Neural processes underlying cognitive control during language production

Tara Pirnia

March 10, 2023

Joint PNC-MLD program
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Leila Wehbe *co-chair*
Nazbanou Nozari *co-chair*
Stephanie Ries-Cornou
Graham Neubig
Paula Clemens

Contents

1 Introduction

1.1 Abstract

Processing external context and producing target responses in conversation requires coordinated monitoring and control processes. For example, when saying, “Please pass me the salt and mustard,” overriding the prepotent “pepper” in favor of the less potent “mustard” requires cognitive control. We study neural activity underlying control processes through experimental manipulations that require the (a) suppression of prepotent word representations (Stroop-like and Picture-word interference tasks) or (b) selection when there is co-activation of word representations (blocked-cyclic picture-naming task) to achieve task goals. We first establish the ability of statistical learning to map participants’ neural responses to condition-related cognitive states. The proposed thesis characterizes the underlying neural processes by leveraging statistical learning methods, neural electrophysiology, and neuroimaging recordings in conjunction with experimental manipulations while participants complete each task. We propose experimental methods that characterize the shared neural features across participants, between specified language processes, and in addition, between language and non-language control mechanisms. Our overall aim is two-fold, (1) to advance theories of cognitive control by the systematic mapping of neural activity to psycholinguistic variables with established experimental paradigms shown to elicit control; (2) to test the central assumptions imposed by theoretical frameworks of word production by examining the temporal dynamics of stages of production.

1.2 Overview

At each point in time, a speaker has various options for what to produce and how to produce it. For example, you can say “I enjoyed the food” or “The pasta was delicious.” Owing to years of research in language production, we know that words compete for production, especially if they are similar, e.g., when you say “pasta,” “panini” competes for production, and if competition is not resolved successfully, production ends in an error. It is often thought that speakers resolve such competition by applying inhibitory control, a process whereby the competitors are suppressed in favor of the target word. But what is the nature of this control? Is it using the same process that speakers apply when they intentionally wish to suppress an utterance that is on their mind? Imagine the pasta tasted awful, but you wished to be polite. In that situation, you willfully apply cognitive control to override the production of the prepotent word “awful” in favor of “good”. While we know that speakers can do both, i.e., suppress prepotent words and resolve competition induced by similar words, we know little about the nature of the control processes underlying these two situations.

Broadly, this work looks at how the stages of word production interact in time and space to inform existing theoretical frameworks of language production. We aim to advance theories of cognitive control to characterize the underlying neural processes using machine-learning techniques, applied to EEG and functional neuroimaging recorded during well-established experimental paradigms. Specifically, we characterize the neural features within and across participants, between specified language processes, and in addition, between language and non-language control mechanisms during high and low-conflict conditions. We examine the temporal dynamics and spatial location by systematic mapping of the neural activity to psycholinguistic variables to delineate the relationship between stages of word-production.

1.3 Thesis Aims

Aim I. We examine cognitive control in language production during a Stroop-like task. Such tasks require the suppression of prepotent words via cognitive control to successfully accomplish the task's goals. First, we tested the feasibility of using statistical learning methods to study the neural processes underlying this phenomenon from electrophysiological recordings. Once established, we implement these tools to characterize the temporal dynamics of task-related control processes. Subsequently, we ask if there is something more general about these states that is common to all humans by examining the temporal variability and generalizable neural features between study participants.

Aim II. To understand more broadly the neural processes of control within language, we study how conflict is induced in different language production processes. Here, we look to conflict states because of contextual similarity. We first characterize the temporal dynamics of high- and low-conflict states due to the co-activation of semantic and phonologically similar word representations by implementing the methods described in Aim I. Next, we focus on the degree to which high-conflict states are generalizable within language production. Specifically, we address (1) whether the observed neural signatures of conflict in semantically similar tasks are generalizable to those of phonologically similar conditions (2) if the decoded features from contextual similarity conditions are generalizable to conflict states decoded during Stroop-like tasks through a series of cross-classification procedures.

Aim III. We examine the generality or specificity of high versus low conflict states between linguistic and non-linguistic domains to understand how the control recruited during language production aligns with the broader control literature. We present a second experimental data set in which participants complete two interleaved experimental tasks that elicit control in language production and visual-motor domains independently. The proposed work leverages the high spatial resolution of a different neuroimaging modality, fMRI recordings, to answer whether and to what extent are the neural correlates of control shared in language and visual-motor domains. We describe the planned whole-brain and searchlight analysis to localize control within and between the two domains.

2 Approach

2.1 Empirical Data

2.1.1 Picture-Stroop paradigm

Participants completed a blocked-cyclic picture-naming task where they named a small set of pictures that are presented in blocks of repeated picture pairs (Belke et al., 2005; Schnur et al., 2009; Nozari et al., 2016). Across the paradigm, a six-way cross with two Stroop-like conditions and three contextual conditions. The Stroop-like conditions consisted of half-blocks where participants saw one of two pictures (e.g., cake/pie) and either named them by their canonical names (congruent naming trials) or by the name of the other picture (incongruent naming trials). Contextual conditions were presented block-wise, where picture pairs presented in each block were either semantically related (e.g., cake/pie), phonologically related (e.g., tie/pie), or unrelated (e.g., tie/rake). In total, ten unique pictures were presented in an interleaved fashion, with ten repetitions of each (Figure 2.1a).

2.1.2 The Congruency Sequence Effect Paradigm

In experimental paradigm 2, participants completed two cognitive tasks: (1) picture-word interference (PWI), a linguistic task (Schriefers et al., 1990), interleaved with (2) Simon task, a non-linguistic visuospatial task (Simon & Rudell, 1967). The interleaved tasks were structured such that 1-back and 2-back results allowed for across-domain and within-task examination of cognitive control, respectively (Freund & Nozari, 2018). Each of the two tasks included congruent and incongruent conditions and required the inhibition of a prepotent response (Hirschfeld et al., 2008; Schriefers et al., 1990). During each PWI trial, participants were presented with line drawing (picture) and a superimposed word and tasked with naming the picture (target). In congruent trials, the superimposed word corresponded with the picture name and, thus the target. In incongruent trials, the superimposed word did not match the picture name and thus acted as a distractor.

