Situational demands and emotional significance during language processing

A dissertation by
Nathaniel Delaney-Busch

In Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy in Psychology

TUFTS UNIVERSITY

February 2016

Advisor: Gina Kuperberg, M.D., Ph.D.

Abstract

The meaning of a word is more than definition. The comprehension of the words "love" or "death" entails a deep understanding of the social and emotional implications, as wells as contextual and motivational significance. In this dissertation, we explore the neuroscience of how emotional content, local context, and task demands influence word processing. In chapter I, we describe analytic approaches to these questions that model word processing as the convergence of the particulars of the item, the comprehender, and context. In chapter II, we apply these techniques to investigate how the semantic processing of infrequent words is facilitated if that word is also highly emotional, regardless of whether attention is oriented towards semantic features or emotional features. In chapter III, we present two experiments that investigate the role of valence in semantic processing by implementing a full cross of semantic priming and affective priming, finding radically different patterns of effects that suggest the two may rely on distinct mechanisms. And in chapter IV, we investigate how the semantic priming effect adapts to the local context as participants implicitly learn the statistical contingencies, using a novel trial-by-trial adaptation analysis that shows the evolution of the semantic priming effect through time. Overall, these data suggest that semantic processing is fundamentally supported by all pertinent knowledge, including knowledge of emotional significance and implicit contextual expectations. Language comprehension is as richly textured as the comprehenders.

Dedicated to my grandparents

Acknowledgments

My work for this dissertation over the past six years was guided, inspired, and supported by more outstanding individuals than I could possibly name. I'd like to thank my mentors here at Tufts University and back at UC Davis, particularly my insightful committee and my wonderful advisor Dr. Gina Kuperberg, for always helping me grow in the right direction. I'd like to thank my lab, for always challenging me to be better and for providing an environment where it's possible to thrive. In particular I am tremendously grateful to my amazing undergraduate volunteers, who made this work possible. I'd like to thank my family, for instilling in me the sense of purpose and community that allowed me to make it this far (especially my parents, to whom I owe everything), and my friends, for always keeping me sane and grounded. And last but certainly not least, I'd like to thank my wife Siobhan McRee. I could not ask for a more insightful, encouraging, compassionate, and meaningful partner. Thank you for everything!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Emotion is central to deriving coherent meaning from a wide universe of possible stimuli and events (Osgood, Suci, & Tannenbaum, 1967). Even for language, the appropriate comprehension of even simple utterances like "I love you" and "he passed away" requires consideration of the profound social and emotional implications just as much as the knowledge about the properties of love and death. Emotionally significant speech can resonate with our memories and our hopes (Bradley, 1994), our fears and our joys (Bradley & Lang, 2007), and even our most deeply-held values (Van Berkum, Holleman, Nieuwland, Otten, & Murre, 2009).

As a consequence, both the process of comprehension and the resulting message that is actually comprehended are likely richly nuanced. There is a huge array possibilities for how the particulars of the word content, individual speakers and comprehenders, and contexts can interact during language processing, before even considering social implications and pragmatic inferences. In essence, word processing is a fundamentally multilevel process. Event-related potentials (ERPs), direct online measures of neural activity recorded at the surface of the scalp, have been essential in beginning to uncover how these levels influence word processing as it unfolds through time. But much remains largely unknown. In the present dissertation, we systematically explore interactions between and across these levels to better understand how emotion and context contribute to language processing.

But first, we detail how such multi-level research questions are more suitably addressed by the use of multi-level analytic techniques. Mixed-effects regression, for example, unites the analytic approach to inference with the theoretical motivations behind the questions. We discuss how these trial-level regression models are similar to and different from traditional subjects ANOVAs and item regressions, outline the procedure for fitting such models, and work through an example using real experimental data (from Delaney-Busch, Wilkie, & Kuperberg, *under review*) and example R code.

Then, we continue chapter 1 by discussing a number of specific issues that must be considered when using this technique, providing recommendations and references to further discussion where appropriate. This section includes detailed consideration of how data should be collected and treated (including normality and outliers), how theoretically-motivated models can be constructed and utilized (including model selection, the computation of p-values, and the specification of random effects), and how to appropriately handle spatial factors, temporal factors, and ordering effects. In each discussion, the theoretical issues summarized are linked to the applied usage in chapters 2 and 4, where the questions under discussion are the most suitably addressed by regression-based techniques. And finally, while much of this chapter reviews and consolidates existing literatures, we also contribute a novel analysis of the practical effects of different assumptions about the treatment of electrodes in a regression, using tens of thousands of data permutations that are each fit by the different models.

In chapter 2, we apply these modeling techniques to better understand the processing of single emotional words. It has long been suggested that emotional

information can at times be "privileged" over other information during stimulus processing — the so-called *affective primacy hypothesis*. It may be advantageous in some circumstances to maintain a vigilance for emotionally significant stimulus features, particularly when a highly motivationally significant stimulus is generally rare in occurrence (and thus would tend to be surprising when encountered). For example, if "avalanche!" is shouted in your direction, the emotional features (and thus motivational significance) may be valuable to extract as quickly as possible. Some recent behavioral (Kuchinke, Vo, Hofmann, & Jacobs, 2007) and neuroimaging (Mendez-Bertolo, Pozo, & Hinojosa, 2011; Scott, O'Donnell, Leuthold, & Sereno, 2009) studies have found evidence of such a facilitation for emotional words. Specifically, *infrequent* emotional words appear to be treated by the brain as more frequent than they actually are. We call this the "avalanche effect". The present investigation was designed to better characterize the avalanche effect, particularly by specifying which particular dimension of emotion appears to be most responsible for the facilitation. We attempted to isolate the effect of valence from the effect of arousal on the N400 lexical frequency effect. And further, we hypothesized that the avalanche effect would be larger in experimental contexts that encourage the deployment of attention towards emotional features. Regression-based approaches to ERP analysis, where single-trial amplitudes are modeled using predictors based on information about the items, subjects, and experimental context, were used address these questions.

In chapter III, we expand our investigation of emotional words to the simplest of contexts: a single preceding word, called the "prime". Priming paradigms have offered

critical insights into how word meanings are stored and utilized during language comprehension (Herring et al., 2013; Neely, 1991). Among a number of more nuanced questions, they indicate that the processing of a word can be facilitated by either a semantically related *or* an emotionally related prime. However, the extent that affective priming and semantic priming rely on shared or distinct mechanisms is presently unclear (Klauer & Musch, 2001, 2003). Specifically, the mechanisms behind affective priming are less well understood than those behind semantic priming.

The present experiments address this gap. We implemented a full crossing of semantic priming and affective priming under two different tasks, utilizing a large, carefully controlled set of stimuli. We considered a number of possibilities that could help explain the divergent ERP literature on affective priming. First, emotion could be nonrandomly rated to particular semantic features, such that valence correlates with semantic content. Second, previous ERP studies could have found N400 affective priming effects because emotional features (such as valence) are actually stored as part of a word's meaning and activated during semantic processing. And third, affective priming could instead be a consequence of facilitation during *evaluative* processing (as required to make typically task-motivated decisions) rather than semantic processing. In addition to providing information about how semantic priming and affective priming relate to one another during word processing, this design also allows for the assessment of how semantic priming is influenced by emotion more generally. And lastly, we considered how these priming effects might be influenced by task.

Finally, in chapter IV, we investigate how word processing adapts over time to the particulars of the context. Prediction appears to be a central feature of language comprehension (Kuperberg & Jaeger, In Press). When *semantic* information is activated prior to bottom-up input (i.e. predicted), the semantic processing for the incoming word is typically facilitated. The strength of this semantic priming effect depends on the extent to which the prime can pre-activate the semantic features of the target. As such, the semantic priming effect is likely sensitive to predictive validity as participants adapt to the likelihood of seeing highly related targets over time. However, adaptation of the N400 semantic priming effect to changes in predictive validity has not to our knowledge been directly observed. We directly assessed the trial-by-trial adaptation in two ways, utilizing both local regression and mixed-effect models.

Overall, these studies suggest that semantic processing is broadly supported by the rich experiences and patterns available to the comprehender, including knowledge of emotional significance and implicit contextual expectations. Comprehension through definition is as limited as paint by numbers. Language comprehension is as richly textured as the comprehenders.

CHAPTER I

AN INTRODUCTION TO MIXED-EFFECT MODELS FOR ERP ANALYSIS

Analysis of Event-related Potentials (ERPs)

Typically, event-related potentials (ERPs) are derived by averaging across time-locked trials. Brain activity that is orthogonal to the stimulus presentation can be assumed to average out to zero as the sample size increases, while brain activity that is related to the stimulus presentation is captured by the relative amplitudes of the resulting averaged waveforms. Typically, EEGs from trials of the same "condition" (or level of a categorical factor) are averaged together within subjects (a so-called "subjects analysis"), though it is also relatively common to average together EEGs from different participants for each individual item (a so-called "items analysis"). The resulting waveforms from the subjects and/or items analysis are then used for data visualization and for inferential hypothesis tests.

Specifically, waveforms from the subjects analysis can be further averaged across subjects to get "grand mean" waveforms, a visualization that can be interpreted as the average evoked potential for a particular category of stimulus across items and subjects. Comparisons of the grand mean waveforms between two categories of stimuli can reveal amplitude differences in particular time windows and/or at particular locations on the scalp. ERP "components" are systematic differences that can be reliably evoked by particular types of contrasts that are thought to differentially activate the same or similar neural generators (Luck, 2014). For example, the N400 ERP component typically appears

as an amplitude difference around 300-500ms over centro-parietal electrodes in response to words or pictures that differ in their semantic expectedness (Kutas & Federmeier, 2011; Kutas, Van Petten, & Kluender, 2006), where unexpected stimuli elicit a larger negative amplitude than expected stimuli. Though components are often described in terms of their positive- or negative-going direction, local maximums for individual waveforms are actually underinformative and depend considerably on the choice of reference (Luck, 2014) and the overlap between components: the *contrast* between two categories is the primary signal of interest.

In a subjects analysis, differences between waveforms are commonly assessed by averaging over time points within a time window of interest to get "mean amplitudes" for each electrode for each stimulus category for each subject, and utilizing these mean amplitude values for a within-subjects (randomized block) ANOVA (Luck, 2014). Though it is also possible to evaluate differences in other properties of the waveform, including peak amplitude, peak latency, point-to-point amplitudes and more, the mean amplitude effects are generally the most stable across studies (Luck, 2014). During the ANOVA procedure, it is common to average mean amplitudes from sets of individual electrodes into small "regions" of the scalp, as licensed by the component of interest. Overall, this approach has been a powerful tool in visualizing ERP effects and conducting inferential statistics. It is well-suited for most studies in which categorical independent measures are hypothesized to modulate an a priori ERP component, such as in chapter III where categorical semantic priming and affective priming effects are assessed on the N400 component.

However, in recent years, a number of research questions have arisen that are less elegantly addressed with this classic subjects- or items-analysis, such as questions pertaining to stimulus features that are not categories at all, but are rather measured on a scale, such as word frequency or length (number of letters). Ranges on these scales can be "binned" into categories (such as "low frequency words"), but at the cost of potentially arbitrary cutoffs and often reduced statistical power. In this example, the frequency variation *within* each bin would be treated as noise, despite being the primary feature of interest. Further, this sort of binning cannot easily address questions about the shape of the relationship between the ERP amplitudes (such as the N400) and the independent measures (such as frequency), because any inferences would need to extend over the "gaps" between the means of each bin. An items analysis is more appropriate for such questions, where the ERP amplitude is regressed against scale measures of the independent predictors, such as frequency (Laszlo & Federmeier, 2010). In fact, items analyses can address most of the same questions that subjects analyses can, in addition to having the added benefit of being able to use scale variables and test specific types of relationships.

But there are a number of recent research questions that the typical items analysis cannot well address either. For instance, questions about how effects differ across individuals. Essentially, the limitations that subjects analyses have for addressing questions about the properties of items (such as the effect of word frequency) are the same as the limitations that items analyses have for addressing questions about the properties of subjects (such as the effect of verbal fluency or working memory). Though

questions pertaining to *either* features of the subjects or features of the items can be addressed by simply choosing the appropriate approach, this leaves no elegant way to characterize interactions between these levels, such as whether the N400 frequency effect is influenced by verbal fluency. Similarly, it is challenging in an items analysis to incorporate other cross-level interactions, such as learning or adaptation and other context effects. Finally, neither items- nor subjects- analyses alone can elegantly handle unbalanced designs (e.g. where particular trials must be rejected due to artifact).

In the present chapter, we discuss a mixed-effects regression approach to ERP analysis (using unaveraged trial-level data) that combines the strengths of both items- and subjects-analyses while addressing many of their weaknesses. It is well-suited to address particular types of questions of interest to psycholinguists and affective neuroscientists (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013; Hauk, Davis, Ford, Pulvermuller, & Marslen-Wilson, 2006), including questions pertaining to effects of continuous variables (e.g. the relationship between valence and the emotional late positivity component, which is thought to be a U-shaped function, where pleasant and unpleasant words elicit a higher positivity than neutral words), individual differences (e.g. the effect of anxiety on emotion processing), and change over time (e.g. learning or adaptation).

We will first present the theoretical motivation behind the use of mixed-effects regression for ERP analysis, both for hypothesis testing (”explanation”) and functional generalizations (”prediction”). Then, we discuss practical considerations for conducting a regression-based analysis of ERP data, including feature selection, outliers, model

construction and selection, the handling of spatial and temporal factors, the handling of sequence effects, the computation of inferential statistics, and the interpretation of such effects. This will include step-by-step examples (with R syntax) using real experimental data.

Regression Techniques

An ANOVA is a special case of a larger family of statistical regression procedures, known as the general linear model. Specifically, an ANOVA using averaged data is functionally identical to a linear regression with categorical independent variables. As such, every common technique used to analyze ERP data with ANOVAs has a corresponding way to express the technique as part of a regression model (Smith & Kutas, 2015a). For example, finding an average amplitude elicited by certain groups (e.g. across subjects or items) is identical to fitting a regression model to the levels of that group (Smith & Kutas, 2015a). Conducting a hypothesis test for an interaction in an ANOVA is identical to conducting a hypothesis test for an effect-coded categorical interaction in a regression. And even including continuous nuisance factors in an ANCOVA is identical to including those same factors as predictors in a regression model. Other than possibly some intuitiveness and familiarity (which are of course important for effective scientific communication), nothing is lost when expressing ANOVAs in regression form. However, much additional flexibility is gained, as we can exploit the remainder of the techniques licensed by the general linear model. The basics of linear regression is described in appendix A.

