Are higher-level processes delayed in second language word production? Evidence from picture naming and phoneme monitoring

Wouter P. J. Broos, Wouter Duyck, & 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 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 W.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

There are clear disadvantages in the speed of word production and recognition in a second language (L2), relative to the first language (L1). Some accounts claim that these disadvantages occur because of a slow-down in lexical retrieval and phonological encoding. But it is also possible that the slow-down originates from a later part of the production process, namely articulatory planning or articulation. We used a phoneme monitoring task to study the time course of conceptualization, lexical retrieval, and phonological encoding during language production in the absence of articulation. First, we demonstrated that there was indeed an L2 disadvantage of 102 ms in a picture-word interference (PWI) task with phonologically related and unrelated distractor words. Next, participants from the same population performed a combined phoneme monitoring task / PWI task with the same stimuli: they monitored for the occurrence of a phoneme in a picture name while ignoring a distractor word. In both the PWI task and the combined phoneme monitoring/PWI task, there was phonological facilitation, suggesting that both tasks are similar up to the level of phonological encoding. Importantly, L2 speakers were not slower in phoneme monitoring than L1 speakers. These findings suggest that the slow-down typically observed in L2 speech production may not be situated at phonological or pre-phonological stages of speech production, but rather in a later stage of speech production.

Keywords: self-monitoring, picture word interference task, phonemic overlap, second language processing

Introduction

Speaking in one's native language seems to be effortless: we can produce the right words quickly and accurately. However, when having to speak in a second language, we tend to speak slower and be more error-prone (Van Hest, 1996). For instance, several studies reported that picture naming in a second language (L2) is slower than in a first language (L1) (Gollan, Montoya, Cera, & Sandoval, 2008; Starreveld, de Groot, Rossmark, & van Hell, 2014). There are several hypotheses explaining these L2 disadvantages, but they often have in common that L2 speakers would be slower because they have difficulty retrieving the words from the mental lexicon. However, a slow-down in picture naming does not necessarily imply that lexical processes are slower, as this task not only involves higher-level speech planning processes, but also includes lower-level processes such as articulatory planning and articulation (Hanulová, Davidson, & Indefrey, 2011). The aim of this study is to test whether L2 speakers are indeed slower because of difficulties in higher-level processes such as conceptualization, lexical retrieval, and phonological encoding or alternatively, whether the slow-down is situated further downstream in the speech production process.

Multiple studies have shown that L2 speech production is slower, more disfluent, and more prone to errors than L1 speech (Gollan & Silverberg, 2001; Poulisse, 1999; Poulisse, 2000). Poulisse (1999), for instance, found exactly 2000 slips in 35 hours of English (L2) speech production while only 137 slips were found in the same amount of time in L1 speech. Furthermore, a proficiency effect was found in that more proficient L2 speakers made fewer errors than speakers that were less proficient in their L2. Additionally, L2 speakers made more errors in content words than L1 speakers. The Tip-of-the-Tongue (TOT) phenomenon, where speakers cannot find a word they are certain they know, also occurs more frequently in L2 than L1 speakers. Gollan and Silverberg (2001) tested monolingual English speakers and bilingual Hebrew-English speakers by presenting them with descriptions of words. The bilingual participants showed a higher TOT rate than monolingual speakers in both languages.

One hypothesis that explains the slow-down in L2 speakers is the weaker-links hypothesis (Gollan et al., 2008). The weaker-links hypothesis starts from the observation that bilinguals necessarily have to divide language practice across two languages, so that lexical representations of L2 words (and to a certain extent L1 words) are weaker and less detailed (Finkbeiner, Forster, Nicol, & Nakamura, 2004; Gollan et al., 2008). As a consequence, it is more difficult for bilinguals to access linguistic representations in L2 which results in slower and less accurate retrieval of words. In addition, this leads to weaker activation spreading to other processing levels in L2 speakers. Gollan and Silverberg's (2001) TOT study suggests that higher-level processes such as lexical retrieval are more difficult in L2 than in L1. Their findings are consistent with the notion that competition between translation equivalents causes TOT but also with the claim that less frequent word use causes this phenomenon. Additionally, Gollan, Montoya, and Fennema-Notestine (2005) asked whether the L2 slow-down would still be present if Spanish-English bilinguals (whose dominant language was English) would repeatedly name the same pictures in a picture naming task. The findings were compared to those of English monolinguals. Consistent with the weaker-links hypothesis, the L2 slow-down disappeared in the bilingual group with practice: they were still significantly slower than the monolinguals for the third repetition but no significant differences were found for the fifth repetition. Ivanova and Costa (2008), however, tested a group of monolinguals Spanish speakers, a group of Spanish-Catalan bilinguals whose dominant language was their L1 (as opposed to a bilingual group whose dominant language was the L2 as in Gollan et al. 2008) and a group of Catalan-Spanish bilinguals. A slow-down was found when comparing the monolingual Spanish group and the bilingual Spanish-Catalan group in that the bilinguals were slower in naming pictures in both their L1 and L2 as opposed to the monolinguals. The bilingual Catalan-Spanish group was also slower at naming pictures than the monolingual group. Moreover, the L2 slow-down was not resolved in either of the bilingual groups after five repetitions, a finding that does not support the weaker-links hypothesis.

Alternatively, it is also possible that L2 delays in production occur farther downstream (i.e., during phonetic planning or articulation). After all, the processes involved in articulation are clearly effortful and time consuming (i.e., they take longer than planning according to Indefrey and Levelt's (2004) time course analysis of speech production) making them a possible candidate for L2 disadvantages. One reason articulation in L2 might be particularly effortful is the need to program and execute speech motor commands that are unusual or nonexistent in L1. Simmonds, Wise, and Leech (2011) reviewed L2 speech production with regard to articulation and the integration of motor and sensory aspects of non-native speech. They argue that the articulation of non-native phonemes is particularly difficult for L2 speakers (see also Alario, Goslin, Michel, & Laganaro, 2010). Hanulová et al. (2011) reviewed picture naming studies that used several experimental designs and also argue for the L2 disadvantage in picture naming to be situated at the post-lexical level. Hence, the difficulties that L2 speakers encounter are not necessarily situated at the semantic or phonological stages of speech production, but their underlying cause may be later during the process. We will refer to this possibility as the articulatory delay hypothesis.

There has been empirical support for the articulatory delay hypothesis. Hanulová, Davidson, and Indefrey (2008) for instance, performed an ERP study where Dutch-English bilinguals were asked to perform a delayed naming task in a go/no-go paradigm. The go/no-go paradigm in this study entailed that participants either do or do not press a button, depending on a particular decision that had to be made. Before pressing the button, participants were asked to either decide if the depicted object was manmade or natural or whether the picture name started with a particular phoneme (see Schmitt, Munte, and Kutas (2000) for a dual go/no-go task). Whether the button was pressed or not depended on the decision. This way, the paradigm reveals the time course of both semantic and phonological information of the picture that is present on the screen at that time. The N200 was the main component of interest since this has been argued to reflect response inhibition (Jodo & Kayama, 1992). The rationale behind this particular paradigm is that participants can only inhibit a response if there is enough information to do so, leading to corresponding N200 responses. The timing of these responses can then be used to determine when semantic and phonological activation is present. Hanulová et al. (2008) did not find a significant difference between the intervals between semantic and phonological N200 responses in L1 or L2 (also see Guo & Peng, 2007). This does not support the existence of a slow-down in the L2, at least up through phonological retrieval of the initial phoneme. It rather suggests that the slow-down occurs later in the speech production process.