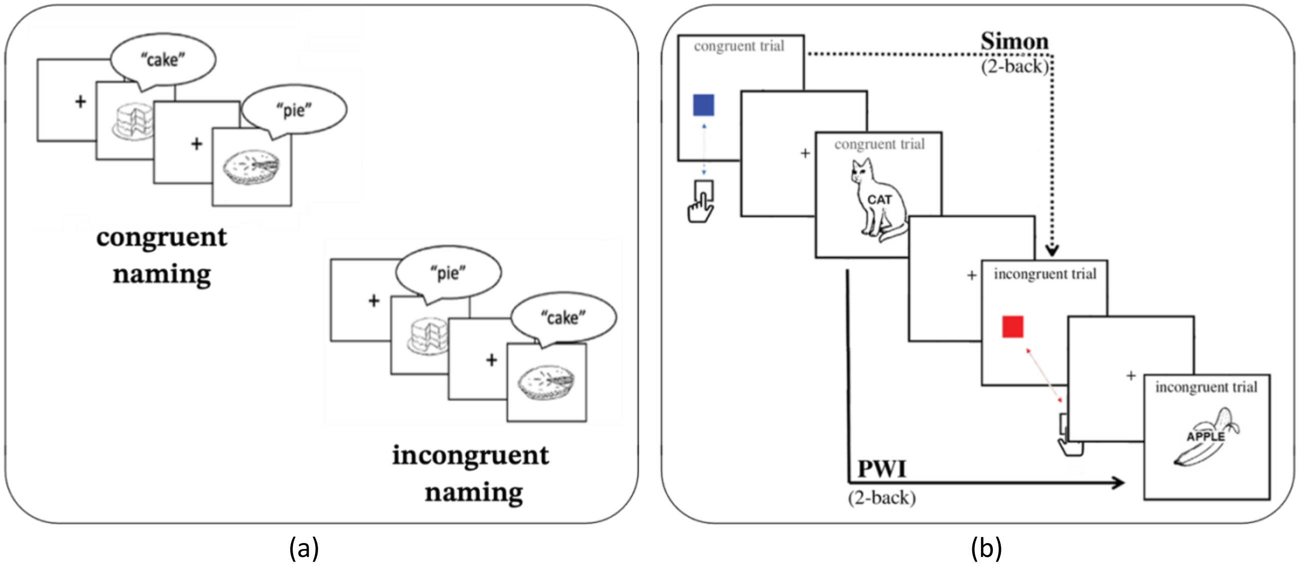


Figure 2.1 Experimental Paradigms (a) Blocked-cycling picture naming task, and (b) Picture-word interference and Simon task.

2.1.3 Electroencephalography

Aims 1 and 2 examine the neural activity recorded with electroencephalography (EEG), a non-invasive technique that measures electrical signals generated by neurons from electrodes attached to the scalp. Notably, EEG provides high-temporal resolution (milliseconds) measures of the dynamic neural signatures that are associated with the brain states of participants as they complete each trial. The following work uses a 128-channel EEG recorded as participants completed the Picture-Stroop task (2.1.1) with analysis completed on sensor space.

2.1.4 Functional magnetic resonance imaging (fMRI)

For aim 3, we utilize a second modality to record brain activity—functional magnetic resonance imaging (fMRI). fMRI is an imaging technique that measures changes in blood flow in the brain. As participants perform each task, the neurons require greater oxygen-rich blood, causing the magnetic properties detected by fMRI to change. While fMRI does not have the temporal resolution of EEG, it provides high-resolution spatial information on the changes in brain activity during task performance.

2.2 Analysis of neural activity

In the following work, we utilize machine learning classifiers to better decode the dynamics of neural activity underlying language production. Supervised classifiers provide a map from neural activity (e.g., EEG and fMRI) to particular task conditions that information on when and where stimulus-evoked responses occur in the brain. Importantly, the proposed thesis demonstrates how machine learning classifiers can directly test and, moreover, inform the theoretical frameworks in the study of language production. Specifically, we utilize four classifiers: linear ridge classifiers, Support Vector Machines (SVM), decision trees (DT), and logistic classifiers. Briefly, each of which provides a different method for classification—linear ridge classifiers use a linear model with regularization to reduce overfitting; SVM allows for a non-linear approach to separate data points; decision trees utilize a tree-like structure to make class decisions; and a logistic classifier uses a logistic function to model the probability of each condition. In the work below, we initially test each classifier's performance in decoding each experimental condition from the recorded neural activity and use the best-performing classifier to test subsequent hypotheses.

Cross-validation was implemented by splitting the data sets to ensure the independence of each model's training from its test performance. Specifically, we utilized a nested-cross-validation scheme to split each data set into a train, validation, and test set to train the model, tune the models' hyperparameters, and to evaluate the models' performance, respectively.

To measure the localization of temporal information, we used a sliding window analysis. Sliding window classification involves dividing the neural activity during each trial into a series of overlapping windows and then classifying each window's condition (Fyshe et al., 2019; King & Dehaene, 2014). Allows us to learn how information evolves over time by decoding distinct features across the duration of an EEG recording. By examining the generalizability of classifiers at each subsequent window of time, we can identify when pertinent information is observed and for how long it is sustained.

The performance of each classifier was measured by classification accuracy by comparing the predicted class labels with the actual class labels to calculate the accuracy score. The accuracy score was the ratio of the number of correctly classified instances to the total number of instances in the test data. For statistical analysis.

Significance was determined using a non-parametric, cluster-wise statistical test that accounts the relationship between adjacent time windows. For each model, cluster-size distributions under the null hypothesis were acquired through permutations for each subject. After thresholding classification accuracy of the permutation results at 55%, we defined clusters as accuracy values obtained along all possible train and test windows in from the temporal generalization matrices. The null distribution was determined by the maximum cluster size at each permutation (Maris & Oostenveld, 2007). Clusters were defined in the observed results as those accuracy results that were above 55% accuracy and at adjacent time train and test time windows. Cluster sizes of p-value <.05, determined by the null distribution were then determined as significant.

3 Cognitive Control during Production

3.1 Overview

A lot of times, we produce words that are expected and highly predictable; for example, we say, "Please pass me the salt and *pepper.*" But sometimes, we do not intend to produce the most predictable word, for example, I may need salt and *mustard*. Overriding the prepotent "pepper" in favor of the less potent "mustard" requires cognitive control.