While ANOVAs share the same statistical roots as regression-based techniques, the linear model has a number of additional capabilities that are useful for particular types of questions. First and foremost, continuous measures can be used for independent variables. For instance, "Valence" and "Arousal", the two main dimensions of emotion (Osgood et al., 1967), are often expressed on 9-point scales (Bradley & Lang, 1999; Warriner, Kuperman, & Brysbaert, 2013). Though formally an ordinal measure, the "distances" between steps on the scale are generally thought to be approximately equal, and thus it is common for valence and arousal to be treated as scale measures within affective space (Bradley, 2000; Bradley & Lang, 1999).

A large number of studies have found that neutral low-arousal items tend to elicit a smaller late positivity in ERP waveforms than pleasant or unpleasant high-arousal items (Citron, 2012; Hajcak, MacNamara, & Olvet, 2010; Herbert, Kissler, Junghöfer, Peyk, & Rockstroh, 2006), but the use of ANOVAs for these analyses requires imposing a potentially arbitrary category boundary over the continuous valence and arousal scales (e.g. between "neutral" and "pleasant" valence) and treats within-category variance as noise. Further, it disconnects the motivating theoretical orientation, in this case the circumplex model of affect (Lang & Bradley, 2009; Russell, 1980), from the operationalization of the measures. Instead, we might use a regression to examine how the late positivity changes as a function of continuous emotion predictors. For instance, we might expect that the late positivity increases linearly as a function of arousal (where higher arousal items elicit a larger positivity than lower arousal items), while relating in a polynomial manner to valence (specifically a second-order function where more pleasant

and unpleasant items elicit a greater positivity than less pleasant or unpleasant items).

Though the inferences appear very similar, the regression-based approach has several potential advantages: it no longer requires the assumption that the between-category variance is representative of the whole variance, we gain additional information about the best-fitting model shapes (e.g. a second-order valence effect as opposed to a linear arousal effect), we usually increase statistical power by incorporating what would have been within-group variance, and the theoretical orientation is unified with the measurement operationalizations.

It's important to note, however, that the interpretation of the regression coefficients is slightly different for continuous data than it is for categorical data. In ERP analyses using categorical data, the fitted coefficient reflects the difference in expected mean amplitude between two categories. In contrast, in ERP analyses using continuous data, the fitted coefficient reflects the change in amplitude that is expected per change in the continuous predictor. For example, an arousal effect on the LPC would be expressed in terms of change in LPC amplitude per unit arousal (e.g. the slope of a regression line), while a valence effect on the LPC might be expressed in terms of change in LPC amplitude per unit of squared valence (e.g. the "steepness" of a U-shaped polynomial, or the slope of the second-order term of a line).

Often in psycholinguistics (as with all sciences), we are interested in identifying the effect of some variable above and beyond what could also be accounted for by some correlated nuisance factors. For example, we might be interested in the effect of lexical frequency on the N400 after controlling for word length or bigram frequency (see "model

selection" below). In traditional ERP analyses, this is usually done by "matching" nuisance factors across levels of the experimental factor, e.g. matching bigram frequency (on average) across both high- and low-frequency groups of words. However, such a procedure can quickly become difficult as the number of nuisance factors grows (though there are handy algorithms for automating the matching procedure, such as van Casteren & Davis, 2006).

More pressingly, however, the tight constraints on stimuli may become strict enough to make the final stimulus set unrepresentative of the desired level of inference. The random sampling of stimuli is important for drawing generalizations across stimuli (e.g. of "English language processing"), but selecting the specific stimuli that meet the imposed criteria may limit what can actually be generalized from a study (as the sample may be determined ahead of time in large part by the constraints, and thus nonrandom).

In contrast, regression-based approaches (including items analyses) do not necessarily require strict matching procedures. Terms in a multiple regression are interpreted as the relationship between the predictor and the outcome (i.e. ERP amplitude) holding all else in the model equal. In other words, fitted coefficients capture the variance explained by that predictor above and beyond all other predictors. Variance that is shared between two or more variables is expressed in the overall r^2 value of the whole model, but is attributed to neither of the individual variables. This is essentially comparable to using an ANCOVA to control for a nuisance factor. As such, "nuisance" factors are simply added to the regression model in the same way as any other predictor, and the interpretation of the effect of interest will reflect having controlled for the other

variables. As the interpretation of any fitted effect in a model fundamentally depends on the other variables included in the model, however, it is critical to report all included variables, and to draw inferences that reflect the model composition (e.g. the N400 frequency effect holding word class and orthographic neighborhood equal). It should also be noted that regression models with correlated inputs can influence coefficient estimates, sometimes considerably (Wurm & Fisicaro, 2014).

Finally, the use of regression techniques allows for the overt specification of the outcome distribution, where an ANOVA approach does not. The same set of principles that apply to regressing against normally distributed outcomes (such as ERP data) can also apply to binomial distributions like "correct" and "incorrect" behavioral responses, or to Poisson distributions like counts (e.g. "number of disfluencies in a speech stream"). This makes use of the "generalized" linear model, where regression models are "linked" to particular outcome distributions using a "link function" (Hastie, Tibshirani, & Friedman, 2009). As such, the same model that tests how the N400 is affected by lexical frequency during a lexical decision task could also be used to test how decision accuracy is affected by lexical frequency, nicely unifying the analytic framework for the normally-distributed ERP amplitudes and the binomially-distributed behavioral responses (Jaeger, 2008).

At present, it is common to simply sum correct responses (then derive a "percentage correct") and conduct an ANOVA over these aggregate accuracy scores that is on its face similar to the typical ERP analysis. Unfortunately, such an approach is often anticonservative (Jaeger, 2008), assigning probability mass in the test distributions to

events that could not possibly occur (such as an accuracy rate higher than 100%). The fundamental problem is that the actual data is binomially distributed and thus bounded, while the ANOVA assumes normally distributed data that is not bounded. Instead, the binomial outcomes can be appropriately modeled using logistic regression, which is easily implemented with the "logit" link-function (Jaeger, 2008). In this way, the regression approach can be tailored to accommodate the most commonly encountered data, with an interpretation that is relatively consistent across forms.

Because of these benefits, regression is now a routine analysis in many fields, including behavioral psycholinguistics. Here, we simply apply the same approach to ERP data, with a focus on psycholinguistics.

Fixed and Random effects in ERP analysis

With the intercept and parameters derived by regression (see appendix A), the fitted model can be used both for the generation of predictions (i.e. expected values of y for novel values of x) and for explanatory purposes, describing the nature of some relationship between the predictors and the outcome (Shmueli, 2010). In psycholinguistics, we are often interested in the latter. Drawing an inference using a regression model is comparable to drawing inferences from other statistical techniques: we manipulate the operationalized factor x (such as word frequency), collect observations of the outcome y (such as N400 amplitudes) and fit a model the describes how y changes in response to the manipulated change in x (e.g. a decrease in N400 amplitude that corresponds with an increase in log frequency).

To generalize this effect across individuals, a representative random sample of subjects is drawn for testing, and to generalize this effect across some portion of the language, a representative random sample of linguistic items is drawn for testing. The experimental effect of interest, such as the N400 frequency effect, is expected to replicate (with some uncertainty) to new samples of subjects and/or items, and if manipulated in the experimental design by the researcher, causality can be inferred. These experimental predictors that a) are determined by the researcher and b) replicate across novel samples can be considered as a "fixed" contributor to the ERP amplitudes.

However, just as N400 amplitudes may differ across levels of frequency, so too might N400 amplitudes differ on average from person to person or from item to item. Critically, these subject and item differences are not expected to directly replicate to new samples of subjects and items. "Subject 10" might have an above-average N400 amplitude (across all items) in one study, but we would not expect "Subject 10" in a different study to necessarily also have an above-average N400 amplitude (assuming that subjects are randomly sampled and thus the probability that subject 10 is the same person is vanishingly small). Rather, the differences in outcomes across particular subjects and items are sample-specific (and thus experiment-specific), because particular subjects and items are the randomly sampled points from the underlying population distribution. So while estimates for "fixed" experimental effects are fundamentally similar across samples (i.e. reflect the population parameter mean), estimates for particular subjects and items are necessarily different across samples (i.e. reflect the random sampling from a population distribution). In short, while experimental factors like frequency and word

class are “fixed” in their relationship to the outcome, the contributions of subjects and items to the outcome are “random”, and particular to the idiosyncratic sample the experiment happened to take. The only stable feature of subject and item effects is the variance: how different, on average, we expect subjects and items to be from one another when we generate a sample.

This distinction between “fixed” effects like frequency and word class and “random” effects like subject and item is central to appropriate statistical analysis, and thus to inference. Consider a thought experiment where we show 1000 words of various frequencies to two subjects. Both the subjects and items (words) have been randomly sampled, and we are interested in determining the fixed (and generalizable) effect of word frequency on the N400. After collecting the data, we construct a simple regression model with N400 amplitude as the outcome and frequency as the predictor, along with an intercept. As described above, we intend to infer a frequency effect across subjects and items (i.e. for speakers of English in similar contexts). However, assume that the participants naturally happened to differ in their average N400 amplitudes across all of the words (regardless of frequency). Perhaps “subject A” has a mean N400 amplitude of -4μv and subject B has a mean N400 amplitude of -1μv. This sort of variation is typical of subject-level ERP amplitudes.

Once we fit the regression model, even if change in N400 amplitude is actually relatively well determined by frequency, the model will report a very high degree of residual error, because the single best-fit line will split the difference between these two subjects (and thus “miss” for both subjects). Further, once we plot the residuals of our

model, we see an immediate problem: the residuals are not normally distributed. Rather than most of the error density being slightly above and slightly below zero with just a few outliers, *most* of the error is either significantly below the best-fit line, or significantly above the best-fit line, with relatively little error that is close to zero. Worse, there is a pattern to the bias: residual error for subject A is predominantly negative (i.e. we are predicting N400 amplitudes that are not negative enough), while residual error for subject B is predominantly positive (i.e. we are predicting N400 amplitudes that are too negative). This clearly violates the independence assumption of regression: the residuals are not independent if they "clump" around subjects. Overall, the model fit is underestimated, and the assumptions of both normality and independence are violated. It should be readily apparent that simple regression is not well suited for repeated-measure designs, because it fails to account for dependencies in the data (where data from any given subject will be more similar to other data from that subject than it will to the remainder of the data from other subjects).

The primary issue in our thought experiment is a failure to consider our units of sampling. Any given subject will randomly differ from other subjects, and any given item will randomly differ from other items. This "clumpiness" about subjects and items violates the assumptions of regression and can lead to erroneous conclusions (Baayen et al., 2008; Barr et al., 2013). One way to incorporate information about subjects and items is to enter them into the regression as well. Because every term in a regression is interpreted as being marginal to all other terms in the model, this should allow us to infer the marginal effect of frequency above and beyond the variance in N400 amplitudes that

can be explained by the particular subjects and items we happened to include in our samples. However, a brief consideration will also show that this may be problematic. First, the resulting effects for subjects and items are relatively uninterpretable: it doesn't make much sense to generalize the difference between subject A and subject B, because nobody else is likely to have an interest in these particular subjects specifically (though this may be merited in cases of extremely limited populations, such as aphasia patients that are to be tested numerous times in the future and for whom individual information is compelling). And second, we quickly encounter a limit to our statistical power: a fixed effect for "item" with one thousand levels (or 999 dummy codes) is almost absurd to estimate with only two data points per item, much less to find an effect of frequency that is marginal to these estimates.

Mixed-effects regression was developed to address these problems. Fixed effects like word frequency and random effects like subject and item are both included as predictors, and each term is overtly specified as either "fixed" or "random". Essentially, the random effects structure can be used to identify the variance accounted for by the units of sampling. Since the model contains both fixed and random factors, it is called a "mixed" model.

Fixed effects are still estimated as described above, with coefficients that reflect the generalizable marginal contribution to the outcome. But random effects, where data tend to "cluster" around subjects or items, are accounted for by alternative means, typically "best linear unbiased prediction" (BLUP). BLUP makes adjustments to the levels of a random factor (e.g. across subjects) to account for the dependencies in the

data. Importantly, these adjustments sum to zero. In our thought experiment, it might adjust our estimations of the amplitudes for subject A to be relatively lower and for subject B to be relatively higher. Because the sum of adjustments is 0, the rest of the model (including the fixed effect) is not biased by the procedure, even if the data for subjects A and B are imbalanced. If we were to add a third subject, our best guess of what their mean N400 amplitude would be prior to data collection is still the overall mean of subjects A and B. Further, these adjustments err on the side of being conservative - relatively less informative data are "shrunk" towards the mean, and thus overall variance in the BLUP adjustments for a random factor are less than the "true" variance along that random factor (though the BLUPs asymptotically approach the true effects as sample size increases)(Jiang, 1998).[1] In the problem above, where "word" has 1000 levels with only 2 instances at each level, the shrinkage factor ensures that the actual predicted amplitudes for those words is a product of both the knowledge of the mean amplitudes for the individual words *and* the knowledge of the typical amplitudes *across* words (where all words are assumed to be randomly sampled from the same distribution). Further, the adjustments are maximally "unbiased" in the sense that the average bias across all of the potentially "true" values for each level of the random factor is zero (Robinson, 1991). Provided these constraints - of linear unbiased adjustments (with shrinkage) that sum to zero - the actual by-item and by-subject adjustments are determined simply by minimizing mean squared error.

[1] This "shrinkage" is a consequence of using a bayesian prior at zero, as the BLUP procedure is related to an empirical Bayes estimate

Since the BLUPs are unbiased and sum to zero, there is no "main effect" of a random factor to express. Some subjects will have an above-average N400, and some will have a below-average N400, but the average by-subjects adjustment is 0. Instead, it is common to report the *variance* of random factors - how different individual subjects (or items) are from one another. This captures how big of an adjustment was required for that random factor, on average. It also reflects the motivation for accounting for units of sampling. Random factors have levels that are sampled randomly from some population distribution. In practice, even if people show the same average frequency effect, they will not yield identical mean N400 amplitudes - some will have amplitudes above average, and some will have amplitudes below average. We can account for these differences in our sample in order to better estimate the fixed effect of frequency on the N400. But if we collect a new sample of subjects from the original population, even though the particular adjustments will differ, the typical size of the adjustment required should remain comparable, because the subjects are sampled from the same distribution (i.e. the variance in the random factor is expected to replicate, not the values for particular levels).

In sum, adding a random factor for "subject" in this manner essentially fits a regression line to the data after it has already been adjusted for each particular participant (or more generally, each unit of sampling). Inferences about the fixed effects pertain to the mean amplitude expected for particular values of the fixed factor, and are drawn across subjects and items. Inferences about *particular* subjects pertain to the mean amplitude expected for particular levels of the random "subject" factor. And finally, inferences about the random effect of "subject" in general pertain to the variance in

amplitudes expected across levels of the random "subject" factor.