To test whether the slow-down in L2 is situated at a pre-phonological or post-phonological stage, our study used the *phoneme monitoring task in production*. In this task, participants silently extract a word from their mental lexicon and respond with a button press if that name contains a target phoneme. Arguably, this task involves the planning stage up through phonological encoding, but not articulatory planning or actual articulation. As the participants do not have to produce speech in the task, it is highly unlikely that they will plan articulation. The phoneme monitoring task was introduced by Wheeldon and Levelt (1995) who aimed to determine the time course of phonological encoding. Participants first memorized Dutch-English translation pairs, such as *lifter-hitchhiker*. Once the pairs were remembered correctly, the experimental phase began in which a phoneme and an English word were presented auditorily. The participants were asked to press a button if the phoneme was present in the Dutch translation of the English word they just heard. Participants reacted significantly faster to the target phoneme if it was present in the first syllable of the Dutch translation (e.g., /l/) than when it was situated in the second syllable (/t/), indicating that the monitoring process is sequential. Furthermore, there was a significant slowdown in reaction time between the first and last phoneme of the first syllable, whereas there was no such difference in the second syllable. This suggests that phoneme monitoring speeds up from the second syllable onwards.

The phoneme monitoring task has also been used in bilingual speakers (e.g., Colomé, 2001) and in combination with distractor words (e.g., Ganushchak & Schiller, 2008), as is the case in our experiments. Colomé (2001) used the phoneme monitoring task to investigate whether activation of lexical entries and their corresponding phonemic representations spreads to the non-target language in bilinguals. Catalan-Spanish bilinguals decided whether a particular phoneme was present in the Catalan name of a target picture. The participants were slower in rejecting phonemes that belonged to the Spanish translation than those that were absent in both languages. This is explained by arguing that the picture activated a concept that is shared by Catalan and Spanish, which in turn activated not only the name of the picture in both languages but even the phonemes occurring in those names.

In sum, the literature on phoneme monitoring suggests that the task taps into speech planning (up through phonological encoding), that it can be used with picture stimuli (also see Özdemir, Roelofs, & Levelt, 2007) in speakers using a second language, and in combination with a picture-word interference task, all of which are features of the experiments reported below.

In the present study, we use the phoneme monitoring task with the purpose of isolating the stages of lexical retrieval and phonological encoding from the stages of articulatory planning and articulation. That is, phoneme monitoring arguably requires the speaker to retrieve the target word and spell out its phonemes, but it does not require articulatory processing. If the L2 disadvantage often observed in speech production is situated at the stages of lexical retrieval or phonological encoding, we expect bilingual L2 English speakers to be slower in phoneme monitoring than monolingual L1 English speakers. However, if such delays primarily reflect differences in articulatory processing, we expect no difference in phoneme monitoring times between languages. One possible caveat is that phoneme monitoring is a metalinguistic task (Vigliocco & Hartsuiker, 2002), which does not necessarily tap into all processes of normal speech production. To deal with this potential issue, our experiments test whether phoneme monitoring is sensitive to two speech planning variables. First, Levelt, Roelofs, and Meyer (1999) argued that phonemes in an earlier position are available earlier than phonemes in a later position. Hence, in the phoneme monitoring task, word-initial phonemes should be detected more quickly than word-final phonemes (as was the case in Wheeldon & Levelt, 1995). Second, speech production is influenced by phonological overlap of a distractor word both at the beginning and the end of a word (Meyer & Schriefers, 1991) and this facilitation effect occurs during phonological encoding (Levelt et al., 1999). If the phoneme monitoring task in our study taps into regular word form retrieval, then reaction times should be affected by phonological overlap between the distractor word and picture name.

Specifically, six conditions will be used in the following experiments, resulting from crossing three different amounts of phonological overlap between distractor word and picture name (double, single, and no overlap) with two places where the target phoneme can be placed (onset or coda). We predict that reaction times will be shorter if the target phoneme is placed in onset position (e.g., /b/ for picture *bag*) as opposed to coda position (e.g., /g/ for picture *bag*). Moreover, reaction times will also be shorter if there is more phonological overlap (e.g., *bag-bug*) than when there is less (e.g., *bag-bin*) or no overlap (e.g., *bag-rod*) between picture name and distractor word. According to hypotheses that assume an L2 slow-down during lexical retrieval and phonological encoding, a language effect should be seen in that the bilingual L2 speakers are slower than the monolingual L1 speakers. Furthermore, slower planning also suggests that facilitation in L2 speakers should be stronger if the phonemes between the picture name and distractor word overlap. As those representations are weaker in L2 speakers, they should benefit more from overlapping phonemes because there is more room for facilitation, relative to L1 speakers. In other words, phonological overlap might be more beneficial to L2 speakers as the weaker-links hypothesis presumes that the lexical representations are weaker and the retrieval of these representations is slower.

Before we report the speech monitoring experiments, we will first verify whether L2 speakers of English are indeed slower at naming pictures than L1 speakers. As the speech monitoring tasks involved the presentation of distractor words, we also presented distractor words in the picture naming task, rendering it a picture-word interference (PWI) task. The participants in the PWI task were English monolingual L1 speakers and Dutch-English bilingual L2 speakers. Participants that were tested in the combined PWI/phoneme monitoring task originated from the same population. In sum, the PWI and phoneme monitoring experiments were kept as similar as possible. We hypothesized that L1 speakers will be significantly faster in naming pictures than L2 speakers. Moreover, we expected a phonological facilitation effect and possibly stronger phonological facilitation for a larger amount of phonological overlap.

Experiment 1: Picture Word Interference

Methods

Participants

Thirty-five monolingual English L1 speakers (male = 9 / female = 26, mean age = 34) and 48 bilingual Dutch-English L2 speakers (male = 10 / female = 38, mean age = 20) participated in the experiment. Participants, mostly students, were recruited from the participant pools of the University of Leeds and Ghent University, respectively. Participants were monetarily compensated for their participation. There was a small subgroup of monolingual participants over 40 years of age, which increases the mean age of that group. Participants all reported to have normal hearing, normal to corrected-to-normal sight, and not to have dyslexia. 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 filled in a questionnaire and were asked to rate their English proficiency on a scale from one (very poor) to seven (very good). An overview of the participants' proficiency scores can be found in Table 1 below. The table shows that there is slightly more variation in English ratings compared to Dutch ratings, but their L2 level seems to be rather homogeneous. Mean language proficiency across measures was significantly higher in Dutch than in English (t(80.37) = 8.67 p < .001).

[Insert Table 1 around here]

Materials

Fifty black and white line drawings of objects were presented together with the same number of distractor words of which 25 pictures were *target* pictures (see Appendix A for a list of target stimuli). The experiment consisted of five blocks in total and every target picture was presented 12 times during the entire experiment[1]. All picture names

[1] Only half of these pictures were analyzed because of the experimental design of Experiment 2. In that experiment, a phoneme monitoring task had to be performed. The phoneme was present in the picture name in half of the trials and absent in the other half. Since we wanted to keep the set-up of Experiment 1 as similar as possible to that of Experiment 2 (Experiment 2 was conducted first) we only analyzed the trials where the phoneme was present. Therefore, only half of the pictures were analyzed in the end, leading to a total of 7200 target trials (25*12*48/2 = 7200).

and distractor words were monosyllabic nouns with a CVC-structure. The mapping between phonology and orthography was regular for all picture names and distractor words.

Three different overlap categories were created that differed in phonological overlap between picture name and distractor word: double overlap, single overlap, and no overlap. Double overlap consisted of a picture-word pair in which the consonants of both the onset and coda were identical (e.g., *bag-bug*). Single overlap had only one phoneme in common between the picture and distractor word in either onset (e.g., *bag-bet*) or coda (e.g., *bag-fog*). Finally, no overlap contained a picture name and a distractor word without any phoneme in common (e.g., *bag-rod*). Note that Experiment 2 uses the same stimuli, but with an additional factor, namely position of the target phoneme (see Table 3). This position coincides with the locus of overlap in single overlap (e.g., for the pair *bag-bet* the target phoneme would be the /b/). For the sake of comparison with these further experiments, we included position as a factor in the design, although this factor was of course only meaningful in single overlap.