In the chapter, we use a picture-naming paradigm to create situations where speakers produce a prepotent word (e.g., say "cat" when they see a cat) or, conversely, suppress the prepotent word in favor of a different word (e.g., say "dog" when they see a cat), and record their EEG data. Our assumption is that the cognitive state differs between these two conditions: the former is a low-conflict situation as the picture brings the prepotent label to mind, whereas the latter is a high-conflict situation because the evoked label is different from the target word. We then use machine-learning techniques to answer three questions:

(1) Can we EEG data to successfully distinguish between high- and low-conflict states in speakers? To answer this question, we compare the performance of three standard classifiers, a support vector machine (SVM), a decision tree (DT), and a regularized ridge classifier. (2) What is the timeline of overriding a prepotent word in favor or a less potent word? Serial theories of word production assume that lexical selection is done by 200-300 ms after viewing a picture (Indefrey, 2011; Indefrey & Levelt, 2004). Conversely, interactive theories of production predict lasting effects of lexical selection even during further steps of production (Dell, 1986; Pinet & Nozari, 2023; Ries et al., 2021). To adjudicate between these accounts, we examine the timeline of cognitive control measured by decodability using the optimal of the three classifiers in a sliding-window approach. (3) Finally, we test the generalizability of neural signatures of low vs. high conflict states across speakers. To do so, we do examine if conflict can be decoded from individual participants using models trained on all other participants. Moreover, we examine the dynamics of cross-participant classification in time as compared to within-participant classifier performance.

To summarize, the analyses proposed in this chapter will allow us to combine EEG data with machine learning techniques to identify the optimal classification method for identifying low vs. high conflict states within individual speakers, map out the timeline of such an effect, and test the generalizability of this effect across speakers.

3.2 Picture-naming paradigm

Using paired picture stimuli, participants completed a Stroop-like picture-naming task with (1) congruent low-conflict trials or (2) incongruent high-conflict trials (see Figure 2.1). In the low-conflict condition, participants are tasked with saying the word matching the picture presented. These trials are expected to require little cognitive control as the presented item is congruent with the task goal. During the high-conflict trials, participants are presented with the same pictures, but the target production is a non-matching word—one belonging to the alternate picture. Such trials are expected to require greater cognitive control as the target goal and the presented picture are both highly activated, increasing conflict.

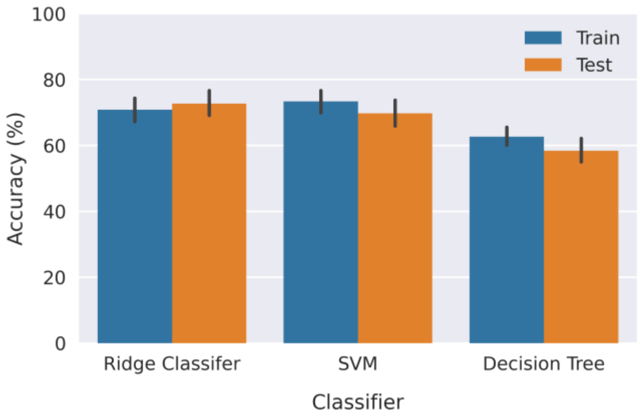


Figure 3.1 Comparison of prediction accuracy from three classifiers, Ridge classifier, SVM, and Decision tree. The training accuracy determined by the inner cross-validation is shown in blue. Test accuracy, based on held-out blocks, is shown in orange. (SVM=support vector machine)

Introducing these two conditions allows us to investigate production processes associated with conflict-associated cognitive control.

3.3 Control in high and low-conflict language production tasks

The neural signatures of low and high conflict states are different in the brain and can be robustly decoded via classical ML methods (i.e., phase conditions can be classified from EEG at above-chance accuracy).

Methods: We compared the performance of three classifiers in determining conflict—decoding congruent and incongruent conditions of Experiment 1. Each classifier was trained and evaluated on the average of 8 repetitions of EEG recordings spanning the full trial time window (cue to 1000ms). EEG trials and corresponding labels were split into train, validation, and test using a 5-fold nested cross-validation scheme, stratified by trial blocks, Accuracy is calculated and averaged per fold across participants. We measured model performance through classification accuracy for each test set. Significance was measured through permutation testing to determine the distribution under the null hypothesis and define significant confidence intervals per model.

Results: Of the three classifiers, ridge and SVM both performed significantly above chance (69.6%, SD=10.6 and 72.6%, SD=10.2, respectively), but DT classifier did not (53.8%, SD=11.6). Moreover, utilizing linear regression, the ridge classifier decoded conditions significantly better than both SVM and Decision Tree (see Figure 3.1).

Implications: Our results indicate that the neural signatures of conflict, as measured through simple classifiers, can decode congruent and incongruent trials with a ridge regression classifier performing optimally. Overall, the results imply that despite signal noise, the condition effects are robust at a group level and suitable for subsequent analyses.

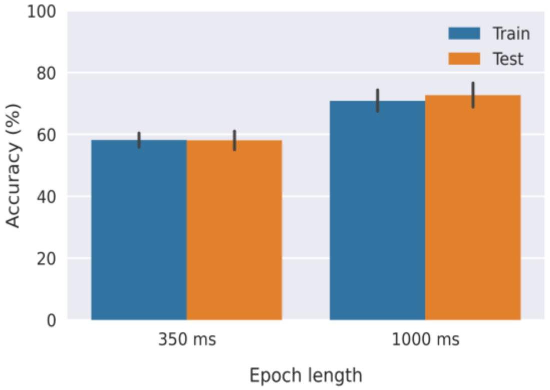


Figure 3.2 Comparison of Ridge classifier performance within an early epoch (350ms) vs. full trial (1000ms). The train accuracy is shown in blue. Test accuracy, based on held-out blocks, shown in

3.4 Temporal dynamics of conflict

When cognitive control occurs from the time of stimulus presentation and word production is not certain. Based on Indefrey and Levelt's (2004) model of language production, lexical selection happens at around 250-300ms after stimuli presentation. In a strictly serial model (Levelt et al., 1999), lexical selection eliminates other lexical competitors and their influence on further processes. If control resolves lexical conflict completely, as stipulated by serial models, we expect the classification peak around the lexical selection time window. If, on the other hand, the system operates more continuously, with the lexical effect persisting after lexical selection, we would expect better classification in a larger time window.

