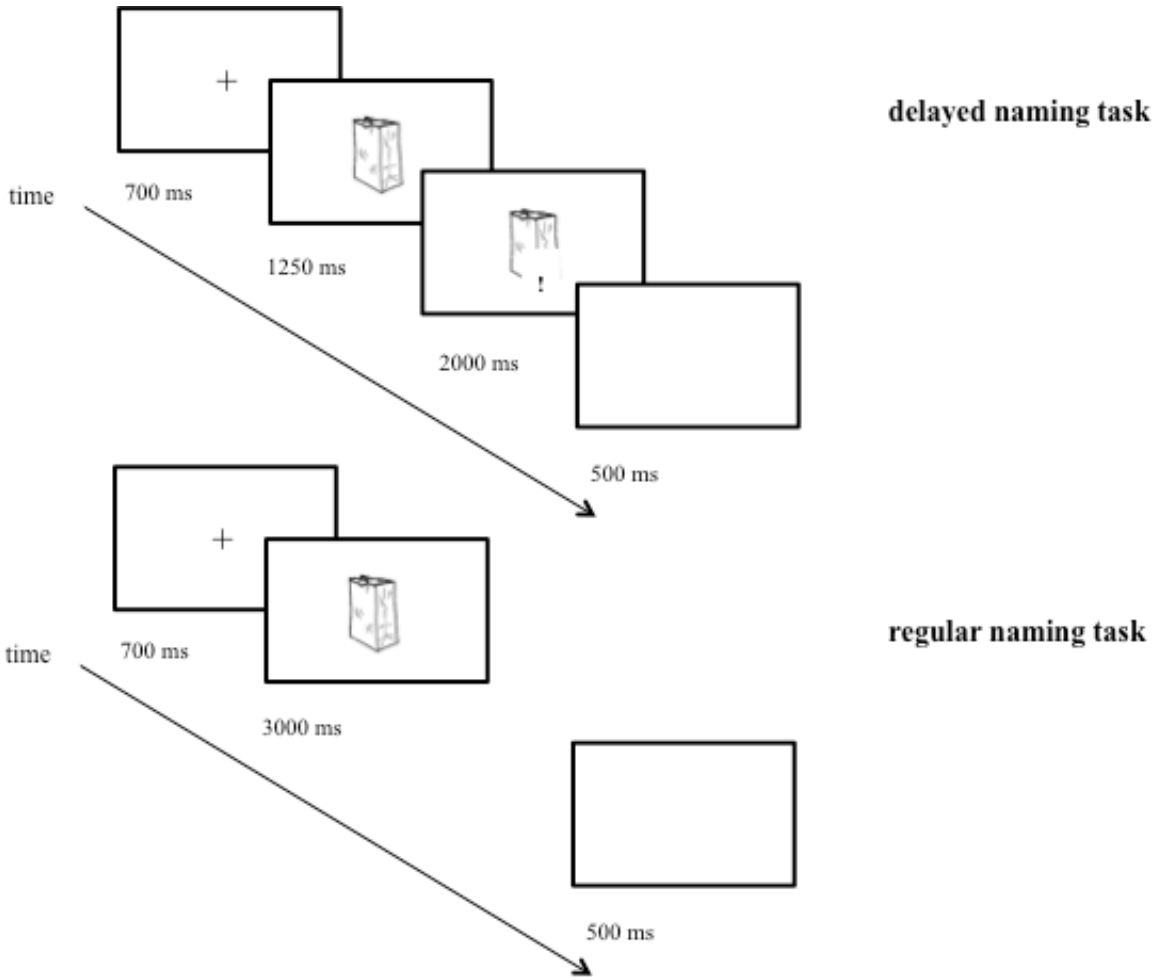
Figure 3. Mean observed reaction times in the between-subjects comparison (error bars denote the standard error away from the mean).