Procedure

Participants were seated in a silent room and were placed in front of a computer screen. The pictures were presented in the middle of the screen (width and height both set at 75% in E-prime 2.0) and participants were asked to name the pictures as soon as they saw the picture appearing on the screen. The distractor words (Times New Roman, 26, set at width 25% and height 15% in E-prime 2.0) were presented across the lower half of the pictures. The pictures were taken from the Severens, Van Lommel, Ratinckx, and Hartsuiker (2005) database.

The experiment consisted of a familiarization phase, a practice phase, and an experimental phase. During the familiarization phase, participants were simultaneously presented with each picture and its name. Participants were asked to look at the pictures without responding. The practice phase contained three trials that were added before the experimental phase began. Pictures and distractor words used in this phase were not presented in the experimental phase. During the practice and experimental phase, a fixation cross was presented on the screen for 250 ms after which the picture and distractor word were shown for 3000 ms. The next trial was started after a blank screen was presented for 1000 ms. Reaction times were measured as soon as the picture was presented on the screen. The experiment took twenty minutes to complete. Figure 1 represents the procedure of the trials.

[Insert Figure 1 around here]

Data analysis

Before the data were analyzed, trials were deleted because of incorrect, non-fluent, or missing responses. Fifty-five out of 7200 trials (L2 data set) were not properly recorded by E-Prime 2.0 and could therefore not be analysed. The computer program Praat (Boersma & Weenink, 2017) and the software package Chronset (Roux, Armstrong, & Carreiras, 2016) were used to determine the response latencies. Chronset is an automatic speech recognition program that uses phonetic information to determine speech onset. Some participants spoke rather softly, leading to a subset of trials where the program could not determine speech onset. These trials were annotated by hand (1803 trials). A subset of the data that Chronset annotated (415 trials) were also manually annotated while a correlation analyses was performed on these trials. This way, the accuracy of the Chronset package could be objectively measured. The correlation between the hand-coded and automatically coded speech was 0.9 meaning that Chronset was quite accurate in determining speech onset. L1 speakers made 155/5250 mistakes (2.95%) whereas L2 speakers answered 365/7145 trials (5.11%) incorrectly. These trials were removed from the data set.

Reaction times that fell above or below 2.5 standard deviations away from the mean per overlap category and speaker were also deleted from this data set. This amounted to 369/11875 trials (3.11%) meaning that a total of 11506 trials were used for the final analyses. The data set was analyzed 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) packages 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). Sum coding was used for all analyses where the mean of all factors amounts to zero. Type II Wald Chi square tests were conducted in order to calculate main effects and interaction effects. The function 'lsmeans' was used to determine significant differences between all different contrasts. Additionally, we conducted traditional ANOVAs on aggregated data per subject (F1) and item (F2). These showed an almost identical pattern of results (see Appendix C for summary tables). The R-scripts and data sets for the F1/F2 analysis (and the linear mixed effects analysis) can be found on Open Science Framework (https://osf.io/7jncs/).

Results

Reaction times

The fixed factors that were included in the final model were Language, Degree of Overlap, and Position. Interactions were added for all fixed factors. The factor Language consisted of two levels (L1 and L2), Degree of Overlap consisted of three levels (no overlap, single overlap, and double overlap), and Position involved two levels (onset and coda). The factor 'Trial Number' was added as covariate to account for a potential decrease in reaction time due to learning that could occur because of repeated exposure to the same pictures. Random slopes were included based on the 'maximal random effects structure' approach, as suggested by Barr, Levy, Scheepers, and Tily (2013). This means that the maximal random slopes structure would consist of the three-way interaction of Degree of Overlap, Language, and Position for item (Picture) and the two-way interaction of Degree of Overlap and Position for subject (Subject). Note that Language could not be added as random slope to Subject as this was a between-subject variable. What is also important to mention is that by including the three- and two-way interactions as random slopes, fixed effects (and lower level interactions) are added automatically because of the way in which R handles factors (see Levy, 2014) . The maximal model did not converge and we therefore implemented the forward selection procedure of Barr et al. (2013) to determine the final model.

We started the forward selection procedure by creating several models in which each model contained only one random slope for either subject of item. This random slope could be a random slope of a main effect or an interaction effect. These models were run and only if they converged were they compared to the null model (a model without random slopes). If the p-value fell below .2, we added the random slope to the null model. After all converging models were tested for significance, we ran the new 'null' model (base model) to see if it converged. If the base model did not converge, we removed random slopes with the highest p-value in a stepwise manner until it converged. If the base model converged, we compared the base model to models that contained random slopes of models that did not previously converge or were not significant before (base model + random slope of non-converging/non-significant model). If one or several of the comparisons between the base model and other models were significant, we created a new base model and repeated the process until no other model converged. The final model contained the random slopes of the fixed factors Language, Position, and Degree of Overlap for item (Picture) and the random slope of Degree of Overlap for subject (Subject). No interactions of fixed factors were added as random slopes. Type II Wald Chi square tests were conducted in order to calculate main effects and interaction effects.

[Insert Figure 2 around here]

As shown in Figure 2, L1 speakers are clearly faster in naming pictures than L2 speakers and this effect was indeed significant (χ^2 (1) = 16.73, p < .001). Degree of Overlap also showed a significant main effect (χ^2 (2) = 29.16, p < .001). The factor Position did not reach significance (χ^2 (1) = 0.57, p = .45), but note again that this distinction was only meaningful for single overlap, where it indicated the place of overlap (onset vs. coda). A substantial learning effect was seen where participants named the pictures faster at the end of the experiment (χ^2 (1) = 146.64, p < .001). None of the interaction effects were significant (p-values > .1). As is clear from Figure 2 and from the lack of interaction between Position and Degree of Overlap, there seems to be similar phonological facilitation from begin-related and end-related phonemes.

Accuracy

Fixed factors that were included in the final generalized linear mixed effects model were Language, Degree of Overlap, and Position. Interactions for all fixed factors were included. An attempt was made to include a maximal random effects structure, but the model did not converge. The final model only contained Degree of Overlap and Language as random slope for item (Picture) but no random slopes were added for subject (Subject). Note that the model automatically uses logistic regression. Type II Wald Chi square tests were conducted in order to calculate main effects and interaction effects.

[Insert Figure 3 around here]

Figure 3 reveals that L1 speakers are significantly more accurate than L2 speakers ($\chi^2(1) = 7.07$, $p = .008$). The interaction of Language and Position was significant as well ($\chi^2(2) = 10.79$, $p = .005$) suggesting that the difference in accuracy between onset and coda is smaller in L2 than in L1. No other main effects or interaction effects reached significance (all p-values > .1).

Discussion

Experiment 1 has confirmed that there is indeed an L2 delay when naming pictures in a picture-word interference paradigm. The difference between L1 and L2 speakers was exactly 102 ms. This finding is further supported by model comparison, which showed that there was evidence for the model that includes Language as a factor. We found no evidence to suggest that phonological overlap in onset position yields more facilitation than overlap in coda position. Finally, analyses on accuracy data revealed that L2 speakers made more mistakes than L1 speakers when naming the pictures. No speed/accuracy trade-off is seen in L2 speakers since both their reaction times and accuracy scores are lower than those of L1 speakers.

In sum, Experiment 1 shows that in this population and with these picture-word stimuli there is an L2 delay in picture naming of 102 ms. Furthermore, there was a classical phonological facilitation effect in both L1 and L2 of comparable magnitude. Since Experiment 1 has confirmed the L2 delay during picture naming, Experiment 2 below will focus on pinpointing the locus of this delay in the speech production process. This experiment will use a phoneme monitoring task to tap into speech production processes in the absence of articulation. To check whether the paradigm taps into normal production processes there were again phonologically related and unrelated phonological distractors; we expect to see phonological facilitation in phoneme monitoring too.