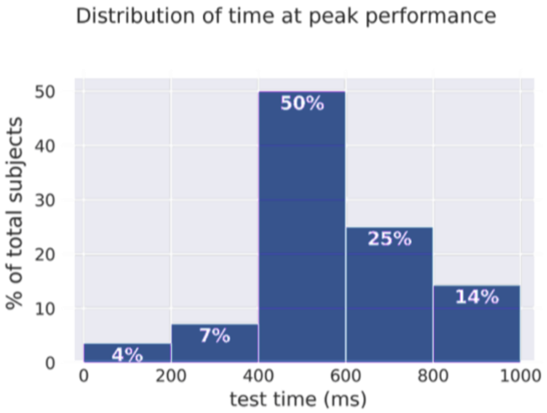


Figure 3.3 Percent distribution of participants' peak performance train-test time window. Train-test time window where the classifier performs optimally per participant is shown in histogram with 200ms bin width.

3.4.1 Group-level decoding of conflict states

To adjudicate between these two possibilities, we utilize methods from Section 3.3 to compare the decodability of cognitive control within only the lexical selection time window to that of classifiers observing the entire length of the trial. If the implementation of cognitive control is limited to the lexical selection period, classifier performance is not expected to improve when a longer time window is observed. In contrast, if control processes extend beyond the theoretical period, extending the observed time window will improve the decodability of control.

Methods: We compared the performance of ridge classifiers trained and tested on the full trial window (cue to 1000ms) to that of those trained and tested on expected lexical selection window (cue to 350ms extended beyond 300ms to account for response time variation).

Results: We found that the classification accuracy of models trained and tested on full trials was significantly higher when the entire time window was used, as opposed to an earlier window usually considered relevant to lexical selection (see Figure 3.2).

Implications: The results suggest that neural signatures of conflict at the group level, measured by classification of congruent and incongruent trials, are observed throughout picture-naming to word production timeline, extending beyond the expected "lexical selection" time window as described by Indefrey and Levelt (2004). Based on these results, we reject the hypothesis that all information attributing to high vs low conflict conditions is limited to the proposed lexical selection time window in every participant.

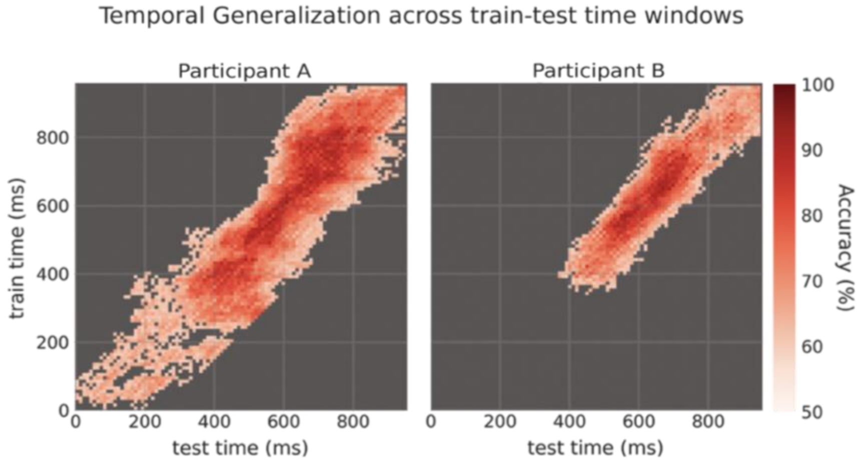


Figure 3.4 Temporal generalization plots in example participants, A (left) and B (right) show classifier train-test performance averaged on CV folds. y-axis: classifiers' training epoch time window, x-axis: test epoch time windows, color map: performance accuracy threshold at >55% accuracy.

3.4.2 Individual timelines in cognitive control

Results of presented under Section 3.4.1 suggest the differences between high and low conflict states throughout the trial. There are two possible reasons for this: (a) all speakers show this extended timeline, or (b) different speakers show peak accuracy at different times, resulting at an extended window at the group level. To distinguish between these two possibilities, we examine if the decoded temporal features underlying control are variably localized at the participant level.

Methods: We implemented a sliding window analysis utilizing a ridge classifier trained on 50ms segments of EEG trials. Each full 1000ms trial is segmented into overlapping time windows of 50ms duration and 10ms step size. An individual classifier is trained on epochs within a 50ms time window and tested on all epochs in all time windows. The peak test accuracy is identified across classifiers trained and tested along an expanded diagonal per participant. Temporal localization is defined as the test time window when the peak test accuracy occurs. For each model, cluster-wise significance was measured after non-parametric multiple-comparisons correction with a p<0.05. Participant-wise distributions of cluster sizes under the null hypothesis were acquired through permutations, with an initial classification accuracy cutoff of 55%.

Results: We found that the ridge classifier successfully decodes high and low-conflict conditions, which are expected to recruit cognitive control, when given only 50 ms of EEG recordings. The observed time of peak accuracy per participant is dispersed across the entire 1000ms trial window. While some participants show early peak test accuracy, most peak after 400 milliseconds (Figure 3.3). Two examples of participant-level results are shown in Figure 3.4.

Implications: The dynamics in classification accuracy across time indicate that even lexical competition resolution, which has been traditionally expected in early (200-250 ms) time windows, can emerge much later in some speakers, highlighting the critical importance of individual differences in models of language production. The observed variability in peak time windows, reinforced by the findings from Section 3.4.1, motivates further participant-level.