Types of random effects. There are two primary types of random effects in common use: random intercepts and random slopes (see figure 1 in Barr et al, 2013).

Random intercepts account for how mean outcomes differ between levels of a random factor. In our thought experiment above, expected N400 amplitude (the outcome measure) was adjusted up or down depending on the idiosyncrasies of each participant. In that regression model, adding or subtracting from the outcome measure is equivalent to adding or subtracting from the intercept used in the model. The intercept is always equivalent to the outcome measure (ERP amplitude) when all other predictor variables are equal to 0. It's the "baseline" estimate. Now, assume that the word frequency variable in the thought experiment was centered. The intercept of the model would then be the mean N400 amplitude across all of the data. Since the BLUPs for the random effect of subject capture the extent to which we expect that any particular subject has N400 amplitudes above or below average, we simply adjust the overall intercept by the BLUP in order to get our expected N400 amplitudes for that particular subject. Thus, this type of by-subjects adjustment is called a "random intercept". The remaining fixed effect is then the contribution of frequency to the N400 that remains after accounting for by-subject differences in mean N400 amplitudes. This reflects our understanding that our sample of participants is drawn from a distribution of possible participants that vary in their mean N400 amplitudes.

Similarly, the random slopes account for how mean *effects* differ between levels of the random factor. In our thought experiment above, the expected *effect* of frequency

on N400 amplitudes (the outcome measure) could be adjusted up or down depending on the idiosyncrasies of each participant. Over a large number of subjects, we expect more frequent words to elicit smaller (less negative) N400 amplitudes than less frequent words. This frequency effect is an experimental manipulation that we expect to generalize across subjects, and is best treated as a fixed effect. However, some participants will show a frequency effect that is bigger or smaller than the average frequency effect. For example, one subject might show a 1μv decrease in N400 amplitude per unit increase in frequency, while a second subject might show a 0.5μv decrease in N400 amplitude per unit increase frequency. If we know a word's frequency and want to predict what N400 amplitude will be exhibited by a particular subject, it makes sense to adjust our prediction based on that subject's particular frequency effect. A "random slopes" factor allows the regression coefficient for frequency to be adjusted up or down for each participant. Again, all of these adjustments must sum to 0: a line that needs to show a bigger-than-average frequency effect for every participant is clearly not the line of best fit. Overall, random slopes reflect our understanding that the frequency effect observed in our sample of participants is drawn from a distribution of possible participants that vary in their N400 frequency effects. Not everyone will show exactly the average frequency effect, but the average frequency effect can be estimated from the data by determining the fixed effect of frequency that remains after accounting for the randomly different frequency effects observed in the sample.

It's important to note that while random intercepts just require specifying a single factor (e.g. "subject" or "item"), random slopes require specifying two factors, because

they capture how the effect of some fixed factor differs in the data across levels of a random factor. In typical nomenclature, a "by-subjects random slope for frequency" is a (fixed) frequency effect that is adjusted for each subject. Because the BLUPs sum to zero, it is important to note that a by-subjects random slope for frequency does not replace the fixed effect of frequency in a model. On the contrary, a primary use of mixed models is to find the marginal fixed effect that remains after accounting for other dependencies in the data (such as those induced by the units of sampling). So a near-complete random effects structure for the N400 frequency effect may include random intercepts for subjects and items, along with a by-subjects random slope for frequency, all included alongside the fixed effect of frequency, as in

$$Amplitude \sim (1|Item) + (1|Subject) + (Frequency|Subject) + Frequency + i$$

where random factors are enclosed in parentheses and the "|" indicates both the nature of the adjustment to the left (where a "1" indicates an adjustment to the intercept i and a factor name indicates and adjustment to the slope of the effect for that factor), and the units of sampling to the right (e.g. "subject" or "item").

This process for estimating random effects is different from a model that includes a fixed effects term for subject that interacts with the variable of interest. Though both ostensibly derive a subject-specific slope, there are theoretical and practical differences (Baayen et al., 2008). As described above, the precise method of model fitting is very different, though this primarily stems from the desired treatment of subjects: in random-effects model, subjects are assumed to be randomly sampled from a population, and thus hold important similarities to one another. In a practical sense, this assumption manifests

in the shrinkage factor for BLUPs. Point estimates for individual subjects will be closer to the estimated population mean in a random-effects treatment than a fixed-effects treatment (Bates, 2010). This employs the prior knowledge that any particular sample (i.e. item or subject) was drawn from the same population as other samples (items and subjects). In contrast, a fixed factor for subjects and items makes no such attribution, and there is no shrinkage towards the mean even for small samples. When one is more interested in best predicting the effects that novel subjects will show, rather than predicting the effects that particular subjects will show, a random-effects term for subjects is usually more suitable.

These random intercepts and random slopes tend to improve power, produce better estimates for particular subjects or items, and reduce the false discovery rate during the estimation of the remaining fixed effects (Baayen et al., 2008; Barr et al., 2013). They account for dependencies in the data that result from units of sampling, and alleviate some of the problems that arise from violating model assumptions without accounting for these dependencies. As such, they elegantly allow for repeated measures designs, retain all the benefits of the GLM, and are gaining increasing interest within psycholinguistics (Baayen et al., 2008; Barr et al., 2013).

Naturally, consideration of the units of sampling is not novel. In fact, it was field-standard in psycholinguistics even a half-century ago (Raaijmakers, Schrijnemakers, & Gremmen, 1999). Importantly, all ANOVAs already make assumptions about how sampling occurred across subjects and items. Though these assumptions are often tacit, it is completely possible to construct an ANOVA that accounts for random intercepts and

random slopes (Barr et al., 2013).

However, ANOVAs in psycholinguistics generally involve averaging over either items or subjects, and omitting a consideration of either dependency can lead to anticonservative bias (Baayen, Tweedie, & Schreuder, 2002). Unfortunately, however, ANOVAs can't account for both dependencies at once (Barr et al., 2013). The "minF" test was developed to remedy this deficiency (H. H. Clark, 1973; Raaijmakers et al., 1999). It involves calculating both an F-statistic averaged over items and an F-statistic averaged over subjects in order to derive the appropriate F-statistic for a study with a random sample of subjects and items without inflating the type I error. Because mixed-effects models can incorporate multiple random effects at the same time, they calculate directly what the minF test was designed to approximate: "crossed" random effects of subjects and items. Though there are a number of instances where a simple ANOVA is appropriate in psycholinguistics (such as when every subject is shown different randomly sampled stimuli), the use of crossed sampling of subjects and items is generally the norm in experimental design. For such instances, crossed random effects should also be the norm.

Guided Example: N400 Effects of Frequency and Concreteness

Below, we walk through an example of fitting a mixed-effect regression model to ERP data.

Data. For the present example, we use data from the two experiments described in

chapter 2. Data was collected from 46 participants reading 468 single words that systematically varied on a number of word features, including frequency, concreteness, and word length. Trials were time-locked to the onset of each word, and segmented from 100ms pre-onset to 800ms post-onset. Data was adjusted to a -100 to 0ms baseline, normalized to a subject-specific calibration (where recorded amplitudes on the day of the study were compared against the average of 200 known calibration pulses), and then subject to a 0.1-40Hz bandpass filter.

To provide for the use of regression models, individual trials were left unaveraged. Mean amplitudes were extracted for each trial in 25ms increments from 0ms to 800ms for all 29 electrodes. *A priori* time windows (such as the N400) were created by averaging across the pertinent time points (such as 300-500ms). These arrays of trial-level amplitudes at different time points were labeled with the subject code, the item code, and the sensor code, arranged in long form. Subject-level information (e.g. age, sex), item-level information (e.g. frequency, length, concreteness), and electrode-level information (e.g. region on the scalp) was then affixed to the ERP data for the corresponding subjects, items, and electrodes. This is the appropriate form for regression analysis, where each outcome data point (i.e. mean amplitude for each trial in a given spatiotemporal ROI) is listed alongside the pertinent predictors in the data frame.

Specific considerations for regression analyses are detailed further below. But briefly, the data for the present example was subject to a number of additional procedures to ensure data quality. Each predictor was checked for normality and outliers, all predictors were z-transformed to provide for informative intercepts and comparison of

coefficients, and sets of predictors were evaluated for multicollinearity. All models were fit using the lme4 package (Bates, 2010; Bates, Maechler, Bolker, & Walker, 2014) of the R statistical software program (R Core Team, 2014).

Hypotheses. The existing literature provides for strong *a priori* hypotheses for how frequency and concreteness might influence ERP amplitudes during the N400 time window.

First, we expect that more frequent words to be facilitated during lexico-semantic processing compared to infrequent words. We expect that more frequent words would elicit a smaller (less negative) N400 amplitude than infrequent words (Laszlo & Federmeier, 2014). This would appear as a *positive* relationship between Frequency and voltage: as frequency increases (i.e. has higher positive values), ERP amplitude gets *less* negative (i.e. gets *more* positive). Previous studies have indicated that the relationship between frequency and the N400 appears to be logarithmic (Hauk et al., 2006; Laszlo & Federmeier, 2014; Payne, Lee, & Federmeier, 2015). As such, frequency was first log-transformed, and log frequency values were then z-transformed.

Second, we expect that concrete words will elicit a larger (more negative) N400 amplitude than abstract words, as found by previous studies (Kounios & Holcomb, 1994). This would appear as a *negative* relationship between Concreteness and voltage: as concreteness increases (i.e. has higher positive concreteness ratings), ERP amplitude gets more negative (i.e. gets less positive).

These two effects - the N400 frequency effect and the concreteness effect - are not

novel. By design, we have a clear set of expectations by which to demonstrate the use of regression-based techniques. The positive effect of Frequency and the negative effect of Concreteness should overlap temporally, but should be clearly discernible from one another. One single regression model fits the marginal contributions of frequency and concreteness to the N400 at the same time, above and beyond the random idiosyncrasies of the particular subjects and items used for the study.

As such, the regression model is straightforward. The outcome variable is the N400 ERP amplitude, from 300-500ms. For simplicity, we confine our outcome measures to the centro-parietal mid-region where the N400 effect tends to be maximal (different treatments of spatial factors are discussed under "spatial factors" below). Each trial is thus reflected by a single value for the amplitude of the N400 within an *a priori* spatio-temporal ROI. The *a priori* fixed factors are the standardized log-frequency and concreteness measures. Though we could investigate the interaction between frequency and concreteness if desired, the present example simply assess their marginal additive effects on the N400 component.

The random-effects structure accounts for the dependencies in the data. It has been recommended to include the maximal random-effects structure (Barr et al., 2013) that is licensed by the design and the data (Bates, Kliegl, Vasishth, & Baayen, 2015). Here, we use a near-maximal random-effects design (the actual maximal random-effects structure is discussed under "random-effects structure" below). As the same items were shown to multiple subjects, a random intercept for item was included, and as each subject was shown multiple items, a random intercept for subject was included. Further, both the

frequency effect and the concreteness effect could conceivably differ from subject to subject. As such, we include a by-subjects random slopes for frequency and a by-subjects random slope for concreteness. Because the items (words) used for the present study all had fixed frequency and concreteness values, it does not make sense to include a by-item random slope for frequency or concreteness: more than one frequency value is required to compute a frequency effect on the N400, and thus the single frequency value for each word is insufficient to calculate a word-by-word adjustment to the average frequency effect.

In R, this model is specified as

$$N400 \sim (1|Word) + (1|Subject) + (0 + z.Log\_Freq|Subject) + (0 + z.Concreteness|Subject) + z.Log\_Freq + z.Concreteness$$

By default in lme4, the model is fit using a restricted maximum likelihood estimation (REML). Similar to the MLE procedures (described in appendix A), the REML procedure is not biased by the loss of the degrees of freedom necessary to estimate the fixed effects, and performs well with unbalanced designs (Harville, 1977; Patterson & Thompson, 1971). It is the recommended criterion for fitting mixed-effect regression models (Bates et al., 2014).

To actually run the model with lme4, it is identified as a linear mixed-effects regression using "lmer()", and a source of data is provided to fit the model. Here, we restrict the dataset ("d.r") to a single centro-parietal region (Region 3, as in figure 3) and save the fitted model to a variable ("model1") so that it can be called by other operations:

```
model1 <- lmer(N400 ~ (1|Word) + (1|Subject) + (0 + z.Log_Freq|Subject) + (0 +
        z.Concreteness|Subject) + z.Log_Freq + z.Concreteness,
        d.r[d.r$Region==3,])
```

Calling model1 yields the following output:

```
Linear mixed model fit by REML ['merModLmerTest']
Formula: N400 ~ (1 | Word) + (1 | Subject) + (0 + z.Log_Freq | Subject) +
   (0 + z.Concreteness | Subject) + z.Log_Freq + z.Concreteness
  Data: d.r[d.r$Region == 3, ]
REML criterion at convergence: 139629.4

Random effects:
 Groups    Name                Std.Dev.
 Word      (Intercept)         1.1287
 Subject   (Intercept)         3.7374
 Subject.1 z.Log_Freq          0.4178
 Subject.2 z.Concreteness      0.1963
 Residual                      9.1486
Number of obs: 19156, groups:  Word, 468; Subject, 46

Fixed Effects:
  (Intercept)     z.Log_Freq  z.Concreteness
       1.2520         0.3766        -0.5955
```

As can be seen, both the model and the procedure for model fitting are first described in the output. Then, the four random effects are listed. As described above, only a standard deviation is provided for the random-effects terms. The by-item and by-subject predictions sum to zero, so there is nothing comparable to a regression coefficient that is applicable to random effect terms. Instead, the standard deviations provide a summary of the extent to which the levels of the random factor were predicted to differ from one another, on average. The individual random-effect BLUPs for specific items and subjects can be retrieved by the “ranef” command: “ranef(model1)”. This returns the

by-items adjustment to the intercept for each word, and the by-subjects adjustments to the intercept and two fixed effects for each participant. The regression coefficients and model intercept for particular subjects and items (*after* all adjustments have been applied for that subject and item) can be called by the "coef" command: "coef(model1)". These item- and subject-specific values are not necessarily useful for inference about novel or generalized populations (as they are not generalizable across subjects and items), but allow for the determination of fitted models and predictions for specific items and subjects.