Experiment 2: Phoneme Monitoring

Methods

Participants

Fifty-four monolingual native English speakers (male = 12 / female = 42, mean age = 29) and 43 Dutch-English bilinguals (10 males and 33 females, mean age = 19.6) participated in the experiment. Participants, mostly students, were recruited from the participant pools of the University of Leeds and Ghent University, respectively. Participants were monetarily compensated for participation. None of the participants participated in Experiment 1. Participants all reported to have normal hearing, normal to corrected-to-normal sight, and not to have dyslexia. Table 2 describes English proficiency measures by means of self-ratings in which participants were asked to judge how good they were at writing, speaking, listening, and reading in English on a scale from one (very poor) to seven (very good). The table shows that there is slightly more variation in English ratings than Dutch ratings, but their L2 level seems to be rather homogeneous. Mean language proficiency across measures was significantly higher in Dutch than in English (t(57.43) = 4.98, p < .001).

[Insert Table 2 around here]

Materials

The pictures and distractor words were identical to the ones used in Experiment 1. Additionally, target letters were presented on the screen for the purpose of phoneme monitoring (all letters mapped onto only one English phoneme). Only trials where the phoneme was present in the picture name were considered. Table 3 gives an overview of the experimental conditions. For the yes-answers, either the onset (e.g., /b/ for *bag*) or coda (e.g., /g/ for *bag*) phoneme was selected as the target for phoneme monitoring (depending on the condition). For the no-answers, which served as fillers, a phoneme was selected that corresponded to neither the onset nor the coda (e.g., /l/ for *bag*).

[Insert Table 3 around here]

Table 3 shows examples of our stimuli as a function of degree of overlap and target phoneme location. In order to compare the different degrees of overlap, the same pictures were used twice in every overlap category with the same distractor word except for single overlap (in which case a different distractor was used for onset and coda position).

Procedure

The pictures were preceded by a letter that indicated the target phoneme (presented in Times New Roman, 48 font). The pictures were presented in exactly the same manner as in Experiment 1. Stimuli were presented in a pseudorandom order as there were certain restrictions on stimulus presentation: 1. No more than three trials with correct identical answers could be presented in a row (yes or no) / 2. No more than three consecutive trials were presented where the target phoneme occurred at the beginning or end of the word (onset vs. coda) / 3. Maximally two of the same consecutive target phonemes were presented / 4. The same overlap category did not appear more than twice in a row.

Participants were seated in a silent room and were placed in front of a computer screen. They were asked to perform a phoneme monitoring task while being shown a phoneme and subsequently a picture together with a distractor word. Participants were asked to decide whether the phoneme was present in the English picture name and ignore the distractor word. In order to respond, a button on a response box was pressed; the green button (right) if the phoneme was present in the picture name and the blue button (left) if it was absent. Participants were instructed to keep their hands on the response box in order to limit variation in reaction times as much as possible. Moreover, participants were asked to react as fast as they could but were told to slow down if the speed negatively affected accuracy.

The experiment again consisted of a familiarization phase, a practice phase, and an experimental phase. The procedure of the practice and experimental phase were slightly different than in Experiment 1. During the practice and experimental phase, the participants were asked to decide whether the phoneme that was presented first was present in the name of the picture. A fixation cross was presented on the screen for 250 ms after which the target phoneme was shown on the screen for 1000 ms. Another fixation cross was presented for 250 ms while the picture was shown for 1000 ms. The next trial began when the participant responded. Reaction times were measured as soon as the picture was presented on the screen. The experiment took thirty minutes to complete. Figure 4 represents the sequence of events during a trial. The same procedure was used for both the monolingual and bilingual group. The only exception was that the oral instructions were given in Dutch to the bilingual group (instead of English oral instructions, which were given to the monolingual English group). The written instructions that were presented on the screen in the Dutch bilingual group, however, were provided in English.

[Insert Figure 4 around here]

Data analysis

Twenty-eight trials (out of 8100; 0.3%) were not recorded by E-prime due to technical difficulties. Four participants were excluded from the analysis as they misunderstood the task (which was determined based on excessive error rates in several categories)[2]. The trials that were answered incorrectly were removed first, which amounted to 1497

[2] If more than 20 out of 25 trials were answered incorrectly per category (e.g., double overlap, yes answer), then the participant was excluded from the data set. Four participants answered at least 24 out of 25 trials incorrectly, indicating that they clearly misunderstood the task and were therefore excluded. Other participants showed a range from 0 to 8 incorrect trials out of 25 (although the majority of the participants only answered 1 or 2 trials incorrectly per category).

trials out of 13922 (10.8%). Reaction times that fell above or below 2.5 standard deviations away from the mean per overlap category and speaker were also deleted from the data sets, which amounted to 392 outliers (2.8%). As in Experiment 1, Type II Wald Chi square tests were run in order to calculate main and interaction effects. Further traditional ANOVAs with subjects (F1) and items (F2) as a random factor were run as well; these showed an almost identical pattern of results as the chi square tests (see Appendix C for summary tables). The R-scripts and data sets for the F1/F2 analysis and the linear mixed effects analysis can be found on Open Science Framework (https://osf.io/7jncs/).

Results

Reaction times

A linear mixed effects model was created which contained the fixed factors Degree of Overlap, Position, and Language while Trial Number was included as a co-variate. Interactions of all these fixed factors were added to the model. The maximal model did not converge, so we used the forward selection procedure. The fixed factor Position and the interaction of Position and Degree of Overlap were added as random slopes to item (Picture) whereas the fixed factors Position and Degree of Overlap were added to subject (Subject). There was a main effect of Position (χ^2 (1) = 115.23, $p < .001$) indicating that the target phoneme was recognized faster in the onset than in the coda position. A main effect of Degree of Overlap was also observed (χ^2 (2) = 48.99, $p < .001$). Importantly, the factor Language was not significant (χ^2 (1) = 0.83, $p = .36$).

Thus, this analysis does not support the hypothesis that lexical retrieval or phonological encoding is delayed in a second language. An overall learning effect was observed as well (χ^2 (1) = 271.55, $p < .001$) as Trial Number reached significance. No interaction effects reached significance (all p-values > .1).

Accuracy

The final generalized linear mixed effect model contained the fixed factors Degree of Overlap, Position, and Language. Interactions of all these factors were added to the model. The maximal model did not converge, but forward modelling revealed that the fixed factors Language, Degree of Overlap, and Position could be added to item (Picture) while Position should be added to subject (Subject). There was a main effect of Position (χ^2 (1) = 53.53, $p < .001$) indicating that participants were more accurate at trials where the target phoneme was presented in the onset position. A main effect of Degree of Overlap was also observed (χ^2 (2) = 50.41, $p < .001$). Language does appear to be significant when accuracy is concerned (χ^2 (1) = 6.32, $p = .01$) but note that the L2 speakers were more accurate than L1 speakers. One interaction effect reached significance which was the interaction between Degree of Overlap and Position (χ^2 (2) = 18.52, $p < .001$) indicating that the difference between overlap categories was larger in the coda than the onset position.

Separate analysis L1 and L2

Analyses of reaction times and accuracy scores were are also performed for L1 and L2 speakers separately. The reason for this split pertains to the relation between speech monitoring and speech production processes. In particular, if position effects and phonological effects are also found in these analyses, then this confirms that the same processes are shared between picture naming and phoneme monitoring. It is crucial to verify this claim if one wants to argue for an L2 delay in a particular stage of speech production.