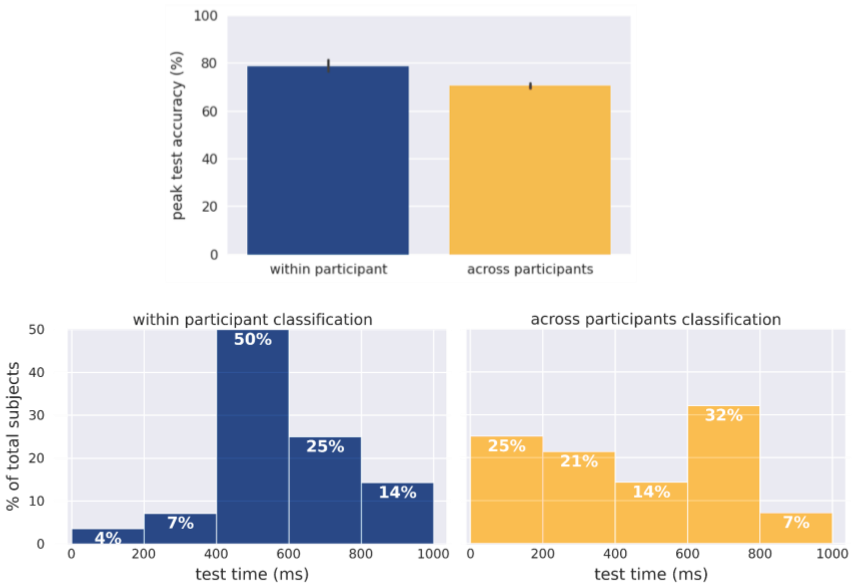


Figure 3.5 Comparison of peak performance accuracy of within- vs. across-participants trained classifiers. The maximal test accuracy is identified across classifiers trained and tested along an expanded diagonal per participant.

3.5 Generalizability of cognitive control across participants

Our results have so far shown that we can successfully decode the neurocognitive states associated with high and low levels of conflict within individual participants. We now ask if these states have a common neural signature in the population. If true, we should be able to successfully decode high and low conflict states using models trained on all but one participant and tested on a held-out participant. Moreover, we expect that the more similar the participants' neural signatures are the greater the across-participant classifiers perform.

Methods: Each within-participant classifier is trained and tested on the same participant. Each across-participants classifier is trained on all but a single left-out test participant. To obtain the peak test accuracy value, trials are segmented into 146 overlapping epochs of 50ms in duration and 10ms step size. Individual classifiers were trained on epochs within a single 50ms time window and tested on all possible epochs from the entire trial length.

Results: In the comparison of the participant level peak performance time window in within-participant versus across-participant classification, we see that classifiers can decode conflict states above chance accuracy in both cases but greater in the within-participant classifiers (Figure 3.5).

Implications: Successful decoding suggests that participants' neural signatures of conflict can be shifted in time. Moreover, there are characteristics shared across participants that vary in temporal presentation from one participant to another.

3.6 Conclusion

Taken together, our findings implicate neural signatures of conflict, and by proxy, control processes occur beyond that time window of lexical selection and are spread across the entire presentation to production timeline Moreover, at the participant level, the temporal localization of control appears highly variable, indicating that individuals may not recruit control at the same time scale. However, despite the variability, we found there are shared characteristics across participants by testing the generalizability of classifiers. In summary, we ask distinct questions in this chapter, directed by theoretical models of language production, and show that linear classifiers, decoding neural signatures, can provide substantive answers about the underlying neural processes.

4 Cognitive control across stages of processing

4.1 Overview

The last chapter investigated conflict and recruitment of control when a prepotent word had to be overwritten by a less potent word. Conflict and the need for control can also arise in other ways within the production system. One important route is through *contextual similarity*—words are generally harder to produce in similar compared to dissimilar contexts (Belke et al., 2005; Nozari et al., 2016; Schnur et al., 2006, 2009) There are two main types of similarity: similarity in meaning, like cat-dog (semantic similarity) or in form, like cat-mat (phonological similarity). These two types of similarity are relevant to different stages of processing. Semantic similarity affects the mapping of semantic features to words. Phonological similarity affects the mapping of words to phonemes. Despite their different loci in the production system, both types of similarity induce behavioral interference through increased conflict with the related competitor (see Nozari & Pinet, 2020, for a review).

This chapter asks four questions: (1) Can we successfully decode low and high-conflict states induced by contextual similarity using the same general approach used in the previous chapter to decode these states in a Stroop-like task? (2) Do language models differ in whether they maintain any modularity to stages of processing or not. Serial models (Levelt et al., 1999), as well as globally modular models (Dell, 1986; Rapp & Goldrick, 2000), posit that semantic-lexical mapping generally precedes lexical-phonological mapping. On the other hand, non-modular models (Strijkers, 2016) reject such temporal segregation. We apply similar TGA techniques as the previous chapter to map out the timeline of semantic and phonological similarity to answer this question. (3) In keeping with the previous section, we ask whether conflict induced by contextual similarity has a common neural signature across speakers. We answer this question by measuring the generalizability of leave-one-out cross-participant classifiers across time and participants. (4) Lastly, we pose a new question in this chapter – Do the different ways in which conflict arises share common neural signatures? We examine this in two parts: (a) whether conflict states induced by different types of contextual similarity have a similar neural signatures and (b) whether the neural signatures associated with contextual similarity share features with the conflict state induced by a Stroop-like manipulation. To answer this question, we carry out a series of cross-classification analyses to measure the generalizability between the high and low-conflict states described in Section 3.2 with conflict induced by semantic and phonological-similarity conditions.

Collectively, the results of the analyses in this chapter will inform us of the decodability of low and high conflict states induced by semantic and phonological similarity within participants and the timeline of such effects, as well as the generalizability of such states across participants. Furthermore, they will shed light on whether conflict states within the language production system have the same neural signature when they arise at different stages of processing or through different manipulations.