However, the primary goal at present is hypothesis testing. Though the point-estimates for each of the fixed effects are provided by the model (and interpreted as standardized betas, as the predictors are all z-transformed), there is no indication of the precision with which these values have been estimated, as necessary for inference. This is not accidental: in such mixed-effect models, while the test statistic can be calculated, the degrees of freedom of the model cannot be exactly specified, and there is no ideal estimation of degrees of freedom that allow for inference without additional assumptions. As such, there is significant disagreement about how best to calculate p-values for mixed models. We discuss some of this controversy and provide some recommended procedures under "computation of p-values" below. One such easily-computed (and suitably conservative under many conditions) approach is the Satterthwaite approximation, which can be called by default from the "summary()" command in the lmertest package (Kuznetsova, Brockhoff, & Christensen, 2015). This yields the following output for model1:

Linear mixed model fit by REML t-tests use Satterthwaite approximations to degrees of freedom [merModLmerTest]

Formula: N400 ~ (1 | Word) + (1 | Subject) + (0 + z.Log_Freq | Subject) + (0 + z.Concreteness | Subject) + z.Log_Freq + z.Concreteness

Data: d.r[d.r$Region == 3,]

REML criterion at convergence: 139629.4

Scaled residuals:

Min	1Q	Median	3Q	Max
-6.5834	-0.6356	-0.0067	0.6325	4.3037

Random effects:

Groups	Name	Variance	Std.Dev.
Word	(Intercept)	1.27391	1.1287
Subject	(Intercept)	13.96806	3.7374
Subject.1	z.Log_Freq	0.17458	0.4178
Subject.2	z.Concreteness	0.03852	0.1963
Residual		83.69625	9.1486

Number of obs: 19156, groups: Word, 468; Subject, 46

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.25201	0.55756	45.76000	2.246	0.029608 *
z.Log_Freq	0.37664	0.10463	73.03000	3.600	0.000577 ***
z.Concreteness	-0.59546	0.08913	86.58000	-6.680	2.2e-09 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:

	(Intr)	z.Lg_F
z.Log_Freq	0.000	
z.Concrtnss	0.000	0.006

As can be seen by comparing this output to the initial output above, all of the fitted estimates are the same. The summary command in lmertest added additional inferential statistics for the fixed effects. We can see from this that our hypotheses were supported. First, frequency showed a significant positive relationship with N400

amplitudes. This indicates that more frequent words tended to elicit less negative (i.e. smaller) N400 amplitudes than more frequent words, all other things held equal. In contrast, concreteness showed a significant negative relationship with the N400 amplitudes. This indicates that more concrete words tended to elicit more negative (i.e. larger) N400 amplitudes than less concrete (more abstract) words, all other things held equal.

This N400 analysis illustrates one important consideration when using regression models for the analysis of ERP data: the signs of the effects can be counterintuitive for negative-going ERP components. Intuitively, more frequent words would yield a "smaller" N400 component, and thus one might expect a negative relationship between frequency and N400 amplitudes. However, because the N400 component is a negativity, "smaller" amplitudes are in fact *less negative* amplitudes. As such, the relationship between frequency and measured ERP amplitudes is positive (where larger frequency values correspond with more positive / less negative ERP amplitudes). We prefer describing real amplitudes in this manner, though care must be taken to explain exactly how positive and negative signs are to be interpreted as they pertain to component directions. However, if desired, researchers may multiply the amplitudes by -1 when analyzing a negativity, such that a "smaller" N400 amplitude actually corresponds with numerically smaller values. The model estimation will be the same, but the point estimates will be of the opposite sign. Some may find this more intuitive. However, we caution that this critically changes the interpretation of one feature in particular: the intercept. When all other terms are standardized, the model intercept is typically the mean

amplitude across the experiment. If the amplitudes have been multiplied by -1, the model intercept is instead the *negative* of the mean (i.e. the distance from the baseline amplitude in the negative direction). Such an intercept does not have an obvious interpretation. As such, we prefer leaving the amplitudes in their untransformed state. However, researchers are ultimately recommended to choose whichever method most intuitively describes their effects, carefully specifying how the sign of the effects are meant to be interpreted.

A number of recent publications have utilized trial-level regression techniques for ERP analysis (Hauk et al., 2006; Hauk, Pulvermuller, Ford, Marslen-Wilson, & Davis, 2009; Laszlo & Federmeier, 2014), including some that have used mixed-effects regression (Payne et al., 2015). Recently, Smith and Kutas published a comprehensive framework for regression-based analysis of ERPs (Smith & Kutas, 2015a, 2015b), and though they do not incorporate random-effect terms, many of the same considerations apply, and the rigorous discussion in these articles can be helpful when analyzing ERPs using mixed-effect regression as well.

Specific Considerations

The use of mixed-effect regression for ERP data requires a number of careful considerations, many of which are unique to mixed models, and/or unique to ERPs. Some of these are discussed below.

Data quality. As with traditional ERP analysis, the quality of the data and the quality of the experimental design are of primary importance. A laboriously rigorous

analysis is comprehensively undone by compromised data. With ERPs, scientifically important inferences can be draw from very modest amplitude deflections: categorical effects of just 1-3 microvolts are not uncommon. At the single-trial level, this slight fixed effect can be embedded within "noise" (or more precisely, amplitude changes not completely attributable to stimulus processing) that is an order of magnitude larger. The signal-to-noise ratio is already profoundly unfavorable. Any increase in the level of noise can easily wash out real effects (though the regression techniques discussed here generally have more statistical power for typical psycholinguistic studies). But more critically, even small *biases* in the generation of this noise can lead to spurious effects. This is not unique to regression analysis, but it is important enough to bear repeating: good inferences require good data (rigorously discussed in Luck, 2014).

Normality. Normally distributed data (specifically, normally distributed errors) is still an assumption of the linear model, regardless of whether data are analyzed using ANOVAs or regressions. Informally, one might expect that a normally distributed outcome would not be well predicted by skewed predictors. As such, all predictors should be assessed for normality and transformed if appropriate, as is standard statistical practice. Formal tests for normality are beyond the scope of this paper, and are discussed extensively elsewhere (Thode, 2002), as are common transformations including the log and Box-Cox transformations. But in the very least, we encourage the use of visualizations prior to data analysis. The ggplot2 package for R (Wickham, 2009) provides very simple kernel density estimation procedures (Hastie et al., 2009), which plots the density of continuous variables without the need for binning a continuous

variable, as required for histograms.

In the example detailed above, kernel density plots clearly indicated that Concreteness was right-skewed. As all concreteness values are 1 or greater (specifically, concreteness was measured using a 7-point likert scale), we applied a natural log transformation prior to standardization. The transformed concreteness values were normally distributed. Here is model1 refit for these transformed concreteness values:

```
model2 <- lmer(N400 ~ (1|Word) + (1|Subject) + (0 + z.Log_Freq|Subject) + (0 +
      zlog.Concreteness|Subject) + z.Log_Freq + zlog.Concreteness,
      d.r[d.r$Region==3,])
```

With the following (trimmed) summary output:

```
Random effects:
 Groups    Name              Variance  Std.Dev.
 Word      (Intercept)       1.30814   1.1437
 Subject   (Intercept)       13.96566  3.7371
 Subject.1 z.Log_Freq        0.17433   0.4175
 Subject.2 zlog.Concreteness 0.03207   0.1791
 Residual                    83.70288  9.1489
Number of obs: 19156, groups:  Word, 468; Subject, 46

Fixed effects:
                  Estimate  Std. Error  df        t value  Pr(>|t|)
(Intercept)       1.25190   0.55758     45.78000  2.245    0.029624 *
z.Log_Freq        0.36991   0.10496     73.94000  3.524    0.000733 ***
zlog.Concreteness -0.56586  0.08876     88.53000  -6.375   8.04e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this case, the difference was minimal. But that cannot be assumed. The precise operationalization of predictors is an important consideration. In chapter 2, we implement a number of theoretically-motivated transformations. This includes the log transformation of concreteness, as described here, but also a discretization for a continuous variable

where the majority of values were either “0” or “1” (and thus the model included terms for “none”, “one”, and “many”).

Outliers. Single-trial ERP data may at times be contaminated with extreme values that passed through artifact rejection, such as large alpha waves. Outliers can skew regression estimates in the same way they skew ERP averages (as a condition average is identical to an ordinary regression over a nominal variable), but can additionally exert a considerable leverage on the estimation of random effects or linear regression through sparse data.

There are a number of ways to handle outliers. A recent best-practices review was published by Aguinis and colleagues (Aguinis, Gottfredson, & Joo, 2013), with simply transferable suggestions for ERP data. First and foremost, major outliers should be checked for “inaccuracies” (such as coding or processing errors). To identify these “error outliers”, Aguinis and colleagues suggest a distance cutoff of two standard deviations, roughly corresponding to the outer 5% of the data. As the trial-level ERP data sets can often be tens or hundreds of thousands of data points, we have increased this to four standard deviations in chapters 2 and 4. Additionally, we suggest that checking for “inaccuracies” in this case necessitates checking the original raw trial-level waveforms for trials identified as being potential outliers. It is possible that the outlier was the result of an artifact that was initially missed during processing. If so, there would be theoretically-motivated grounds for removal.

And second, they suggest investigating “interesting outliers”. Working through the actual waveforms for outlier trials could reveal compelling patterns explaining why

the values are so distinct from most of the other data. But we also recommend investigating larger patterns that may reside in the data through statistical means. It is possible that some of the outliers may not be attributable to *irreducible error*. They could instead be variance that is systematically explained by variables not previously considered: *reducible error*. Such trials should not typically be removed, and may even contribute to novel discoveries.

However, it remains a basic tenet of ERPs that the signal of interest is often buried within noise that can be an order of magnitude larger. If the expected effect is on the order of a microvolt or two, a trial 80 or 90 microvolts above the mean can be inferred to probably contain relatively less of the signal of interest, and relatively more noise. As such, in chapters 2 and 4, we overtly restrict our analysis to roughly the middle 99.99% of the data, rather than all of the data. This changes the nature of the interpretation, and must be reported in the analyses. Further, the exact cutoff is ultimately arbitrary. Nonetheless, we find that this small restriction can reduce the range of amplitudes for some components by nearly half, as very extreme values (with comparatively less influence of the signal) are removed.

Random-effects structure. As discussed in the example analysis above, the random-effects structure for model1 was only "near-maximal". We had incorporated random intercepts for subjects and items, and by-subject random slopes for the two variables, but there is one further covariance that is missing: the relationship between random intercept for subjects and the by-subject random slope predictions. Specifically, it is plausible that the same participants that show smaller N400 frequency effects also

generally show smaller N400 amplitudes - a correlation between the random intercept and the random slopes during estimation. For instance, a general attenuation of both effects and waveforms is the expected pattern when a subject does not pay attention to the task or experiment (Barr et al., 2013). In such cases, it is inadvisable to treat the by-subject random slopes and random intercepts as independent. Instead, we can treat them as being jointly drawn from a bivariate distribution, overtly modeling the covariance between the random effects (Baayen et al., 2008). This is very simple to specify in lme4, as such: "(1 + z.Log_Freq | Subject)". This calculates random intercepts, random slopes, and their covariance. As such, a separate term for random intercept for subjects is no longer required. The model specification would be as follows:

```
model3 <- lmer(N400 ~ (1|Word) + (1 + z.Log_Freq|Subject) + (1 +
        z.Concreteness|Subject) + z.Log_Freq + z.Concreteness,
        d.r[d.r$Region==3,])
```

And the (trimmed) lmertest summary output is:

```
Random effects:
 Groups    Name        Variance  Std.Dev.  Corr
 Word      (Intercept) 1.27425   1.1288
 Subject   (Intercept) 2.21589   1.4886
           z.Log_Freq  0.17447   0.4177    0.28
 Subject.1 (Intercept) 11.83108  3.4396
           z.Concreteness 0.03942 0.1985   0.33
 Residual              83.69540  9.1485
Number of obs: 19156, groups:  Word, 468; Subject, 46

Fixed effects:
                 Estimate  Std. Error  df          t value  Pr(>|t|)
(Intercept)      1.25181   0.55909     45.04000    2.239    0.030141 *
z.Log_Freq       0.37594   0.10462     73.10000    3.594    0.000589 ***
z.Concreteness -0.59687  0.08924     86.24000    -6.688   2.15e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared to model1, the random-effect structure for model3 includes an additional "Corr" column, reporting the correlation between the random intercepts and random slope predictions for the specified terms. As indicated by the residual variance of the two models, these covariances were probably not a hugely important source of variance. However, the inclusion of this covariance is recommended in many circumstances (Baayen et al., 2008; Barr et al., 2013; Bates et al., 2015), and the estimation of the fixed effects is influenced by the decision. As such, it is important to consider during analysis.

Bates and colleagues (Bates et al., 2015) have recently provided an excellent overview of considerations in determining the appropriate random-effects structure. Though failing to account for dependencies in the data can radically inflate the false-positive rate (Barr et al., 2013), stretching the model complexity beyond what is licensed by the data can also lead to inappropriate interpretations. They suggest that an iterative reduction in model complexity for the random-effects structure can protect against such risks.

Computation of p-values. There is significant disagreement about the best way to conduct inferential hypothesis tests using mixed models. Specifically, it is unclear how exactly p-values should be calculated for individual terms, as the degrees of freedom (and thus the distribution of the test statistic) are not precisely defined for a mixed model. Though an influential paper by Barr and colleagues suggested deriving the distribution of the test statistic through Markov Chain Monte Carlo sampling (Barr et al., 2013), this

capability has since been removed by the creators of lme4 for being too unreliable (Bates et al., 2014). They have instead partially implemented a model-based bootstrapping procedure (Bates et al., 2014). This estimates standard error by iteratively resampling from the data, with the option to either assume sphericity in the random-effect terms (parametric bootstrapping) or sample from the residuals (semi-parametric bootstrapping). However, these approaches are somewhat experimental for this particular application, and may be anticonservative in some cases (J. S. Morris, 2002).

Pending further development of bootstrapping approaches, we suggest other procedures where the degrees of freedom can be estimated using well-understood and relatively conservative approximations (Bates et al., 2014). The "lmertest" package implements two common approximations for mixed-effects models (Kuznetsova et al., 2015).