Reaction times. The final linear mixed effects model for the L1 speakers contained the fixed factors Degree of Overlap and Position, and Trial Number as co-variate. An interaction of Degree of Overlap and Position was also added to the model. As the maximal model did not converge, we applied the forward selection procedure. The final model for the L1 data set contained the random slopes of the fixed factors Degree of Overlap and Position and its two-way interaction for the random intercept item (Picture). The random slopes of the fixed factors Position and Degree of Overlap were added to the random intercept subject (Subject). The structure of the final model for L2 speaker was the same as that of L1 speakers, with the exception that the random slope of the fixed factor Degree of Overlap was not added as random slope to subject (Subject). Figure 5 below depicts the observed reaction times for L1 speakers (upper panel) and L2 speakers (lower panel) as a function of Position and Degree of Overlap.

[Insert Figure 5 around here]

As shown in Figure 5, participants responded significantly faster to trials where the phoneme was positioned in onset position of the picture name than where it was placed in coda position. This was true for both L1 speakers (χ^2 (1) = 105.60, $p < .001$) and L2 speakers (χ^2 (1) = 53.41, $p < .001$). There was also a main effect of Degree of Overlap in both groups (L1: χ^2 (2) = 48.94, $p < .001$, L2: χ^2 (2) = 24.05, $p < .001$). A strong learning effect was also seen in both monolinguals (χ^2 (1) = 167.19, $p < .001$) and bilinguals (χ^2 (1) = 193.75, $p < .001$) as there was a main effect of Trial Number. The interaction effect was not significant either the mono- or bilingual group (p-values > .1).

Separate analyses per position. As Figure 5 shows a trend that the difference between double and single overlap is descriptively larger in L1 speakers than L2 speakers, we conducted separate analyses per position (one analysis for the onset data and one for the coda data). Hence, the package 'lsmeans' was used to focus on differences between overlap categories within a particular position. In the onset, the contrast between no overlap (no) and double overlap (do) as well as no overlap and single overlap (so) was significant for both L1 and L2 speakers (L1 do vs. no: β = -91.32, *SE* = 16.85, t = -5.42, $p < .001$ / L1 no vs. so: β = 52.06, *SE* = 15.87, t = 3.28, p = .007 / L2 no vs. do: β = -83.74, *SE* = 19.79, t = -4.23, $p < .001$ / L2 no vs. so: β = 62.29, *SE* = 21.51, t = 2.90, p = .02). Importantly, a significant difference was seen for the contrast between single and double overlap but only for the L1 speakers (β = -39.26, *SE* = 11.45, t = -3.43, p = .003). In the coda, there was only a significant difference between the no overlap and single overlap condition for L1 speakers (β = 60.95, *SE* = 16.49, t = 3.70, p = .002). No significant differences were found for L2 speakers.

Accuracy. Fixed factors that were included in the final generalized linear mixed effects model of L1 and L2 speakers were Degree of Overlap and Position. Interactions of these fixed factors were added to the models. The maximal random slope model did not converge for either the L1 or L2 data set and we therefore used the forward selection procedure. In the L1 model, Position was added as random slope to subject (Subject) while the fixed factor Degree of Overlap and the interaction of Degree of Overlap and Position were added as random slopes to item (Picture). In the L2 model, the random slope of Position was added to both subject (Subject) and item (Picture). Type II Wald Chi square tests were used to determine significance of main and interaction effects.

[Insert Figure 6 around here]

Generalized linear mixed effects model. Figure 6 illustrates that participants were more accurate if the target was situated in onset than coda position. Indeed, the effect of Position was significant for both L1 (χ^2 (1) = 57.38, p < .001) and L2 speakers (χ^2 (1) = 15.54, p < .001). The factor Degree of Overlap also reached significance for L1 (χ^2 (2) = 33.77, p < .001) and L2 speakers (χ^2 (2) = 21.29, p < .001). Additionally, there was a significant interaction effect of Position and Degree of Overlap in both L1 (χ^2 (2) = 6.21, p = .04) and L2 (χ^2 (2) = 8.98, p = .01) indicating that the differences in accuracy between overlap categories is larger in the coda than the onset position.

Separate analyses per position. As with reaction times, potentially significant differences between contrasts were measured. In the onset, significant differences were found between no overlap and double overlap for both L1 and L2 speakers (L1 do vs.

no: *β* = 0.78, *SE* = 0.22, z = 3.53, *p* = .001 / L2 no vs. do: *β* = 0.74, *SE* = 0.19, *z* = 3.83, *p* < .001) in which participants were more accurate in the double than the no overlap category. The L2 data set also revealed a significant difference between no overlap and single overlap (*β* = -0.43, *SE* = 0.18, *z* = -2.37, *p* = .047). In the coda, there was a significant difference between no overlap and double overlap and no overlap and single overlap for L1 speakers (do vs. no: *β* = 0.49, *SE* = 0.18, z = 2.64, *p* = .02 / L1 no vs. so: *β* = -0.59, *SE* = 0.14, z = -4.24, *p* < .001). L2 speakers, however, showed a significant difference between no overlap and single overlap (*β* = -0.61, *SE* = 0.16, z = -3.81, *p* < .001) and between double overlap and single overlap (*β* = -0.43, *SE* = 0.16, *z* = -2.66, *p* = .02).

Analysis Language x Task interaction

A final analysis was performed to further support the notion that the L2 disadvantage found in picture naming is not found during phoneme monitoring. In order to strengthen this claim, we observed whether an interaction between language groups and the tasks reached significance. For this particular analysis, we combined the two data sets and made a new linear mixed effects model. The fixed factors in this model were Language, Position, Degree of Overlap, and Task. Interactions of all fixed factors were added to the model. Once again, we applied the maximal random effects approach (which did not yield a model that converged) and used forward modelling to determine the final model. However, we did not include the four-way interaction for random slope determination, as the model would otherwise take days or even weeks to run (and most likely not converge). Moreover, this interaction was theoretically almost impossible to interpret.

After forward modelling, the final model contained the interaction of Degree of Overlap and Task as random slope for item (Picture) while no random slopes were added to subject (Subject). Type II Wald Chi square tests were used to determine significance. The interaction between Language and Task was indeed significant (χ^2 (1) = 4.62, *p* = .03).

Discussion

Experiment 2 demonstrated a clear effect of Position, which entails that participants responded more quickly when the target phoneme occurred in the onset than in the coda position of the picture name. This result is consistent with findings of Wheeldon and Levelt (1995) who also found an effect of phoneme position on reaction time. Additionally, participants were faster in the overlap category where both phonemes in onset and coda position overlapped (double overlap) and where only one phoneme overlapped (single overlap) compared to the category without any overlapping phonemes (no overlap). That is to say, phonological overlap facilitates the speech planning process, which is in line with what we found in Experiment 1. This suggests that the phoneme monitoring task follows the time course of phonological planning, supporting the assumption that these reaction times can be used to compare this planning stage in the different groups. The interaction effect of Degree of Overlap and Position shows that the facilitation effect is stronger in the onset position than the coda position. Furthermore, contrast analyses testing for both onset and coda position showed that there was a significant difference between no overlap and the other two categories. Yet, only L1 speakers responded faster to the double overlap category than the single overlap category in the onset position. Finally, accuracy scores were largely consistent with the reaction time data: the longer the reaction time, the higher the chance of a wrong answer.

The combined L1/L2 analyses allowed us to see whether the same effects arose when taking both data sets together (verifying the strength of the effects) and most importantly whether phoneme monitoring is slowed down in L2. The pattern of results was indeed similar to those obtained in the separate analyses for each language. Most importantly, no main effect of Language was found for reaction times. Moreover, model comparison showed that Language did not improve the model fit. Thus, L2 speakers are not significantly slower at phoneme monitoring than L1 speakers, suggesting that any L2 disadvantage in word production happens downstream from lexical and phonological planning processes (see below). This is confirmed by the significant interaction of Language and Task. Unexpectedly, language was a significant factor when considering accuracy scores in that L2 speakers were more accurate in the coda position than L1 speakers. This might be explained by arguing that L2 speakers benefit more from the distractor words if there is phonological overlap while less interference is seen when there is no overlap. This is consistent with weaker L2 lexical representations.