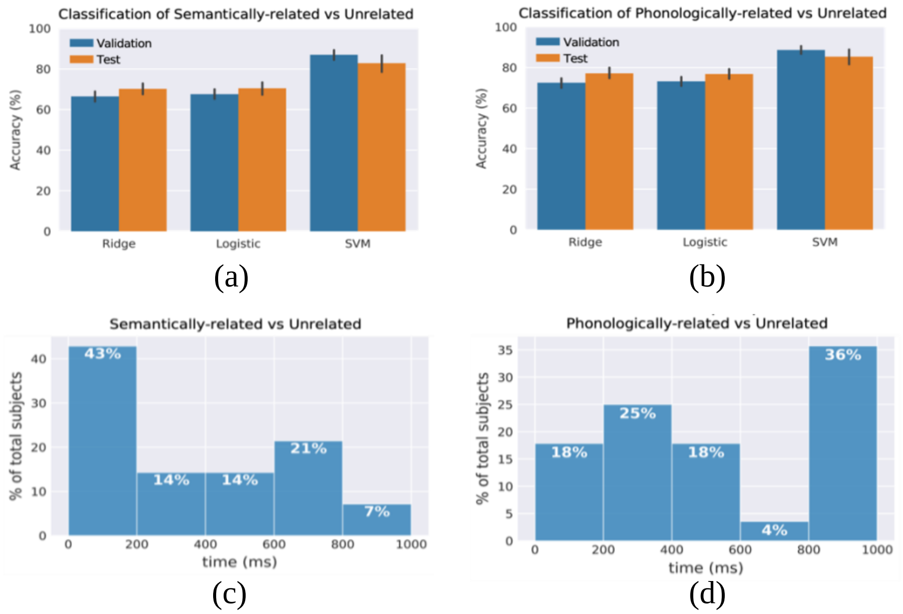
4.2 Experimental paradigm

We examine how changes in the relationship of co-activated item representations affect EEG recordings in a related-unrelated picture naming task. Specifically, participants are presented with two pictures, presented multiple times (see Section 2.1). which they are asked to name. Blocks contain pairs of items that are phonologically, semantically related, or unrelated. Notably, using paired sets on only a small number of pictures allows us to disentangle the effect of a limited set of psycholinguistic processes associated with monitoring and conflict resolution without requiring multiple cognitive processes.

4.3 Completed work: Temporal dynamics of contextual similarity

We first ask if it is possible to distinguish contextually similar vs. unrelated conditions from neural activity recorded with EEG. We did so by applying the approach from Chapter 3 - training three classifiers and evaluating their ability to decode the presented conditions. More specifically, we examined how accurately a ridge classifier, a logistic classifier, and an SVM classified each condition in the two sets: semantically related vs. unrelated, as well as phonologically related vs. unrelated.

In Aim 1, we averaged all repetitions of unique picture-target pairs and stratified train-test-validation block-wise, then completed a nested cross-validation (CV) scheme. Because congruent and incongruent of picture-names were paired within experimental blocks, block-wise splitting ensures that the picture presented remains independent of the intended target production (i.e. each picture presented with a CV fold appears in both congruent or incongruent conditions). Implementing this scheme while stratifying contextually-similarity conditions (i.e., distributing related and unrelated context conditions equally across CV sets) introduces an interaction between unique pictures-names and relatedness within the training sets that are inverted in the test set. The cross-validation scheme was changed to maintain picture-context independence, allowing for conflict due to contextual-relatedness to be learned. Specifically, three exemplar sets were created, each composed of an average of 2-3 repetitions of all unique picture-target stimuli. The cross-validation scheme was then trained, tested, and validated using an exemplar set, each



(a) (b)

(c) (d)

Figure 4.1 Comparison of prediction accuracy from three classifiers, Ridge classifier, SVM, and Decision tree trained and tested on full experimental trial (1000ms). (a) semantically related and (b) phonologically related interference conditions. The training accuracy determined by the inner cross-validation is shown in blue. Second row figures show the distribution of time at which the peak test accuracy occurs for (c) semantic and (d) phonological interference conditions. (SVM=support vector machine)

containing an average of all possible unique trial repetitions, allowing all unique picture stimuli to be independent of contextual similarity. The nested CV approach allowed for hyperparameter tuning and evaluation of the model to remain independent. The resulting classification accuracies were measured as an average of the three possible combinations of train-test-validation splits.

When decoding semantically related vs unrelated conditions, ridge, logistic regression, and SVM classifiers all performed significantly above chance (70.2%, SD=7.8, 70.4%, SD=8.2, and 81.7%, SD=10.8 respectively). Moreover, the SVM classifier decoded semantically related, high-conflict conditions significantly better than both ridge and logistic regression classifiers (**Error! Reference source not found.**a). Similarly, in decoding phonologically related and unrelated conditions, the ridge, logistic regression, and SVM classifiers both performed significantly above chance (77.2%, SD=6.9, 76.8%, SD=6.9, and 85.4%, SD=10.4 respectively). Moreover, the SVM classifier decoded phonologically related, high-conflict conditions significantly better than both ridge and logistic regression classifiers (**Error! Reference source not found.**b).

From these results, the high-conflict states, induced by the co-activation of linguistically related stimuli are distinguishable from the unrelated condition using any of the three classifiers. Notably, contextual interference is much more decodable than interference due to the phase reversal shown in Chapter 3. To ensure this difference is not due to changes in the cross-validation scheme, we repeated analysis from Section 3.3 using the three-fold, 2-3 repetition averaging scheme (Section 4.2) and did not find significant improvements in phase reversal classification.

4.3.1 Temporal variability of participant response:

Subsequently, the optimal classifier type was then used in a sliding window analysis was completed to evaluate the temporal location of at optimally performing time windows. This analysis required that each full 1000ms trial be segmented into epochs with overlapping time windows of 50ms duration and 10ms step size, respectively. For each participant, 146 independent classifiers are trained, one on each possible epoch (segmented trial data from a single 50ms time window) and tested on all possible unobserved epochs.

Significance was calculated by non-parametric measurement of the null distribution and cluster correction. The null distribution was determined on the performance accuracy of classifiers trained and tested on 500 permutations, threshold at a minimum of 55% accuracy and multiple comparisons based on cluster size and a max $p<0.05$.

Results from the sliding window classification analysis indicate that participants implement control on varying timelines. Moreover, the peak decoding of semantically and phonologically related interference have distinct temporal distributions. The majority of participants showed early classification of semantically related interference, with control most decodable prior to 600ms after picture presentation (**Error! Reference source not found.**c) In the phonologically related condition, most participants' time shows a relatively later peak accuracy, distributed up to 1000ms after picture presentation (**Error! Reference source not found.**d).

Based on these results, the classification of control recruited under semantic and phonological similarity conditions are both independently viable. Together with the findings in Chapter 3, our preliminary work supports examining cross-condition classification to inform our understanding of generalizable processes in cognitive control within high-similarity conditions.