First, degrees of freedom can be estimated using the Welch-Satterthwaite approximation (Satterthwaite, 1946; Welch, 1947). The degrees of freedom are derived from a method for pooling standard error in cases where there is more than one variance component. This is the default method used by the lmertest package (as of version 2.0-29). When lmertest is loaded into R, it masks the "summary" command from lme4 and instead adds the approximated p-values to the output, along with additional information. The Welch-Satterthwaite approximation for the fixed-effects terms in model1 is as follows:

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.25201	0.55756	45.76000	2.246	0.029608 *

```
z.Log_Freq      0.37664        0.10463        73.03000   3.600        0.000577 ***
z.Concreteness -0.59546        0.08913        86.58000  -6.680        2.2e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As can be seen here, this approach provides an approximated degrees of freedom from which standard error and a p-value can be calculated (given the test statistic). Both frequency and concreteness significantly influence the N400 component over the central mid-region.

Second, p-values can be derived from a Kenward-Roger degrees of freedom approximation (Kenward & Roger, 1997). Essentially, the Kenward-Roger approach adjusts a Wald statistic to reduce bias due to smaller sampling sizes, a scaling that behaves exceptionally well under a variety of circumstances (Chawla, Maiti, & Sinha). It is also intuitive, actually producing the exact F test when used for fixed-effect linear models or mixed-effect ANOVAs (Alnosaier, 2007). However, it can be more computationally intense than the Satterthwaite approximation, and is generally more conservative. We recommend it primarily for formal *a priori* hypothesis testing for slightly smaller models in cases where a conservative approach is desired, while the Satterthwaite approximation may be sufficient for many instances of data exploration or big-data analyses. The lmertest package utilizes a Kenward-Roger approximation as implemented by the pbkrtest package (Halekoh & Højsgaard, 2014). It can be utilized in lmertest by specifying the Kenward-Roger approximation during the production of a model summary, e.g. “summary(model1, ddf="Kenward-Roger")”.

Model selection. Defining a model appropriate for both the question and the data can be difficult. Each term added to the model changes the interpretation and the estimates for every other term, and it's very easy to overfit a model by adding too much complexity. However, there are a number of simple considerations we find helpful.

First, clarity is required about the goal of the model: is it going to be used to predict, or to explain (Shmueli, 2010)? In Shmueli's formulation, explanatory modeling is the practice of using data to test causal hypotheses about particular theories. Most of our efforts in psycholinguistics generally pertain to explanation: we are interested in the ERP effects elicited by novel manipulations of items and variables in order to understand how stimuli are processed. In contrast, predictive modeling is the practice of using existing data to predict future data with the highest precision. Many machine-learning algorithms (such as such as gradient boosting, random forest models, and vowpal wabbit) are typically used for predictive purposes,. These models often sacrifice interpretability for predictive capability, and many of the most accurate models in refereed competitions (including for EEG data, see Barachant & Cycon, 2015) are weighted ensembles of different composite techniques, where formal inference may be unfeasible in any practical sense, but predictive capability is higher than even the most state-of-the-art academic solution.

Mixed-effect regression can be used effectively both to explain and predict. But the two require different approaches to model building.

Explanation, specifically formal hypothesis-testing, typically involves inferring a particular causal relationship between a predictor and the outcome, above and beyond

certain alternative explanations. Multiple regression is well suited for hypothesis testing. Any particular effect is interpreted as the contribution of that factor in explaining the variance in the outcome when all else in the model is held equal. For example, in model1, the frequency effect is marginal to the concreteness effect, and vice versa. If further terms are added to the model to account for further alternative explanations for the observed effect, the overall model fit is expected to increase. In this way, we can infer the marginal effect of some predictor holding a number of other nuisance factors equal. This provides a theoretical motivation for more expansive models, where nuisance factors can be controlled.

But the exact impact of including additional predictors on the estimation of other terms in the model depends on the nature of the relationship between the predictors (Wurm & Fisicaro, 2014). Some correlated terms can suppress estimates, while others may magnify estimates. To illustrate, here is model1 after adding the (log-transformed) bigram frequency ("zlog.N2_F"):

```
model4 <- lmer(N400 ~ (1|Word) + (1|Subject) + (0 + z.Log_Freq|Subject) + (0 +
        zlog.Concreteness|Subject) + z.Log_Freq + z.Concreteness + zlog.N2_F,
        d.r[d.r$Region==3,])
```

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.25194	0.55746	45.70000	2.246	0.029594 *
z.Log_Freq	0.41431	0.10528	74.60000	3.935	0.000185 ***
z.Concreteness	-0.61362	0.08806	89.60000	-6.968	5.22e-10 ***
zlog.N2_F	-0.21660	0.08525	462.2000	-2.541	0.011390 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

After accounting for the error attributable to the bigram frequency effect (where words with higher-frequency orthographic competitors yield larger N400 amplitudes), both the lexical frequency effect and the concreteness effect were estimated to be larger. However, consider when word length is also added to the model ("z.Length"):