Contrast comparisons showed that L1 speakers responded faster to the double than the single overlap category in onset position. However, L2 speakers show no difference in reaction time between single and double overlap in the onset position. Further evidence for the claim that picture naming and phoneme monitoring tap into the same processes is the finding that both L1 and L2 speakers reacted faster to target phonemes in the onset position than in the coda. As discussed in more detail below, a possible explanation for the double/single overlap effect in L1 is that L1 and L2 speakers show a difference in the amount of feedback between the word and phoneme level. If L2 speakers have less feedback of activation (or weaker activation spreading) between the word and phoneme level, this might result in an absence of such a difference.

General Discussion

This study is the first to systematically compare the PWI task and phoneme monitoring task using the same pictures, allowing us to ascertain potential differences in earlier stages of L1 and L2 speech production. Specifically, we asked from which processing level the slow-down that is typically seen in L2 speakers during speech production originates (Gollan, Montoya, Cera, & Sandoval, 2008; Starreveld, de Groot, Rossmark, & van Hell, 2014). Before this question could be answered, we first needed to verify that there is indeed an L2 disadvantage during picture naming in this population and with these stimuli. Experiment 1 revealed a delay of 102 ms for L2 speakers compared to L1 speakers. In Experiment 2, we asked participants to perform a phoneme monitoring task in order to pinpoint the cause of the L2 delay found in Experiment 1. This task was used here as a measure of the speed of lexical retrieval and phonological encoding. Most importantly, this time we did not observe a significant difference as the difference in reaction times between L1 and L2 speakers amounted to only 9 ms. This suggests that the L2 delay observed in Experiment 1 is not located in any of the processes that the naming and monitoring tasks have in common.

Turning to theoretical implications, the absence of the language effect in the monitoring task cannot be explained by arguing that the distractors make naming the pictures easier as we found an L2 delay in the picture naming task. Moreover, the no overlap category also rules out this possibility. Additionally, the absence of a reaction time difference is unlikely to be a result of lack of experimental sensitivity as the position of the target phoneme very clearly modulates reaction times in both L1 and L2. In fact, every single analysis of the phoneme monitoring tasks has shown that the position of the target phoneme in the picture name is of paramount importance: participants reacted faster in both L1 and L2 when the target phoneme was placed in onset position than when it was positioned at the coda. This L2 finding is in line with the monolingual findings of Wheeldon and Levelt (1995) who found that assignment of the initial phoneme of the first syllable preceded assignment of the initial phoneme of the second syllable, regardless of word stress.

The number of overlapping phonemes also influences reaction times as trials with overlapping phonemes between the picture name and distractor word yielded significantly faster reaction times than if no phonemes overlapped. Interestingly, in the onset position L1 speakers responded faster in the double overlap category than the single overlap category. This is not observed in the L2 speakers and suggests that there is more feedback between the word and phoneme level in monolingual L1 speakers than in bilingual L2 speakers (see below). As for the coda position, the difference between double overlap and the other categories is larger for L1 speakers than L2 speakers. The facilitation effect (as well as the position effect) are evidence for the notion that the phoneme monitoring task taps into processes of speech planning.

For the monitoring tasks, we hypothesized that the reaction times would be shorter if the target phoneme was positioned at the onset of the picture name as opposed to the coda. Moreover, we predicted that in both the picture naming and monitoring tasks, the amount of phonological overlap would modulate reaction times in such a way that participants would be faster if more phonemes between the picture name and distractor word would match. Both hypotheses have been confirmed as reaction times were shorter for onset position and when phonemes overlapped. According to hypotheses that argue for a slow-down in lexical retrieval and phonological encoding, L2 speakers should be slower than L1 speakers. Importantly, we did not observe a language effect in that L2 speakers were not significantly slower than L1 speakers in the phoneme monitoring task. This suggests that the speed of speech planning (at least up through phonological encoding) might not be so different between monolingual L1 and bilingual L2 speakers, even when the latter are unbalanced bilinguals that live in a strongly L1-dominant environment. Yet, we did not find evidence for the claim that facilitation effects due to phonological overlap were stronger for L1 speakers than L2 speakers. We found no significant interaction effects between Language and Degree of Overlap.

The lack of a language effect in monitoring speed does not support hypotheses which claim that earlier stages of speech planning in bilinguals are slower. This finding suggests that the slow-down that is typically seen in bilinguals during picture naming might be situated at the post-phonological stage of speech production, namely articulation. Indefrey and Levelt (2004) performed a meta-analysis of several studies that focus on the time course of the process of word production and that map this process onto brain areas. According to the time course analysis, the retrieval of the lemma takes somewhere between 150 and 225 ms., while articulatory planning takes between 217 and 530 ms. This suggests that articulatory processes take up much more time than lemma retrieval, indicating that there might be a larger chance for a potential slow-down to be situated at the articulatory stage. Moreover, any difference in the time course of lemma retrieval between L1 and L2 might simply be too small to be observable since the lemma is already retrieved rather quickly, which might explain why no differences were found in monitoring times. During L2 speech production, however, a different phonemic inventory has to be activated. This change might explain the L2 disadvantage during speech production.

On the one hand, Simmonds et al. (2011) argue that difficulties in L2 speech production originate from articulation instead of phonological encoding. They argue that the most difficult aspect of L2 production is the accent with which it is pronounced. L2 speakers who learn their L2 after adolescence almost always maintain a non-native accent, which is nearly impossible to correct. On the other hand, studies that show evidence for the weaker-links hypothesis (Gollan et al., 2008; Kroll & Stewart, 1994; Starreveld et al., 2014) claim that earlier processes of speech production are delayed. Yet, these are all based on experiments in which a picture naming task was used. In these instances, L2 disadvantages are found for speech production where the slow-down is explained by arguing that speech planning up through phonological planning is slower in L2 than L1 speakers. However, we did not find evidence for differences between L1 and L2 speakers in earlier stages of speech production, although we do not deny that L2 speakers might have trouble during lexical retrieval (see Gollan et al. 2001).

Finally, contrasts comparisons revealed that the single and double overlap category significantly differ in the onset position in the L1 but not the L2 speakers (although descriptively the latter group showed the same pattern). We suggest the following explanation. When the participants see a phonologically related distractor word (e.g., *bed*) this pre-activates the overlapping phonemes (/b/ for target *bag*), facilitating production of those phonemes. But as is clear from the picture-word interference task (Experiment 1), an end-related distractor word (e.g., *rug*) facilitates the naming latency too, even though the word-beginning was not primed. This suggests that part of the phonological facilitation effect is caused by a further mechanism, possibly one involving lexical representations. On that account, the distractor's phonemes partially activate the target's lexical representation (i.e., phoneme-to-word form feedback, as assumed in Dell, 1986) and this would be true for both beginning-related and end-related phonemes. As the target word would have a higher activation level, the process of spelling out the phonemes can be speeded up. This explains why there is more facilitation in the double than single overlap category, both in the PWI data (Experiment 1) and in the phoneme monitoring data for the onsets (Experiment 2). The reason why this facilitation is not seen in the coda position is that the monitoring process takes longer to reach the coda of the word, allowing it to catch up for the delay in a less related vs. more related category. A possible explanation for why the gradual facilitation effect is not reliable in L2 is that the amount of feedback between the word and phoneme level might be somewhat smaller in L2 speakers than in L1 speakers. Even though the distractor word has the onset and coda phoneme in common with the picture name, the coda phoneme does not send (enough) activation to the word level. This in turn means that the word level does not send this information back to the phoneme level efficiently enough to make a difference in reaction time.