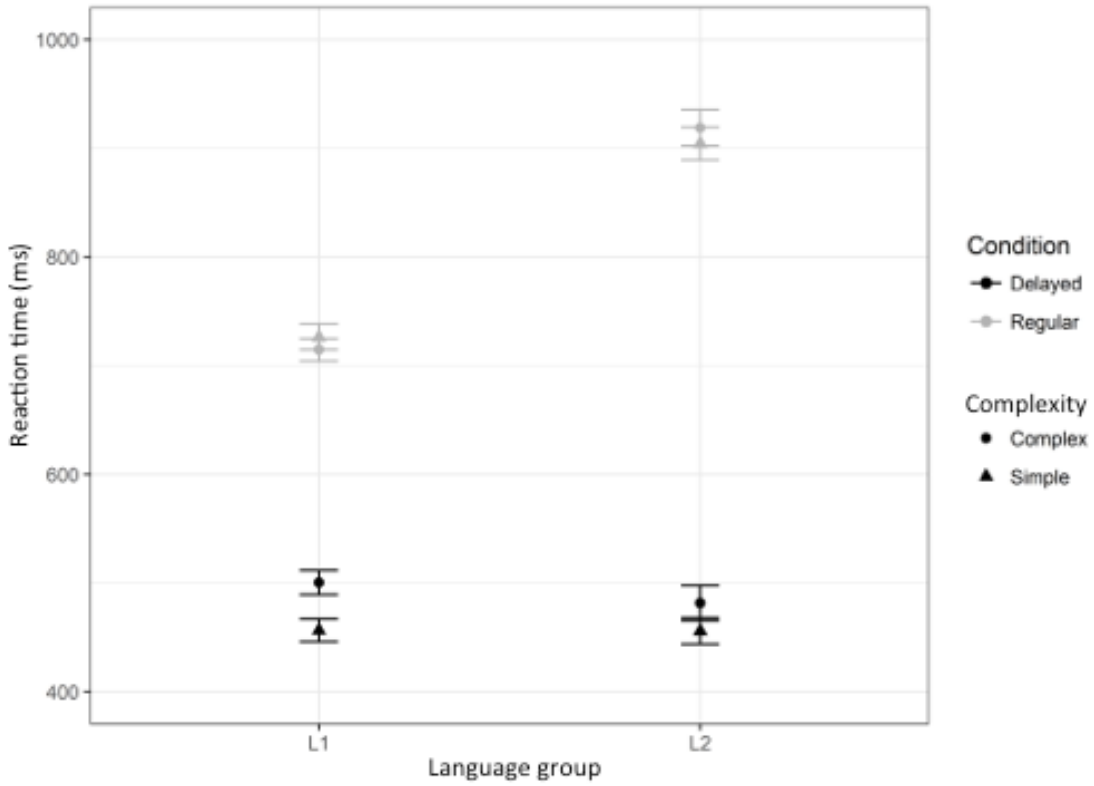
Figure 4. Mean observed accuracy in the between-subjects comparison (error bars denote the standard error away from the mean).