```
model5 <- lmer(N400 ~ (1|Word) + (1|Subject) + (0 + z.Log_Freq|Subject) + (0 +
        zlog.Concreteness|Subject) + z.Log_Freq + z.Concreteness + zlog.N2_F +
        z.Length, d.r[d.r$Region==3,])
```

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	1.25228	0.55737	45.70000	2.247	0.0295 *
z.Log_Freq	0.44656	0.10427	72.10000	4.283	5.60e-05 ***
z.Concreteness	-0.56577	0.08725	86.90000	-6.484	5.23e-09 ***
zlog.N2_F	-0.08900	0.08881	460.80000	-1.002	0.3168
z.Length	0.38070	0.08857	460.90000	4.298	2.10e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

In this data set, concreteness and length correlated with r = 0.128, and when length was added as a predictor, the *marginal* effect of concreteness was lower than with model4. Similarly, length and bigram frequency correlate with r=0.334, and when length was added as a predictor, the *marginal* effect of bigram frequency was attenuated to nonsignificance. In essence, the explanatory variance shared between them was attributed to neither. This is only true of the individual model terms: the *overall* model fit was superior for model5 than for model4, as expected when more explanatory variables (and

model complexity) is added. This can even be shown formally with lmertest, where the "anova()" command can perform a chi-squared test on the overall fit of two models to determine if there exists a difference.

When performing explanatory modeling, it is recommended to include the nuisance factors that are required for the desired inferences. For example, if we simply wanted to infer that frequency and concreteness have an opposing influence on the N400 across a large sample of the English language, model1 is sufficient. If we instead are interested in the precise nature of the N400 frequency effect *after* having controlled for a number of nuisance factors, models such as model5 may be more appropriate.

Importantly, because of the influence on the resulting inferences, it is actually appropriate in many cases to *leave even nonsignificant nuisance factors in the model,* so long as those factors are theoretically motivated. This should become apparent when considering *which* inference is actually desired. For instance, even if there is no evidence from model5 that bigram frequency has any marginal effect on the N400 above and beyond the effects of concreteness, frequency, and length, we still remain interested in whether *frequency* has any marginal effect on the N400 above and beyond the effects of concreteness, bigram frequency, and length. In other words, whether or not a nuisance factor is significant in the presence of the other predictors is often actually mostly irrelevant to the inference of interest, which is whether the experimental effect is significant in the presence of the nuisance factors.

However, it can still be justified in some cases to drop a nuisance factor. For instance, if the addition of a predictor clearly does not result in a better model fit. Though

it is typically preferable to define an *a priori* model for hypothesis testing, there may be cases where a certain predictor is both clearly unrelated to the outcome and the model as whole is too complex for the size of the data set. If the model complexity should be reduced (Baayen et al., 2008; Barr et al., 2013; Bates, 2010; Bates et al., 2015; Bates et al., 2014; Wurm & Fisicaro, 2014), such terms could be justified in being dropped first. However, we note that in explanatory modeling, main effects should not be dropped prior to interaction terms involving that effect, as this complicates the interpretability that is the primary focus of explanatory modeling (Shmueli, 2010).

Because of these considerations, there will also be cases when two factors may each be significant on their own, but both will be non-significant when both are included in the model. This is common for highly correlated predictors (see discussion of multicollinearity in Wurm & Fisicaro, 2014). This is a feature, not a bug, and should inform the inferences that are drawn. Suppose that two factors (x1 and x2) are highly correlated with each other and with the outcome, but neither show marginal significance when included in a model together. One could interpret this to mean that the variance shared by x1 and x2 predicted the outcome, but that the data were insufficient to distinguish their independent contributions (i.e. the variance unique to each variable).

In contrast, predictive modeling requires a significantly more restrictive approach to model complexity (Shmueli, 2010). The primary motivation is to identify which features would best predict future data. As such, non-significant terms are often removed. But more incisively, k-fold cross validation can be used to isolate which terms (and which models) actually predict novel data the most accurately regardless of statistical

significance. For the large datasets common with ERP analysis, k-fold cross validation may become unwieldy. In such cases, it may be appropriate to instead divide the data into a test set and a validation set.

However, it is also common in predictive modeling to loosen the theoretical constraints on feature selection. Automated model selection procedures like the lasso (recently implemented for mixed-effect models in the 'glmmLasso' package) can be used to identify suitable predictors out of a larger feature set. In predictive modeling, it is recommended to start with more information rather than less, and to settle on whichever model minimizes bias and variance. This contrasts with explanatory modeling, where terms are generally interpretable and feature selection is strongly theoretically motivated.

Further considerations of the differences between explanatory modeling and predictive modeling are discussed elsewhere (Shmueli, 2010). Mixed-effect models are valid approaches to both pursuits, including for ERP data. In the present dissertation, we fit both explanatory models and predictive models in chapter 2, and fit only explanatory models in chapter 4.

Spatial factors. In a trial-level ERP data set, each trial will have measures taken from multiple electrodes across the surface of the scalp. The measures from different electrodes have a number of clear patterns: 1) data from any single electrode is likely to resemble other data from that same electrode more so than data from other electrodes, 2) data from nearby electrodes are likely to resemble data from one another more than data from distant electrodes, 3) data from different electrodes in a particular trial are more likely to resemble data from other electrodes during that trial than data from different

trials, and 4) different sensors measure underlying dipole patterns in structured (though not necessarily simple) ways.

It is not clear how best to address these contingencies. There are three primary candidate approaches that have been discussed in the field. First, one can confine analysis to an *a priori* region of interest (ROI), averaging electrodes within that region together at the level of single trials. Consequently, each trial is reflected by a single amplitude measure. Second, one can input electrode (or region) as a fixed effect that interacts with other effects of interest. In this approach, spatial distribution is tested overtly, and the effect of interest is modeled at each electrode or region separately (in addition to the marginal main effects). And finally, one can input electrode as random factor, adjusting the model for the spatial contingencies along sensors.

The treatment of spatial factors as a random effect was recently reported in a rigorous mixed-effect regression analysis of word processing as sentences incrementally unfolded through time (Payne et al., 2015). Among their results was a robust categorical effect of words in coherent versus scrambled sentences on the N400 component, particularly later in sentences, as was the strong *a priori* expectation. In collaboration with original authors Payne and Federmeier, with additional contributions from Dr. Mallory Stites, we utilized the publicly available data from Payne and colleagues to explore the practical consequences of different treatments of the spatial factors.

Specifically, we performed a five-step permutation procedure, as follows. First, we restricted the data set for just the words in the final 25% of scrambled and coherent sentences, where the effect of coherence was largest, in order to allow for faster model

fitting. Second, we randomly selected two electrodes that were used in the original analysis, and restricted the data to just those electrodes. Third, we randomly permuted all of condition labels across sentences, such that the null hypothesis was literally true, and fit all three candidate models (averaged electrodes, fixed-effect electrodes, and random-effect electrodes). Fourth, we repeated steps 2-4 for 10,000 iterations. And fifth, we repeated this whole procedure for two, four, six, and eight electrodes.

We calculated the false positive rate for each of the three models by determining the proportion of t-statistics from each model that exceeded a critical t for that model as determined by a Satterthwaite approximation. As the alpha for the critical t was 0.05, we hoped to see a 5% false positive rate (i.e. 500 false positives from each set of 10,000 random permutations). Lower than that, and the test may have been too conservative. Higher than that, and the test may have been anticonservative.

The results are shown in figure 1. The averaged electrode approach yielded false positive rates of 3.6%-4.2%, and as such may have been conservative. It did not appear to be influenced by the number of electrodes that were averaged together. The fixed-effect electrode models yielded false positive rates of 4.1% - 7.1%, and were generally closest to the 5% target. However, the false positive rate appeared to grow incrementally along with the number of electrodes. Finally, the false positive rate for random-effect electrode models yielded false positive rates of 7.6%-14.8%, and as such may have been anticonservative. Even more so than the fixed-effect electrode model, the false positive rate for the random-effect electrode model appeared to increase with the number of electrodes.

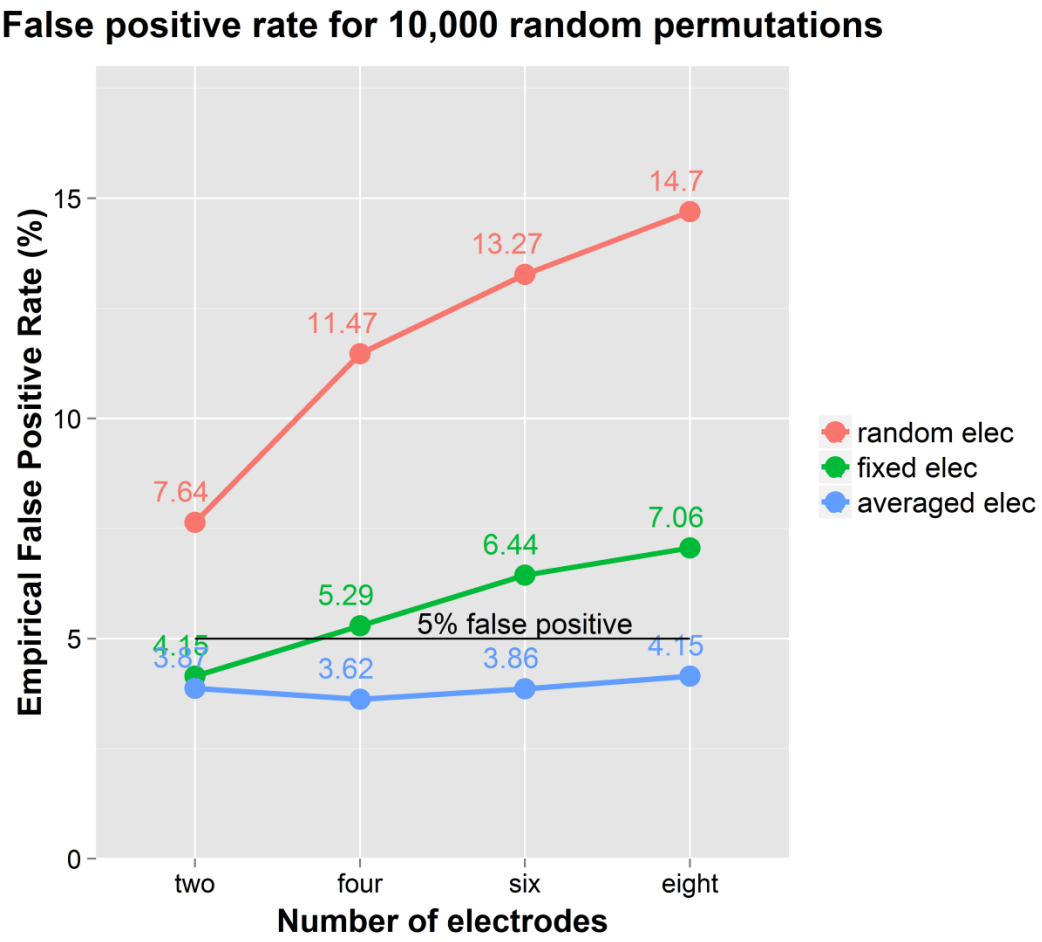

Figure 1 – False positive rate for different spatial treatments in 10,000 random permutations. This shows the proportion of tests exceeding the Satterthwaite-approximated critical t values for a categorical effect where the null hypothesis is known to be true, using published ERP data. Comparable models were fit to the same permutations where electrode was treated as a random factor, a fixed factor, or averaged together into a single region. Random-factor electrode models were generally anticonservative, averaged electrodes were generally slightly anticonservative, and the conservativeness of the fixed electrode treatment depended on the number of electrodes.

Using the Satterthwaite approximation, there was no ideal treatment of spatial factors. Averaging electrodes together tended to decrease power and erred on the side of being overly conservative. In contrast, treating unaveraged electrodes as a random factor

tended to be anticonservative (for the marginal main effect of interest), and could lead to false positives. Finally, entering unaveraged electrodes as a fixed factor was slightly conservative for small numbers of electrodes, but was slightly anticonservative for large numbers of electrodes.

However, we caution that these exact false positive rates are not intended to generalize to other data sets and experimental designs. The models included here were intentionally simplified, and the data set was restricted, in order to improve the speed of iteration (the analysis still required multiple days of computation time even when being processed in parallel on four to eight processing cores). As such, there is no definitive recommendation that can be drawn from these data. However, these data still suggest that averaging electrodes into regions, and/or using small numbers of electrodes (or regions) as fixed factors, may be useful in cases where a conservative approach is desired, depending on design and desired inferences.

It has not escaped our notice that the permutation technique we utilized here could also be utilized to derive empirical test statistic distributions. For instance, the actual effect of sentence coherence for this restricted data (and simple model) for eight electrodes yielded a $t = -8.380$. As shown above, the p-value generated by the Satterthwaite approximation was likely anticonservative (we found that it had a false positive rate of nearly 15% for eight electrodes). So instead, we can generate 10,000 t statistics from the permuted data (as above) and calculate the critical t that leaves 5% of permutations across the upper and lower tails (combined), consistent with a two-tailed hypothesis test. This critical t value, for eight electrodes and the simple model utilized for

this test, was found to be 2.835. Because the test statistic -8.380 was more extreme than the empirical critical t +/-2.835, the null hypothesis can be rejected. A similar calculation can be used to derive an approximate p-value. Specifically, we can calculate the proportion of trials generated by the null hypothesis that exceeded the test statistic of t = +/- 8.380. In this case, the permutation test never once generated a single value this extreme. We can infer from this that the effect was highly significant, with $p < 0.001$. As discussed in the original paper, the effect of sentence coherence is extremely robust (Payne et al., 2015), and we replicate this result using this permutation test.

However, this approach has some drawbacks. For one, it is somewhat impractical. We reduced model complexity and restricted the data set significantly in order to hasten iteration time, in ways that we discourage for actual hypothesis testing. Conducting permutation tests in this manner for more reasonable hypothesis tests could take significant computation time, to the extent that it may be prohibitively unwieldy for use in actual ERP analysis. We estimate that testing the coherence effect with the actual published model using this permutation procedure would take about five weeks on our standard computers for each hypothesis test (i.e. each term in each model, as only one term is permuted at a time).

Temporal factors. Most of our discussion so far has been confined to single time-windows, such as the 300-500ms time window typical for the N400 component. But part of the strengths of ERP data is the ability to investigate how stimulus processing unfolds over time after stimulus onset. There is no single best method for how to handle these temporal features in a regression analysis. However, we have a number of suggested

considerations.

First, one could simply derive the mean average (across time) for some theoretically motivated time window. This approach has a number of strengths (Luck, 2014), which should be familiar to most ERP researchers from the usual analytic approaches. In chapters 2 and 4, we begin our analysis with *a priori* time windows, based on the substantial previous literature on those components (specifically the N400). However, other approaches can also be used to identify meaningful time windows to average over, as discussed in detail elsewhere (Luck, 2014). These include the use of empirical localizers, temporal decomposition, and the treatment of time as a fixed factor.

Second, one could also iteratively fit a single model over successive time points. There are two related applications of this approach. One is the mass univariate approach, where an effect is computed at every possible time point (Groppe, Urbach, & Kutas, 2011a, 2011b; Hauk et al., 2006; Hauk et al., 2009), along with related regression-based approaches like the rERP framework (Smith & Kutas, 2015a, 2015b). This has a number of strengths, as discussed extensively elsewhere (Groppe et al., 2011b; Smith & Kutas, 2015a). However, the rERP framework does not, to our knowledge, support crossed random effects at this time (as described above).

But alternatively, it is also feasible to approximate such frameworks manually, by averaging time into short bins and computing the effect for each bin. This allows for full model specification, including random-effect terms. This is the approach we take in chapter 2. Specifically, we find the mean amplitude for 25ms time windows between 0 and 800ms. This fits a manageable 32 models. For visualization, the point estimates and

confidence intervals can plotted for each window, as we have done in chapter 2. This partially overcomes one of the shortcomings of a regression-based analytic approach: the challenge of visualizing an effect over time (as there is no elegant corollary to "grand mean" waveform for data that is not averaged into categories). For inference, however, the confidence intervals should be subject to a multiple-comparisons correction. There is no single correction that is guaranteed to yield false positive rate of exactly 5% across a family of tests. We generally recommend erring on the side of more rather than less conservative corrections (though the issue is thoroughly discussed elsewhere). In chapter 2, we implement a very conservative Bonferroni correction that was sufficient for our desired inferences.

Sequence and adaptation. Finally, regression models have the capability to incorporate information not only about the items, subjects, and contexts for which any trial-level data point was collected, but can also reflect information about the manner in which these trials were ordered.