One potential limitation of the current study is that the target phonemes that were monitored coexisted with overlapping phonemes of phonologically related distractor words. This might have affected the response latencies in such a way that trials with phonologically related distractor words might inherently be reacted to faster than trials that have phonologically unrelated distractor words. The minor differences between the naming task and phoneme monitoring task might be explained by this discrepancy. Be that as it may, there was still a main effect of Degree of Overlap in the naming task. Moreover, both the position effect and the overlap effect are robust in that they were significant in all analyses of the monitoring tasks. Hence, it is unlikely that this inconsistency would have greatly affected the results and it would certainly not be able to account for the lack of a main effect of Language during monitoring.

The final limitation that needs to be discussed pertains to the nature of the participant groups. In particular, the comparison between the picture naming task and the phoneme monitoring task was based on two different participant groups, as the same speakers did not perform both tasks. However, the results of the questionnaires filled out by bilinguals are very similar between tasks. Additionally, it is highly questionable whether using the same participants for both tasks would yield substantially different results than our experiments. Future experiments using the same participants might be conducted to verify this claim.

Conclusion

We confirmed that there is an L2 delay during picture naming in a picture-word interference paradigm. Moreover, results revealed that the speech monitoring process is sequential. The observed phonological facilitation effects show that the picture-word interference paradigm taps into lexical retrieval and phonological encoding.

Nevertheless, we have not found a difference in phoneme monitoring speed between L1 and L2 speakers, which is not consistent with the hypothesis that the slow-down of L2 speech production is situated at earlier speech planning stages. The lack of a language effect can alternatively be explained by a hypothesis that argues for articulatory delay during speech production.

Acknowledgements

We would like to thank Michael Stevens and Toru Hitomi for their advice on the statistical analyses, specifically on creating and fitting the linear mixed effects models. This paper received funding from the special research fund of Ghent University (GOA - Concerted Research Action BOF13/GOA/032). We would also like to thank the members of the GOA-project for their input and suggestions.

References

Alario, F. X., Goslin, J., Michel, V., and Laganaro, M. (2010). The functional origin of the foreign accent. *Psychological Science*, *21*, 15-20. doi: 10.1177/0956797609354725

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. http://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: http://dx.doi.org/10.1016/j/jml/2012.11.001

Boersma, Paul & Weenink, David (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.29, retrieved 24 May 2017 from http://www.praat.org/

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

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological review*, *93*(3), 283. http://dx.doi.org/10.1037/0033-295X.93.3.283

Finkbeiner, M., Forster, K., Nicol, J., & Nakamura, K. (2004). The role of polysemy in masked semantic and translation priming. *Journal of Memory and Language*, *51*(1), 1-22. doi: 10.1016/j.jml.2004.01.004

Ganushchak, L. Y., & Schiller, N. O. (2008). Brain Error–Monitoring Activity is Affected by Semantic Relatedness: An Event-related Brain Potentials Study. *Journal of Cognitive Neuroscience, 20*(5), 927-940. doi: 10.1162/jocn.2008.20514

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: 10.3758/BF03193224

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: 10.1016/j.jml.2007.07.001

Gollan, T. H., & Silverberg, N. B. (2001). Tip-of-the-tongue states in Hebrew–English bilinguals. *Bilingualism: language and cognition, 4*(01), 63-83. doi: 10.1017/s136672890100013x

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

Hanulová, J., Davidson, D. J., & Indefrey, P. (2008). The time course of word-form encoding in second language word production: An ERP study. In *Poster presented at the 5th International Workshop on Language Production, Annapolis, Maryland*.

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: 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: 10.1016/j.cognition.2002.06.001

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

Jodo, E., & Kayama, Y. (1992). Relation of a negative ERP component to response inhibition in a Go/No-go task. *Electroencephalography and clinical neurophysiology*, *82*(6), 477-482. doi: https://doi.org/10.1016/0013-4694(92)90054-L

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

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences, 22*(01), 1-38. doi: 10.1017/s0140525x99001776

Levy, R. (2014). Using R formulae to test for main effects in the presence of higher-order interactions. *arXiv preprint arXiv:1405.2094*.

Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: effects of stimulus onset asynchrony and types of interfering stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(6), 1146. doi: 10.1037/0278-7393.17.6.1146

Özdemir, R., Roelofs, A., & Levelt, W. J. M. (2007). Perceptual uniqueness point effects in monitoring internal speech. *Cognition, 105*(2), 457-465. doi: 10.1016/j.cognition.2006.10.006

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

Poulisse, N. (2000). Slips of the tongue in first and second language production. *Studia linguistica, 54*(2), 136-149. doi: 10.1017/s002222670100888x

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/

Roux, F., Armstrong, B. C., & Carreiras, M. (2016). Chronset: An automated tool for detecting speech onset. *Behavior Research Methods*, 1-18. doi: 10.3758/s13428-016-0830-1

Schmitt, B. M., Münte, T. F., & Kutas, M. (2000). Electrophysiological estimates of the time course of semantic and phonological encoding during implicit picture naming. *Psychophysiology, 37*(4), 473-484. doi: 10.1111/1469-8986.3740473

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.

Simmonds, A. J., Wise, R. J., & Leech, R. (2011). Two tongues, one brain: imaging bilingual speech production. *Frontiers in Psychology, 2*, 166. doi: http://dx.doi.org/10.3389/fpsyg.2011.00166

Starreveld, P. A., de Groot, A. M., Rossmark, B. M., & Van Hell, J. G. (2014). Parallel language activation during word processing in bilinguals: Evidence from word production in sentence context. *Bilingualism: Language and Cognition, 17*(02), 258-276. http://dx.doi.org/10.1017/S1366728913000308

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

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological bulletin, 128*(3), 442. doi: 10.1037/0033-2909.128.3.442

Wheeldon, L. R., & Levelt, W. J. (1995). Monitoring the time course of phonological encoding. *Journal of memory and language*, *34*(3), 311-334. doi:10.1006/jmla.1995.1014

Appendix A: Target Pictures

tail

pan

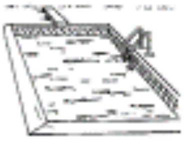
sun

pear

top

pen

saw



pool

rug

net

pot

nut

log

fan

mop

heel

hat

moon

deer

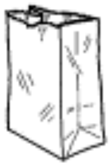
dog

bat

bug

beak

boot



bag

Appendix B: Target Trials

Double overlap - yes

Onset:

bag	bug	b
dog	dig	d
top	tip	t
rug	rag	r
net	not	n
saw	sew	s
beak	book	b
pear	poor	p
fan	fun	f
moon	mean	m
tail	tool	t
bat	but	b
deer	door	d
heel	hail	h
nut	net	n
pot	pit	p
boot	beat	b
pan	pin	p
hat	hit	h
pen	pan	p
sun	sin	s
pool	peel	p
mop	map	m
log	leg	l
bug	big	b

Coda:

bag	bug	g
pear	poor	r
moon	mean	n
tail	tool	l
top	tip	p
deer	door	r
nut	net	t
saw	sew	w
bat	but	t
rug	rag	g
heel	hail	l
net	not	t
fan	fun	n
dog	dig	g
beak	book	k
sun	sin	n
log	leg	g
pool	peel	l
mop	map	p
bug	big	g
boot	beat	t
pan	pin	n
hat	hit	t
pot	pit	t
pen	pan	n

Single overlap - yes

Onset:

bug	bow	b
pool	peak	p
log	lap	l
sun	sad	s
mop	mat	m
hat	hip	h
pot	pub	p
boot	bean	b
pan	pet	p
pen	paw	p
bag	bet	b
dog	dip	d
top	tar	t
rug	red	r
net	nap	n
pear	pool	p
fan	fit	f
beak	boot	b
moon	meal	m
saw	set	s
bat	beg	b
heel	hood	h
tail	tour	t
nut	new	n
deer	doom	d