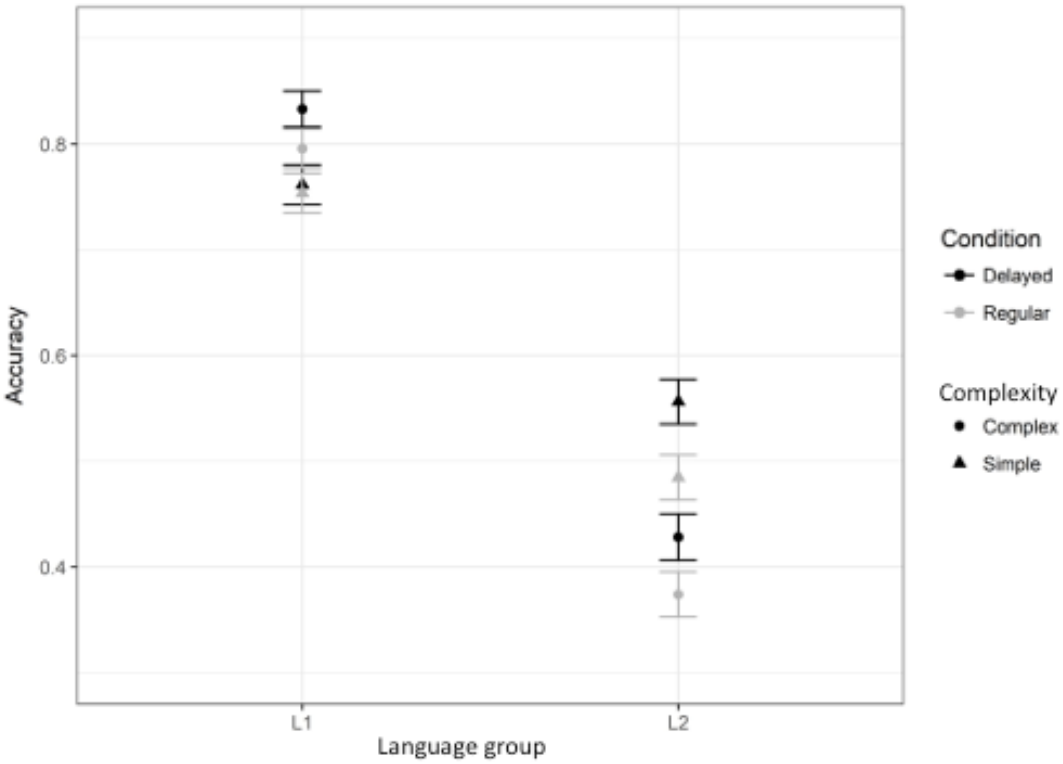
Figure 5. Mean observed reaction times in the within-subjects comparison (error bars denote the standard error away from the mean).

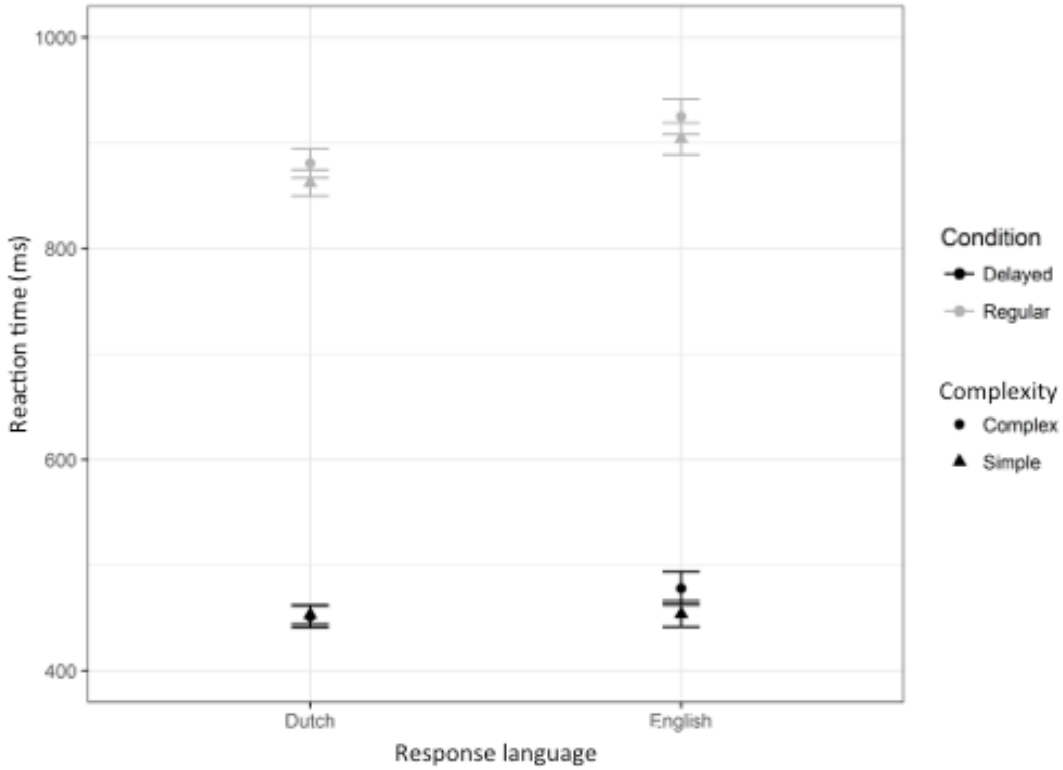
Figure 6. Mean observed accuracy in the within-subjects comparison (error bars denote the standard error away from the mean).