Traditionally, this has been reflected through categorical effects of sequence. For instance, reaction time in a Stroop task (Stroop, 1935) critically depends on the nature of the preceding trial: Stroop effects are smaller following an incongruent trial than a congruent trial (Blais, Stefanidi, & Brewer, 2014; Gratton, Coles, & Donchin, 1992). Such analyses are very simple in a trial-level regression analysis of ERP data: one simply enters a sequence factor into the model, reflecting the identity of previous trials (or their relationship to the present trial).

Further, categorical effects of block can also be included. These can include

manipulations at the level of blocks, but can also include cases where the blocks are essentially the same, but differ in time (where later blocks are presented later in an experiment). These block effects can be included as categorical factor in the model.

Additionally, such sequence effects can manifest in ERPs even over long distances. For instance, the repetition of words in an experiment can attenuate the N400 frequency effect (discussed above) to nonsignificance (Rugg, 1990), despite the repetition occurring only after six other intervening items. In a regression model, this could be addressed by including the repetition number and/or the number of trials since the last repetition etc.

However, there are a number of other ordering effects that are possible in a trial-level regression. For example, an appropriately-designed experiment can allow for the assessment of practice and fatigue effects across an experiment. These can be significant. For instance, a recent metanalysis found that experiments using a greater number of blocks (allowing their participants more frequent rests) found larger behavioral priming effects than otherwise equivalent experiments using a smaller number of blocks (Herring et al., 2013). One possible contributor to this pattern of effects could include a block-final fatigue effect that attenuated the priming effect later within blocks. In a regression model, this could be addressed considering the location of the trial within the block.

Further, experimental effects are also often *expected* to change over time, according to well-described theory. For example, predictions utilized during language processing are thought to adapt over time to the specifics of the input (Fine, Jaeger, Farmer, & Qian, 2013; Kuperberg & Jaeger, In Press; Qian, Jaeger, & Aslin, 2012). Such

adaptation over the course of an experiment can be modeled with trial-level regression. We present two methods for doing this in chapter 4. First, we visualize the change in amplitudes over trials within a block using nonparametric local regression. And second, we implement a polynomial approximation to the adaptation. However, we also suggest that a number of other approaches to the assessment of adaptation may be useful in the study of ERPs, including generalized additive models (Baayen, Tremblay, & Hendrix, 2010a, 2010b) and the implementation of specific theories about the "state" of learning or adaptation at particular times in the experiment (depending on the preceding trials).

Conclusions

In sum, a mixed-effect regression approach to ERP analysis shares a number of similarities to the common subjects ANOVAs and items regressions that are already in wide use. But it offers additional capabilities beyond these common approaches as well, which may be valuable in addressing a number of the specific questions presently facing researchers in psycholinguistics and affective neuroscience. The present dissertation investigates a number of such questions.

In chapter II, we use mixed-effects regression to investigate the continuous and progressive influence of emotion on the N400 word frequency effect (a phenomenon we call the "avalanche effect"), isolate the influence of valence from the influence of arousal, and build a multi-level model to better understand how the avalanche effect patterns across tasks. In addition to formal hypothesis tests, we also build a predictive model of the N400 to derive more generalizable effects in real units, and iteratively refit this model across time windows and electrodes to characterize the exact temporal extent of the

effect.

In chapter IV, we use both local regression and mixed-effect modeling techniques to explore how the N400 semantic priming effect adapts over time to the statistics of the local experimental context. This involves polynomial approximation of learning curves and overt model comparisons. We also characterize the evolution of ERP amplitudes over time, finding evidence of possible fatigue or experiment-level expectation effects.

For chapter III, we chose not to utilize regression-based approaches to ERP analysis. Actual item identity was held constant and carefully counterbalanced in a categorical priming manipulation, diminishing the need for modeling sampling at the level of items.

CHAPTER II

THE AVALANCHE EFFECT: HOW EMOTIONAL WORDS ARE RARELY RARE TO YOUR BRAIN

It has long been suggested that emotional information can at times be "privileged" over other information during stimulus processing — the so-called *affective primacy hypothesis*. Across a number of tasks, emotional stimuli tend to elicit behavioral responses far faster than otherwise equivalent neutral stimuli (Bradley & Lang, 2007; Carretie, 2014; Kunst-Wilson & Zajonc, 1980; Lang & Bradley, 2009; Monahan, Murphy, & Zajonc, 2000). This includes linguistic stimuli, where emotional words show a behavioral advantage in reading time and lexical decision tasks (Briesemeister, Kuchinke, & Jacobs, 2011; Kousta, Vinson, & Vigliocco, 2009). Some have suggested that the ability to rapidly respond to salient or motivationally significant stimuli could carry adaptive significance (LeDoux, 2012; Zajonc, 1980), where the ability to prioritize affective information can be important to survival. This ability to rapidly evaluate the significance of salient stimuli is particularly critical when that stimulus is unexpected, such as for the random occurrence of an infrequent stimulus or event. In the present study, we explored whether this affective primacy manifests as facilitation of semantic processing for infrequent emotional words.

Comprehending a word necessitates the access of stored knowledge about that word's meaning, and it is apparent that not every word's meaning is typically retrieved with same ease. For instance, the meaning of more frequent words is generally retrieved more rapidly and with less effort than less frequent words (Solomon & Howes, 1951).

This frequency effect has been found to robustly influence reading times and lexical decision speed above and beyond features of the wordform like length and number of syllables or phonemes (Chumbley & Balota, 1984; Inhoff & Rayner, 1986), and eye-tracking studies have indicated that more frequent words evoke shorter fixations (Rayner & Duffy, 1986). This is attributable to the expectation by participants that the distribution of words in an experiment will mostly correspond with the distribution of words encountered in everyday life (Laszlo & Federmeier, 2014; Smith & Levy, 2013). Essentially, a word's frequency is equivalent to that word's probably of being encountered; frequency acts as a "default prior".

In addition, it may be advantageous in some circumstances to hold a prior expectation for emotionally significant stimuli, remaining alert for potential threats or appetitive opportunities. There is indeed some indication that emotional content can specifically facilitate the processing even of words (where the content is emotionally significant but the stimulus is just a symbol for the content). The processing of a word, like "avalanche", not only activates stored knowledge about the attributes and necessary preconditions for avalanches, but also activates learned emotional implications. Such affective information may be prioritized during processing under some conditions. If "avalanche!" is shouted in your direction, the emotional features (and thus motivational significance) may be extracted faster than would otherwise be expected (e.g. given that word's frequency, length, orthographic neighborhood etc.).

Some recent behavioral studies have found evidence of such facilitation for infrequent emotional words. In lexical decision tasks, larger frequency effects have been

reported for neutral words than for positive or negative words (Kuchinke et al., 2007), where high-frequency words were all responded to with comparable speed, but low-frequency emotional words were facilitated compared to low-frequency neutral words (see also behavioral data reported by Mendez-Bertolo et al., 2011). Another recent study (Kahan & Hely, 2008) also found that infrequent negative words led to disproportionate slowing during a stroop task, indicating that the meaning of infrequent negative words may have inordinately dominated attention even though participants were tasked with reporting the color of the text and ignore the meaning of the words. We will refer to this processing advantage for infrequent emotional (vs. neutral) words the "avalanche effect".

From one perspective, the "avalanche effect" is simply a bias in expectation for emotional features of the words. A similar effect was recently reported in the comprehension of emotional discourses (Delaney-Busch & Kuperberg, 2013), where an N400 cloze effect (a measure of lexical expectancy given a particular context) was attenuated for emotional words in emotional contexts. If the frequency effect essentially reflects default prior expectations, we might expect emotional words to be differentially facilitated in contexts that emphasize emotional features (Lai, Hagoort, & Casasanto, 2012) even if unexpected (i.e. even if it is infrequent during daily life), just like unexpected emotional words were facilitated in sentences (Delaney-Busch & Kuperberg, 2013) that emphasized emotional features, despite being semantically unpredictable.

The N400 event-related potential (ERP) component provides evidence that word frequency can strongly influence semantic processing of words. The N400 component is generally thought to reflect the neural activity related to processing a word's meaning

given the constraints and expectations elicited by the broader context (Kutas & Federmeier, 2011; Kutas et al., 2006). Word frequency has been shown to have a direct logarithmic relationship to N400 amplitude, such that more frequent words elicit a smaller N400 amplitude than less frequent words (Hauk et al., 2006; Hauk & Pulvermuller, 2004; Laszlo & Federmeier, 2014; Payne et al., 2015). Highly frequent words are easy to retrieve essentially because they are more probable given the relatively uninformative context (Fischer-Baum, Dickson, & Federmeier, 2014; Norris, 2006): the probability of occurrence is expected to match previous experience (Rabovsky & McRae, 2014; Smith & Levy, 2013). And words that are more expected (i.e. frequent) tend elicit smaller N400 amplitudes than words that are less expected. When the context is more informative, such as for supportive sentences, the N400 frequency effect attenuates (Payne et al., 2015; Van Petten & Kutas, 1990b, 1991) because the "default prior" is no longer a suitable expectation for which words are likely to occur next.

Some recent ERP studies have investigated the possibility that low-frequency words could still be efficiently processed if they were sufficiently emotional. One study (Mendez-Bertolo et al., 2011) crossed Emotion (Unpleasant versus Neutral) with Frequency (High versus Low) in a lexical decision task. High frequency words were all responded to at the same speed, but low-frequency unpleasant words were responded to faster than low-frequency neutral words. This corresponded with ERP results peaking around 450ms:[2] neutral words showed a significantly larger frequency effect than

[2] The authors suggest that the component found in this experiment may be a positivity related to the P300/LPP component, as indicated by a factor-analytic decomposition of

unpleasant words, which appeared to be processed with relative ease even if infrequent. One other ERP study (Scott et al., 2009) found that pleasant words yielded smaller frequency effects across a number of early components than neutral words, and appeared to do so during the N400 time window of the grand mean waveforms as well, although statistics in this time window were not reported. While these provide some evidence that the avalanche effect actually occurs during language processing in some situations, these studies leave a number of open questions, and neither have definitively shown an avalanche effect on the N400 component, where frequency effects are expected to occur.

The present investigation was designed to address this gap, and better characterize the avalanche effect. Specifically, we recorded ERPs as participants read a large number of single words, presented in isolation, that varied considerably by both frequency and emotion. In particular, we focused on an a priori N400 time window, conforming to well-established precedents of the frequency effect (Laszlo & Federmeier, 2014) and the functional significance of the component is indicating the activity related to semantic processing (Kutas & Federmeier, 2011). Regression-based analyses of trial-level ERP data allowed us to address a number of open questions about the nature of the avalanche effect. The data set is large (46 participants viewing 468 words) and carefully controlled, the analysis is novel for the topic, and the inclusion of high-arousal neutral words allowed us to disentangle the influence of valence from that of arousal. These data were

---

the spatial and temporal covariances. However, the waveforms show an effect in this time window that appears to be similar to N400 frequency effects from other studies. We suggest that the frequency-arousal interaction in this time window could potentially be interpretable as such.

previously collected for an investigation of the emotional LPC (Delaney-Busch, 2013; Delaney-Busch et al., *under review*), but the investigation of the avalanche effect is novel, and is the primary contribution of this dissertation chapter.

First, it's possible that the avalanche effect is primarily driven by only a particular set of emotional words that categorically violate the frequency effect in certain experimental tasks or contexts. If this is the case, then some emotional words would show the usual N400 frequency effect and others would not. For example, taboo words (Fogel, Midgley, Delaney-Busch, & Holcomb, 2013; Guillet & Arndt, 2009), emotion labels like "happy" (Briesemeister, Kuchinke, & Jacobs, 2014), and biologically relevant words (Aquino & Arnell, 2007; Arnell, Killman, & Fijavz, 2007) have all been implicated as "special" groups within affective space that elicit categorically different effects than other words, though there is also some contrary evidence (Vinson, Ponari, & Vigliocco, 2014). But it's also possible that the frequency effect on the N400 monotonically decreases with emotionality: a difference in scale rather than a difference in kind. In other words, progressively stronger emotional content could yield progressively smaller frequency effects on the N400 as infrequent emotional words are more inordinately facilitated. This would indicate that the avalanche effect is graded corresponding to the magnitude of the emotional significance, and relatively pervasive across the language. To test this novel hypothesis, we used the actual continuous values of the individual stimulus characteristics. In the regression model, the interaction would appear as a significant negative coefficient between frequency and emotionality, where some positive frequency effect is attenuated as emotionality increases.

Second, it's not clear which particular dimension of emotion is responsible for driving the avalanche effect. A large portion of emotional variance is captured by just two dimensions (Osgood et al., 1967; Russell, 1980): valence (ranging from pleasant to neutral to unpleasant) and arousal (ranging from "high" or active to "low" or passive). Both could plausibly provide for a facilitation of infrequent emotional words, as described above. But *which* dimension drives this effect is an empirical question that has not to our knowledge been addressed, either in the behavioral literature or the ERP literature (which generally manipulate both simultaneously).

We attempted to isolate the effect of valence from the effect of arousal on the lexical frequency effect. Considering the role of arousal in guiding the relative attention directed towards different stimulus features (Mather & Sutherland, 2011), we anticipated that arousal specifically might be responsible for attenuating the frequency effect even after valence is held constant. As such, we hypothesized that increasing arousal would decrease the N400 frequency effect, even after having controlled for valence.

And third, the contributions of task to avalanche effect are largely unknown. Nearly all of the previous studies on the topic have used lexical decision tasks, which generally emphasize lexical recognition. We implemented tasks that emphasized deep semantic processing instead. Specifically, we asked whether an overt focus on a particular semantic feature (category membership) might lead to a different avalanche effect from an overt focus on a particular emotional feature (valence). Previous work has suggested that unexpected emotional words may only be inordinately facilitated when in emotional contexts (Delaney-Busch & Kuperberg, 2013), and ERP emotion effects