Coda:

beak	took	k
pear	sour	r
fan	pen	n
saw	now	w
moon	pain	n
top	rap	p
bag	fog	g
rug	leg	g
net	hut	t
dog	hug	g
deer	fair	r
heel	soul	l
nut	rat	t
tail	fool	l
bat	wet	t
hat	get	t
pot	let	t
pen	bun	n
pan	son	n
boot	seat	t
bug	jog	g
mop	gap	p
log	tag	g
pool	deal	l
sun	van	n

No overlap - yes

Onset:

bag	rod	b
net	big	n
dog	jar	d
top	wig	t
rug	mow	r
heel	food	h
tail	seek	t
nut	rim	n
deer	soup	d
bat	fur	b
pool	tear	p
mop	war	m
bug	mad	b
log	run	l
sun	kid	s
pan	sit	p
boot	hair	b
pot	law	p
pen	fat	p
hat	gun	h
pear	hook	p
saw	bet	s
moon	leaf	m
beak	root	b
fan	pig	f

Coda:

sun	kid	n
bug	mad	g
mop	war	p
log	run	g
pool	tear	l
boot	hair	t
hat	gun	t
pan	sit	n
pot	law	t
pen	fat	n
bag	rod	g
top	wig	p
rug	mow	g
dog	jar	g
net	big	t
nut	rim	t
tail	seek	l
deer	soup	r
bat	fur	t
heel	food	l
saw	bet	w
pear	hook	r
moon	leaf	n
fan	pig	n
beak	root	k

Appendix C: Summary Tables ANOVA F1/F2 Analyses

Picture naming:

Reaction Times

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Language	F1: 1, 81 F2: 1, 144	F1: 20.07 F2: 284.80	F1: < .001 F2: < .001
Degree of Overlap	F1: 2, 162 F2: 2, 48	F1: 37.60 F2: 6.08	F1 : < .001 F2: = .004
Position	F1: 1, 81 F2: 1, 24	F1: 1.58 F2: 0.68	F1: = .21 F2: = .42
Language:DegreeofOverlap	F1: 2, 162 F2: 2, 144	F1: 2.06 F2: 0.47	F1: = .13 F2: = .63
Language:Position	F1: 1, 81 F2: 1, 144	F1: 0.09 F2: 0.17	F1: = .77 F2: = .69
Position:DegreeofOverlap	F1: 2, 162 F2: 2, 48	F1: 1.05 F2: 0.26	F1: = .35 F2: = .77
Language:DegreeofOverlap:Position	F1: 2, 162 F2: 2, 144	F1: 0.44 F2: 0.20	F1: = .65 F2: = .82

Accuracy

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Language	F1: 1, 81 F2: 1, 144	F1: 5.43 F2: 12.31	F1: = .02 F2: < .001
Degree of Overlap	F1: 2, 162 F2: 2, 48	F1: 0.32 F2: 0.14	F1: = .73 F2: = .87
Position	F1: 1, 81 F2: 1, 24	F1: 1.27 F2: 0.51	F1: = .26 F2: = .48
Language:DegreeofOverlap	F1: 2, 162 F2: 2, 144	F1: 0.03 F2: 0.003	F1: = .97 F2: = .997
Language:Position	F1: 1, 81 F2: 1, 144	F1: 0.15 F2: 0.03	F1: = .70 F2: = .86
Position:DegreeofOverlap	F1: 2, 162 F2: 2, 48	F1: 6.81 F2: 1.41	F1: = .001 F2: = .25
Language:DegreeofOverlap:Position	F1: 2, 162 F2: 2, 144	F1: 3.14 F2: 0.71	F1: = .046 F2: = .49

L1 monitoring:

Reaction Times

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Degree of Overlap	F1: 2, 106 F2: 2, 48	F1: 27.43 F2: 6.69	F1: < .001 F2: = .003
Position	F1: 1, 53 F2: 1, 24	F1: 160.3 F2: 75.75	F1: < .001 F2: < .001
Position:DegreeofOverlap	F1: 2, 106 F2: 2, 48	F1: 4.84 F2: 1.98	F1: = .01 F2: = .15

Accuracy

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Degree of Overlap	F1: 2, 106 F2: 2, 48	F1: 23.74 F2: 13.15	F1: < .001 F2: < .001
Position	F1: 1, 53 F2: 1, 24	F1: 64.28 F2: 51.68	F1: < .001 F2: < .001
Position:DegreeofOverlap	F1: 2, 106 F2: 2, 48	F1: 7.39 F2: 3.48	F1: = .001 F2: = .04

L2 monitoring:

Reaction Times

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Degree of Overlap	F1: 2, 76 F2: 2, 48	F1: 20.82 F2: 2.68	F1: < .001 F2: = .08
Position	F1: 1, 38 F2: 1, 24	F1: 72.98 F2: 46.09	F1: < .001 F2: < .001
Position:DegreeofOverlap	F1: 2, 76 F2: 2, 48	F1: 2.64 F2: 0.69	F1: = .08 F2: = .51

Accuracy

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Degree of Overlap	F1: 2, 76 F2: 2, 48	F1: 9.55 F2: 10.07	F1: < .001 F2: < .001
Position	F1: 1, 38 F2: 1, 24	F1: 15.04 F2: 10.76	F1: < .001 F2: = .003
Position:DegreeofOverlap	F1: 2, 76 F2: 2, 48	F1: 5.50 F2: 3.03	F1: = .006 F2: = .06

L1 and L2 monitoring combined:

Reaction Times

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Language	F1: 1, 91 F2: 1, 144	F1: 0.11 F2: 3.77	F1: = .75 F2: = .054
Degree of Overlap	F1: 2, 182 F2: 2, 48	F1: 47.96 F2: 4.69	F1: < .001 F2: = .01
Position	F1: 1, 91 F2: 1, 24	F1: 227.78 F2: 66.48	F1: < .001 F2: < .001
Language:DegreeofOverlap	F1: 2, 182 F2: 2, 144	F1: 0.15 F2: 0.31	F1: = .86 F2: = .74
Language:Position	F1: 1, 91 F2: 1, 144	F1: 1.07 F2: 8.22	F1: = .30 F2: = .005
Position:DegreeofOverlap	F1: 2, 182 F2: 2, 48	F1: 6.35 F2: 1.10	F1: = .002 F2: = .34
Language:DegreeofOverlap:Position	F1: 2, 182 F2: 2, 144	F1: 0.99 F2: 0.93	F1: = .37 F2: = .40

Accuracy

Effect	**Degrees of Freedom**	**F-value**	**P-value**
Language	F1: 1, 91 F2: 1, 144	F1: 9.70 F2: 37.02	F1: = .002 F2: = .002
Degree of Overlap	F1: 2, 182 F2: 2, 48	F1: 32.58 F2: 15.66	F1: < .001 F2: < .001
Position	F1: 1, 91 F2: 1, 24	F1: 78.60 F2: 37.09	F1: < .001 F2: < .001
Language:DegreeofOverlap	F1: 2, 182 F2: 2, 144	F1: 1.77 F2: 0.92	F1: = .17 F2: = .40
Language:Position	F1: 1, 91 F2: 1, 144	F1: 12.26 F2: 14.10	F1: < .001 F2: < .001
Position:DegreeofOverlap	F1: 2, 182 F2: 2, 48	F1: 11.46 F2: 3.60	F1: < .001 F2: = .03
Language:DegreeofOverlap:Position	F1: 2, 182 F2: 2, 144	F1: 1.37 F2: 0.54	F1: = .26 F2: = .58