4.4 Proposed Analysis: Generalization across participants, stages of production, and domains

The proposed analysis examines the degree of generalization of language production in three parts: (1) if there is temporal variability in conflict responses across participants during contextual interference conditions; (2) whether the observed neural signatures of conflict in semantically similar tasks are generalizable to those of phonologically similar conditions (3) if the decoded features from contextual similarity conditions are generalizable to conflict states decoded during Stroop-like tasks through a series of cross-classification procedures.

4.4.1 Cross-classification of participants' response

To understand how common the neural signatures between participants are, we propose a cross-participant classification for each of the two contextual similarity conditions, independently. While results from Section 4.3 indicate a group-level effect as well as unique distributions of neural signatures for each contextually similar condition, they also show that there exists variability in the timeline of decoded conflict in at the participant level. We plan to complete a similar cross-participant classification approach as described in Section 3.5 to determine how similar the group-level mappings of high versus low-conflict are. Specifically, we will complete a leave-one-out classification scheme where models are trained on all participants except one and tested on the held-out participants data. We repeat this process, with three-fold cross-validation, to measure how well group-level classifiers' map to each participant individually. If full-trial classification shows greater than chance performance, we will complete sliding window classification to determine how common neural signatures can vary in time. Given the significantly greater than chance performance of classifiers when decoding conflict conditions in both semantic and phonologically related tasks, we expect there to be above chance classification in cross-participant classification for full-trial as well as in sliding window analyses. However, based on the variability in neural signature distribution (Figure 4.1a and c), we expect the timeline of significant classification to be shifted when compared to within-participant temporal generalization results.

4.4.2 Cross-classification between similarity conditions.

The experimental paradigm uniquely allows us to examine the relationship between different conditions which can interfere with word production. In the completed work, neural signatures decoded from semantic and phonological induced conflict at different timelines. These results may indicate that semantic and phonological similarity interfere with word production using the similar mechanisms just at different times or rather different mechanisms entirely. To delineate the two possibilities, we ask "can we decode across all conflict conditions and if so, to what degree do the learned signatures overlap?" We implement a cross-task classification approach using the two optimal classifiers in Section 4.3 for each of the two contextual-similarity conditions and test their performance on the alternate contextual similarity task (i.e., Train a model on semantic similarity. Test how well the model can classify phonological similarity and vis versa). We plan to complete this analysis both at the whole-trial level and at overlapping time windows across the trials.

4.4.3 Cross-classification between context similarity and Stroop-like conditions

Taking the analysis of generalization within language production one step further, we examine how similar are the processes underlying conflict observed during suppression of prepotent words and that of co-activation from contextual similarity. We leverage the 2x4 structure of the experimental design to test how the neural signatures decoded during congruent and incongruent trials in Aim 1 relates to that observed during the contextually similar vs unrelated trials.

This analysis involves first creating three subsets of data based on the task condition: subset 1 - phonological and unrelated conditions (congruent trials only), subset 2 - semantic and unrelated conditions (congruent trials only), and subset 3 - congruent and incongruent trials (unrelated only). We use pairs of subsets to train and test the similarity of neural signatures across tasks. To test the generalizability of conflict in the context conditions to the Stroop-like task we train on subsets 1 and 2 independently and test the models on subset 3. To measure the generalizability of conflict during the Stroop-like task on to the contextual similarity tasks, we train on subset 3 and test on subsets 1 and 2. Each cross-classification analysis will be completed on the full trial epochs. If results indicate greater than chance accuracy, we will complete a sliding window classification protocol as described in 3.2 and 4.2 to characterize the temporal generalization between conditions.

5 Cognitive control across domains

5.1 Overview

In Chapters 3 and 4, we showed that we can successfully apply ML tools to decode low and high conflict states from EEG data, across a variety of conditions, including conflict induced by Stoop-like manipulations and contextual similarity at different stages of production. In this chapter, we take our investigation one step further, both theoretically and methodologically. On the theoretical side, we expand on our question from Chapter 4, and ask whether high and low-conflict states, and the consequent control mechanisms, share a neural signature between linguistic and non-linguistic tasks. Methodologically, we expand our method to fMRI, which allows us to directly compare the neural correlates of control across domains.

The question of whether cognitive control is domain-general or domain-specific, remains controversial. Even using the same methodology, such as the congruency sequence effect (CSE; (Gratton et al., 1992), some have not found the transfer of control between domains (e.g., Freund & Nozari, 2018; Verguts & Notebaert, 2008, 2009), and some have (e.g., Hsu & Novick, 2016). We take up this issue in this Chapter by using a paradigm based on Freund and Nozari (2018), which combines two established tasks to elicit control in two different domains. Specifically, we collect fMRI data from participants as they complete a picture-word interference (PWI) task, which elicits control in the language domain, interleaved with a Simon task, requiring control in the visual-motor domain. By leveraging fMRI's high spatial resolution, we ask where control occurs in the brain. Together, by using machine learning classifiers, we ask if the neural correlates of control are shared between a language and non-language task. To do so, we ask three questions: (1) Can we successfully classify low vs. high conflict states in PWI using the fMRI data? (2) Can we do the same for the Simon task? (3) Critically, can we cross-classify low vs. high conflict states between the two tasks? Above-chance within-task classification, together with at-chance cross-task classification implies domain-specific control mechanisms. On the other hand, above-chance classification both within- and between tasks would be evidence for domain-general control mechanisms. The extent to which the accuracy of within- vs. between-classification differs, can be used to assess the balance between domain-specificity and domain-generality of control.

5.2 Experimental data

The paradigm is structured such that 1-back and 2-back results allow for across-domain and within-task examination of cognitive control within one run. During each PWI trial, participants were presented with a line drawing (picture) to be named (target) along with either a distractor word (incongruent trials) or target word (congruent trials) superimposed at the center of the picture. Distractor words were semantically related to the target as well as matched in word length (see Section 2.1.2).

We are interested in measuring the domain-generality or specificity between language and non-linguistic tasks. In doing so, we examine the generalizability of neural correlates of conflict induced by congruent and incongruent trials within language production and visual-motor domains. In order to do so, we utilize a 2-by-2 experimental paradigm consisting of two tasks (PWI and Simon's) and two control conditions (high and low-conflict). The data acquisition for this study has been completed. It includes fMRI recordings from 24 subjects. The fMRI and structural MRI have been preprocessed, and univariate analysis of cross-control conditions has been completed by a fellow lab member. The proposed analysis focuses on within- and across-domain classification at both a whole brain and localized level.