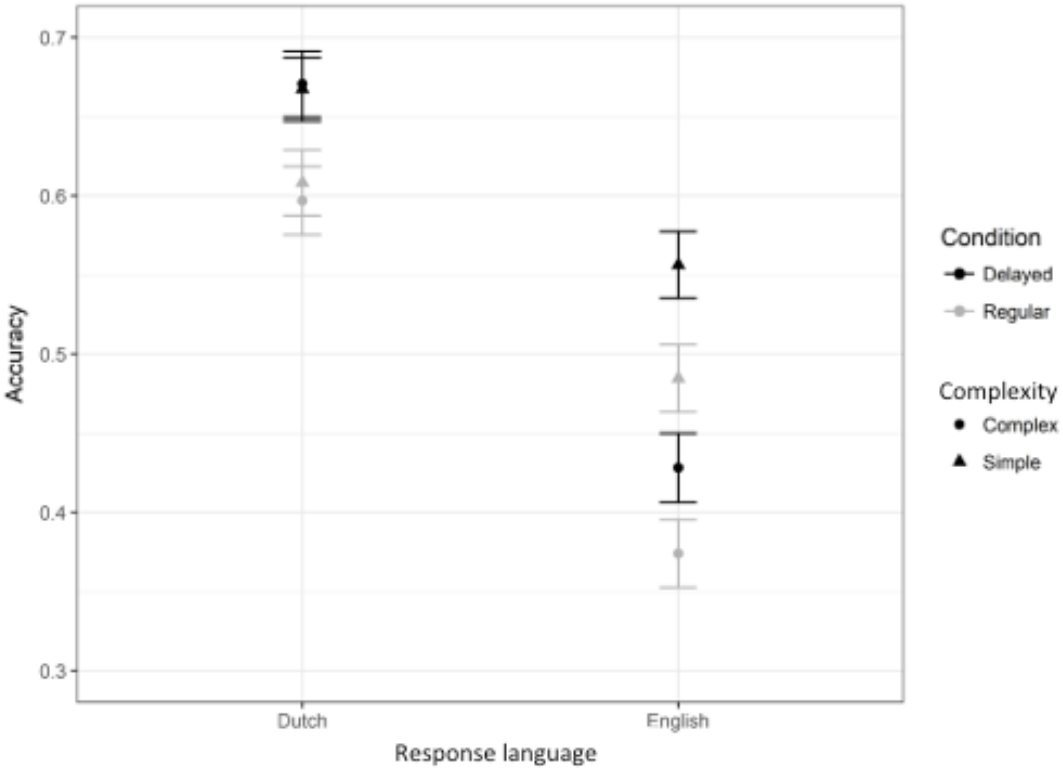

conceptual preparation
in terms of lexical concepts
lexical concept
lexical selection
lemma
morphological
encoding
morpheme
phonological encoding
syllabification
phonological word
phonetic encoding
phonetic gestural score
articulation
sound wave
self-
monitoring
lemmas
MENTAL LEXICON
word forms
SYLLABARY


delayed naming task
time
700 ms
1250 ms
!
2000 ms
500 ms
regular naming task
time
700 ms
3000 ms
500 ms


1000
800
600
400
Reaction time (ms)
L1
L2
Language group
Condition
Delayed
Regular
Complexity
Complex
Simple


0.8
0.6
0.4
Accuracy
L1
L2
Language group
Condition
Delayed
Regular
Complexity
Complex
Simple


1000
800
600
400
Reaction time (ms)
Dutch
English
Response language
Condition
Delayed
Regular
Complexity
Complex
Simple


0.7
0.6
0.5
0.4
0.3
Accuracy
Dutch
English
Response language
Condition
Delayed
Regular
Complexity
Complex
Simple