5.3 Proposed work

We aim to investigate the domain-generality or specificity of cognitive control between cognitive control in language production and visual-spatial tasks. We propose using multivariate pattern analysis (MVPA) to understand the domain-generality—this approach allows for localization of regional activation and identification of distributed patterns of activity *within* and *across* tasks. We will use as input beta values estimated for each voxel at each trial. We will perform both within- and across-domain classification, and for each, we will use perform the classification at both the whole brain and searchlight level (Kriegeskorte et al., 2006).

Language Production models. In these analyses, we will take as input the beta values for single trials during the PWI task. The output will consist of two labels: congruent and incongruent. In the within-domain condition, we train and test on the PWI task data in cross-validation. In the across-domain condition, we train on all the PWI task data and test on all the Simon's task data. We perform these analyses at the whole brain and searchlight levels.

Visuo-spatial models. In these analyses, we will take as input the beta values for single trials during the Simon's task. The output will consist of two labels: congruent and incongruent. In the within-domain condition, we train and test on the Simon's task data in cross-validation. If necessary, we will average the trials to improve the signal to noise ratio. In the across-domain condition, we train on all the Simon's task data and test on all the PWI's task data. We perform these analyses at the whole brain and searchlight levels.

Potential outcomes. There are three potential outcomes:

- *Cognitive control is domain-specific:* This outcome would be supported by our results if the classifiers trained on one cognitive-control task would not be able to decode cognitive-control in the other task. This would imply that the neural correlates underlying cognitive control are generally different across domains.
- *Cognitive control is universal:* This outcome would be supported by our results if the classifiers trained on the PWI task transfers well to the Simon's task, and vice versa. This would imply that the neural correlates underlying cognitive control are generally the same across domains.
- *Cognitive control is primarily domain-specific:* This outcome would be supported by our results if the classifiers trained on the PWI task transfers well to the Simon's task, and vice versa, but only in some regions (based on searchlight analysis). This would imply that the neural correlates underlying cognitive-control are different across domains but that certain brain areas are recruited by shared cognitive-control demands.

Potential pitfalls. The PWI task has only one trial of each picture-name pair. Thus, the signal-to-noise will be lower than in the Simon's task which has the possibility of averaging repetitions of the same trial. The reduced signal-to-noise might prevent classification performance from reaching statistical significance. Further, the across-domain classification results might only be successful in one direction and not the other, which would complicate the interpretability of results. It is also not clear if the searchlight analysis is the most appropriate one, as perhaps an ROI level analysis can be more informative.

6 References

Belke, E., Meyer, A. S., & Damian, M. F. (2005). Refractory effects in picture naming as assessed in a semantic blocking paradigm. *The Quarterly Journal of Experimental Psychology Section A*, *58*(4), 667–692. https://doi.org/10.1080/02724980443000142

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Freund, M., & Nozari, N. (2018). Is adaptive control in language production mediated by learning? *Cognition*, *176*, 107–130. https://doi.org/10.1016/j.cognition.2018.03.009

Fyshe, A., Sudre, G., Wehbe, L., Rafidi, N., & Mitchell, T. M. (2019). The lexical semantics of adjective–noun phrases in the human brain. *Human Brain Mapping*, *40*(15), 4457–4469. https://doi.org/10.1002/hbm.24714

Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General*, *121*(4), 480–506. https://doi.org/10.1037/0096-3445.121.4.480

Hirschfeld, G., Jansma, B., Bölte, J., & Zwitserlood, P. (2008). Interference and facilitation in overt speech production investigated with event-related potentials. *NeuroReport*, *19*(12), 1227–1230. https://doi.org/10.1097/WNR.0b013e328309ecd1

Hsu, N. S., & Novick, J. M. (2016). Dynamic Engagement of Cognitive Control Modulates Recovery From Misinterpretation During Real-Time Language Processing. *Psychological Science*, *27*(4), 572–582. https://doi.org/10.1177/0956797615625223

Indefrey, P. (2011). The Spatial and Temporal Signatures of Word Production Components: A Critical Update. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. https://doi.org/10.1016/j.tics.2014.01.002

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38. https://doi.org/10.1017/S0140525X99001776

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Nozari, N., Freund, M., Breining, B., Rapp, B., & Gordon, B. (2016). Cognitive control during selection and repair in word production. *Language, Cognition and Neuroscience*, *31*(7), 886–903. https://doi.org/10.1080/23273798.2016.1157194

Nozari, N., & Pinet, S. (2020). A critical review of the behavioral, neuroimaging, and electrophysiological studies of co-activation of representations during word production. *Journal of Neurolinguistics*, *53*, 100875. https://doi.org/10.1016/j.jneuroling.2019.100875

Pinet, S., & Nozari, N. (2023). *Different electrophysiological signatures of similarity-induced and Stroop-like interference in language production* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/wq7pf

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review*, *107*(3), 460–499. https://doi.org/10.1037/0033-295X.107.3.460

Ries, S. K., Pinet, S., Nozari, N. B., & Knight, R. T. (2021). Characterizing multi-word speech production using event-related potentials. *Psychophysiology*, *58*(5). Scopus. https://doi.org/10.1111/psyp.13788

Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language*, *54*(2), 199–227. https://doi.org/10.1016/j.jml.2005.10.002

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2009).

Localizing interference during naming: Convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proceedings of the National Academy of Sciences*, *106*(1), 322–327. https://doi.org/10.1073/pnas.0805874106

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102. https://doi.org/10.1016/0749-596X(90)90011-N

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300–304. https://doi.org/10.1037/h0020586

Strijkers, K. (2016). Can hierarchical models display parallel cortical dynamics? A non-hierarchical alternative of brain language theory. *Language, Cognition and Neuroscience*, *31*(4), 465–469. https://doi.org/10.1080/23273798.2015.1096403

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control: Dealing with specific and nonspecific adaptation. *Psychological Review*, *115*(2), 518–525. https://doi.org/10.1037/0033-295X.115.2.518

Verguts, T., & Notebaert, W. (2009). Adaptation by binding: A learning account of cognitive control. *Trends in Cognitive Sciences*, *13*(6), 252–257. https://doi.org/10.1016/j.tics.2009.02.007