generally tend to be largest in valence-evaluation tasks than in semantic categorization tasks (Fischler & Bradley, 2006).

As such, we hypothesized that the avalanche effect would be larger in experimental contexts that encourage the deployment of attention towards emotional features. We investigated this using a task manipulation (valence task vs. semantic task). We overtly tested the effect of task on the main effect of frequency, the main effects of valence and arousal, and their pairwise interactions.

Regression-based approaches to ERP analysis, where single-trial amplitudes are modeled using predictors based on information about the items, subjects, and experimental context, are well suited to address a number of these questions. Most importantly, the hypothesis that emotionality linearly attenuates the N400 frequency effect can be tested overtly, using continuous measures of emotionality and continuous measures of log-transformed frequency statistics. But regression models also allow for the flexibility to assess whether valence or arousal is the primary driver of this interaction. In multiple regression, coefficients are interpreted as the effect of the predictor on the outcome holding all else in the model equal (James, Witten, Hastie, & Tibshirani, 2013), and as such, we can for instance test if there is a frequency by arousal interaction above and beyond any concurrent influence of valence. They also naturally allow for between-group comparisons (such as task effects) that may otherwise be biased by missing data or unbalanced designs. Further advantages of regression-based approaches are discussed extensively elsewhere (Baayen et al., 2008; Baayen et al., 2002; Barr et al., 2013; Smith & Kutas, 2015a, 2015b).

Methods

Data were drawn from a previously published pair of studies assessing emotion processing in single emotional words (Delaney-Busch, 2013). A set of 468 words was composed that systematically varied by Valence (pleasant, unpleasant, or neutral) and Arousal (high, low). However, care was taken to also select words that differed significantly by frequency and concreteness, in addition to part of speech (nouns, verbs, and adjectives were all represented in matched proportions), in order to draw inferences about valence and arousal processing against a large and diverse sample of the English language. As such, these data are ideally suited to also assess emotion-frequency interactions.

Stimulus construction. Valence, Arousal, and Concreteness were all operationalized using 7-point Likert scales in a series of ratings studies (where "1" was "Negative", "Low", or "Abstract" and "7" was "Positive", "High", or "Concrete", respectively). About 20-50 participants were recruited for each ratings study, using the same demographic criteria as the subsequent ERP studies. Frequency was defined as the log of the HAL frequency per million (Balota et al., 2007). Length was defined as the number of letters. Finally, orthographic neighborhood size and bigram frequency were obtained from the MCWord database (Medler & Binder, 2005). Stimulus properties and examples are shown in table 1.

	Unpleasant Low Arousal	Neutral Low Arousal	Pleasant Low Arousal	Unpleasant High Arousal	Neutral High Arousal	Pleasant High Arousal
Valence	2.26 (0.42)	4.20 (0.46)	5.40 (0.36)	1.97 (0.41)	4.12 (0.588)	5.66 (0.394)
Arousal	3.47 (0.43)	3.34 (0.38)	3.39 (0.45)	4.61 (0.56)	4.43 (0.491)	4.75 (0.490)
Frequency	7.81 (1.72)	7.98 (1.82)	8.23 (2.03)	7.91 (1.58)	7.88 (1.806)	7.86 (1.710)
Concreteness	3.82 (0.91)	4.01 (1.07)	3.92 (1.15)	3.79 (0.95)	3.87 (1.073)	3.69 (1.000)
Length	7.06 (2.18)	7.13 (1.47)	6.96 (1.98)	6.92 (1.61)	7.12 (1.546)	7.31 (1.514)
Orth	2.06 (2.78)	1.88 (3.45)	2.15 (3.15)	1.73 (3.21)	1.68 (2.844)	1.45 (2.458)
Orth_F	8.44 (19.1)	6.13 (17.3)	22.90 (50.8)	13.80 (43.4)	8.26 (21.1)	15.85 (56.9)
N2_C	110.31 (89.7)	132.94 (93.1)	111.20 (93.6)	130.75 (108.4)	124.16 (99.7)	132.07 (97.5)
N2_F	811.73 (617.2)	936.24 (678.3)	1035.03 (806.5)	1010.47 (765.7)	884.03 (600.9)	1001.64 (919.0)
Examples	Stingy	Pacify	Serenity	Atrocity	Splashed	Flourish
	Anxiety	Feminine	Loyal	Brutal	Mythical	Delicious
	Ignorance	Random	Peace	Hate	Radical	Success
	Gangster	Sculpture	Tulips	Tyrant	Spicy	Caressed
	Vomit	Apples	Sapphire	Bombs	Samurai	Fireworks
	Garbage	Coffee	Food	Murder	Alien	Champion

Table 1: Stimulus properties and exemplars. Valence, Arousal, and Concreteness were all pre-rated using 7-point likert scales (from "most negative" to "most positive", "least arousing" to "most arousing", and "abstract" to "concrete", respectively). Frequency was defined as the log of the HAL frequency per million (Balota et al., 2007). Length was defined as the number of letters. Number of orthographic neighbors and the number of wordforms that share the same constrained bigrams were drawn from the MCWord database (Medler & Binder, 2005), along with the mean log frequency of the orthographic neighbors and the bigrams. Values are listed as "mean (standard deviation)". Reprinted from Delaney-Busch et al (*in review*).

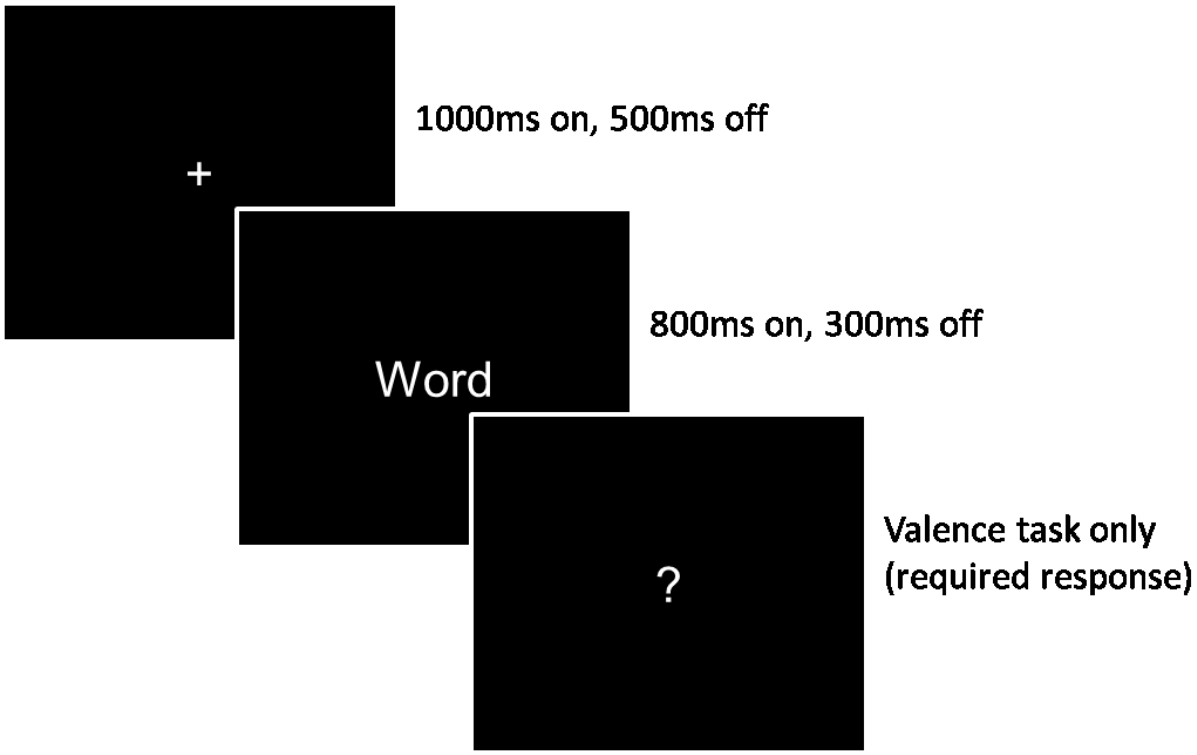

Figure 2: Stimulus presentation. Text was presented centrally on the monitor with white text on a black background. Between each sequence of fixation cross, word, and response demand, the experiment was shortly paused to allow participants time to blink.

Presentation. Stimulus presentation is shown in figure 2. Each trial started with a "blink sign", written as "(- -)", and began when the participant pressed the "advance" button with their right index finger. After pressing this button, a fixation cross appeared at the center of the screen for 1 second, and then disappeared, leaving a blank screen for 500ms. Then, a word was presented on the screen for 800ms. In the semantic categorization task, participants were instructed to press a button with their right thumb as quickly as possible if the word was identified as an animal word (of which 52 were added to the stimulus set, comprising 10% of the total words in the experiment). In the valence categorization task, participants were instructed to wait until a question mark

appeared on the screen 300ms after the offset of each word, and then should indicate whether that word was “positive”, “negative”, or “neutral” using one of three buttons on a button box.

The full stimulus set was divided into 25 self-paced blocks. Between each block, the experiment was paused while “READY” was shown on the screen. Participants were told that they could move their head or hands during this pause only, and that they could continue to the next word by pressing the “advance” button.

Participants. 26 young adults (13 men) participated in the semantic categorization task, and an additional 26 young adults (13 men) completed the valence categorization task. Six participants were rejected due to excessive artifact, leaving 24 in the semantic categorization group and 22 in the valence categorization group (for a total of 46 participants). All participants were right-handed native English speakers (having learned no other language before the age of 5) between the ages of 18 and 25. No participants were taking neuropsychiatric medications and none reported a history of psychiatric or neurological disorders or head trauma. All participants had normal or corrected to normal vision. They were compensated for their time and provided informed consent in accordance with the procedures of the Institutional Review Board of Tufts University.

Data collection and processing. Twenty-nine tin electrodes were held in place on the scalp by an elastic cap (Electro-Cap International, Inc., Eaton, OH). Electrodes were also placed below the left eye and at the outer canthus of the right eye to monitor vertical and horizontal eye movements, and on the left and right mastoids. Target impedance was below 5 kW for all scalp and mastoid electrode sites and below 10 kW for the two eye

channels. The EEG signal was collected with a left mastoid reference, and was amplified by an Isolated Bioelectric Amplifier System Model HandW-32/BA (SA Instrumentation Co., San Diego, CA) with a bandpass of 0.01 to 40 Hz and was continuously sampled at 200 Hz by an analogue-to-digital converter. The stimuli and behavioral responses were simultaneously monitored by a digitizing computer. Trials were rejected if they captured a blink, head movement, disconnected electrode, missing data, or other artifact between 200ms pre-onset and 800ms post-onset.

Statistical analysis. EEG data from each trial was time-locked to each stimulus onset, subject to a 0.1-40Hz bandpass filter, cal-normalized using the average of 100 ten-microvolt pulses (collected for each participant), and adjusted to a -100-0ms baseline. Then, for each trial, the average amplitude was calculated for each 25ms time window between 0 and 800ms (e.g. 0-25ms, 25-50ms, 50-75ms etc.) at each electrode on the scalp. These trial-by-trial amplitudes were then matched with information about the scale properties of the stimulus (e.g. continuous measures of frequency, valence, arousal, concreteness etc.), the properties of the subject (e.g. age, sex, task etc.), and the properties of the electrode (e.g. location on the scalp). ERP components of a priori interest were then composed by averaging over sets of the 25ms segments. These included the N400 (300-500ms), the P1 (75-125ms), the posterior N1 (150-200ms), the P2 (225-325ms), the anterior N2 (275-350ms), and the LPC (500-800ms). Finally, small sets of three electrodes each were averaged into regions: five central regions along the anterior-posterior axis, left and right anterior peripheral regions, and left and right posterior peripheral regions, as shown in figure 3.

For each set of amplitude measures (e.g. for N400 amplitudes), extreme outliers (beyond four sigma from the mean) were removed, leaving the middle 99.99% of the data (see chapter 1 for discussion). A visual inspection of density plots revealed a number of positively-skewed items-level features, including concreteness, number of phonemes, number of syllables, bigram count, and bigram frequency. These were all log-transformed, which necessitated adding "1" to all bigram counts and bigram frequency values prior to log-transforming due to the values of zero. HAL frequency was already log-transformed when it was obtained from the ELP database (Balota et al., 2007). Because of the a priori expectation that both pleasant and unpleasant valence values could yield amplitudes that differed from neutral valence items in the same direction, valence was always included as a second-order polynomial (i.e. every model that included valence also included squared valence). Because a sizeable portion of the words had an orthographic neighborhood count of 0 or 1, it was converted to a nominal variable with three levels of roughly the same size: "none", "one", and "many". All continuous predictors were then z-transformed, in order to provide for more interpretable model intercepts and directly comparable regression coefficients (James et al., 2013), as discussed in chapter 1. Word Class and Region were left as nominal variables.

Multilevel linear mixed-effects regression models were used to analyze these data, using the "lmertest" package 2.0-20 with lme4 (Bates et al., 2014) in R version 3.1.0 (R Core Team, 2014). Hypothesis tests were conducted using an "explanatory model" that controlled for a large number of theoretically-motivated nuisance factors (Shmueli, 2010), listed in table 2. Degrees of Freedom were estimated using the

Satterthwaite approximation in order to allow for the derivation of estimated p-values. To estimate generalizable numerical coefficients for significant effects, a second "prediction model" was composed, which was designed to protect against overfitting. For both models, the maximal justifiable random effects structure was included (Barr et al., 2013). This generally consisted of by-item and by-subjects random intercepts and by-subjects random slopes for frequency (models with additional random slopes usually failed to converge or showed evidence of converging at the boundary).

We expected a very specific pattern of results, where a particular main effect of frequency would be attenuated by arousal. Support for this directional hypothesis would appear as A) a significant positive main effect of frequency, where high-frequency words elicit a lower amplitude than high-frequency words (with a positive sign, where larger frequency values correspond with less negative / more positive amplitudes), and B) a significant interaction between frequency and arousal, where the frequency effect decreases with arousal (with a negative sign, where larger arousal values correspond with decreases in the frequency effect). However, as effects going in the other direction are possible, we utilized a two-tailed alpha for calculation of all significance levels.

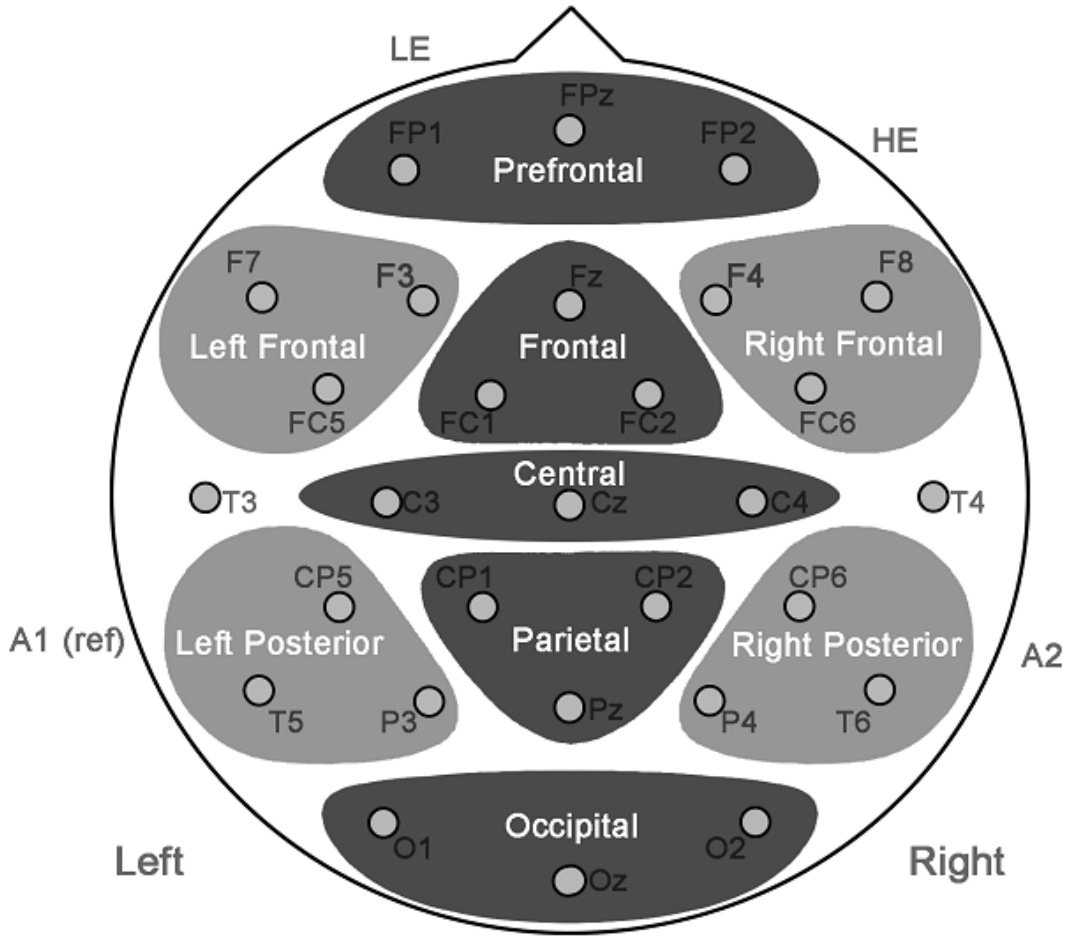

Figure 3: Regions used for analysis, viewed from the top of the scalp.

Results

A priori hypothesis tests were carried out within the N400 time window (300-500ms) along central, parietal, and occipital mid-regions of the scalp, where early N400 frequency effects are maximal (Laszlo & Federmeier, 2014). For the first test, Frequency, Arousal, and Experiment were entered as the three interacting effects of primary interest. Valence, Valence^2, Concreteness, Length, Orthographic Neighborhood size, Bigram count, Bigram frequency, and Word Class were included as nuisance factors. The

maximal justifiable random-effects structure consisted of by-subjects and by-items random intercepts, and a by-subjects random slopes for frequency (where the correlation between random slopes and random intercepts did not exceed an r of 0.2 in any region). Signs of multicollinearity were minimal, and no nuisance factor correlated with Frequency or Arousal greater than r = 0.3. Full results are reported in table 2.

Table 2a: Factors of Experimental Interest		
Fixed Effects	**Beta Estimate**	**P-value**
Arousal	0.055	0.41
Frequency	0.29	< 0.001*
Arousal by Freq	-0.17	0.006*
Experiment	0.54	0.36
Arsl by Exp	0.04	0.59
Freq by Exp	0.04	0.60
Arsl by Freq by Exp	0.04	0.53
Table 2b: Nuisance Factors		
Fixed Effects	**Beta Estimate**	**P-value**
Valence	0.06	0.33
Valence^2	0.07	0.34
Concreteness	-0.25	< 0.001*
Length	0.33	< 0.001*
Nouns	0.16	0.31
Verbs	0.39	0.015*
High Orth_N	-0.26	0.029*

Table 2: Point estimate and p-values for each fixed effect in the explanatory model.

To visualize the assumption of linearity, a loess local regression between Frequency and the N400 amplitude was fit for both high and low arousal words (using the arousal categories from the original study). As shown in figure 4, the linear model appeared to be a good approximation of the frequency effect across different levels of arousal.

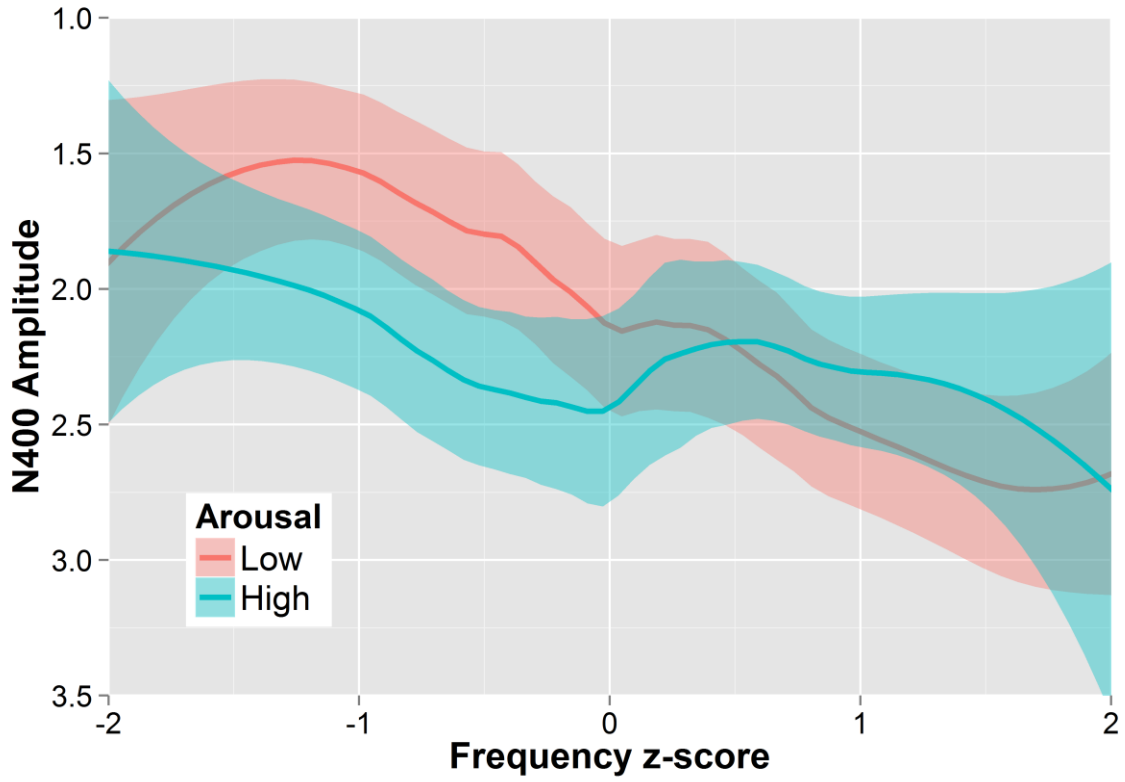

Figure 4: a loess local regression of the N400 component amplitudes over frequency values. A roughly linear frequency effect is discernable for low-arousal words, with higher frequency words eliciting a smaller N400 amplitude, but this effect appeared to be significantly attenuated for high-arousal words. Here, frequency is z-transformed and arousal is binned into two categories.

N400 hypothesis tests. A robust main effect of Frequency was observed over all three regions (central: $\beta = 0.508$, $p < 0.001$; parietal: $\beta = 0.542$, $P < 0.001$; occipital: $\beta = 0.319$, $p < 0.001$), with more frequent words eliciting a smaller N400 (i.e. more positive amplitudes). As anticipated, the main effect of Arousal was not significant in any time window. Nuisance factors that reached significance in this time window included Concreteness, Length, Part of Speech, Valence[2], and Orthographic Neighborhood size. No effects or interactions of Experiment were significant. Critically, an interaction

between Frequency and Arousal reached significance over the occipital region (β = -0.168, p = 0.006) in the hypothesized direction, where the positive N400 frequency effect appeared to attenuate with increased arousal. This remained significant after a Bonferroni correction for multiple comparisons across regions. This interaction was not significant in parietal or central regions. For visualization, the waveforms for the highest 1/3 and lowest 1/3 frequency words for both high and low arousal are shown in figure 5.

For the second hypothesis test, investigating the marginal effect of Valence on the frequency effect, Frequency, Experiment, and a second-order polynomial Valence factor were entered as the interacting effects of interest, with Arousal added as an additional nuisance factor to the ones listed above. No interactions involving a valence term were significant in any region.

N400 prediction model. To better specify the frequency-arousal interaction, a prediction model was fit for the N400 time window over occipital electrodes. Only significant theoretically-motivated factors were entered. This included Frequency, Arousal, the Frequency by Arousal interaction, Concreteness, Length, and Orthographic Neighborhood size, in addition to the same random-effects structure as used above. The standardized Frequency effect was estimated at β = 0.270, with a 95% Wald CI of [0.136, 0.404]. The standardized coefficient for the Frequency by Arousal interaction was estimated at β = -0.165, with a 95% Wald CI of [-0.284, -0.047].

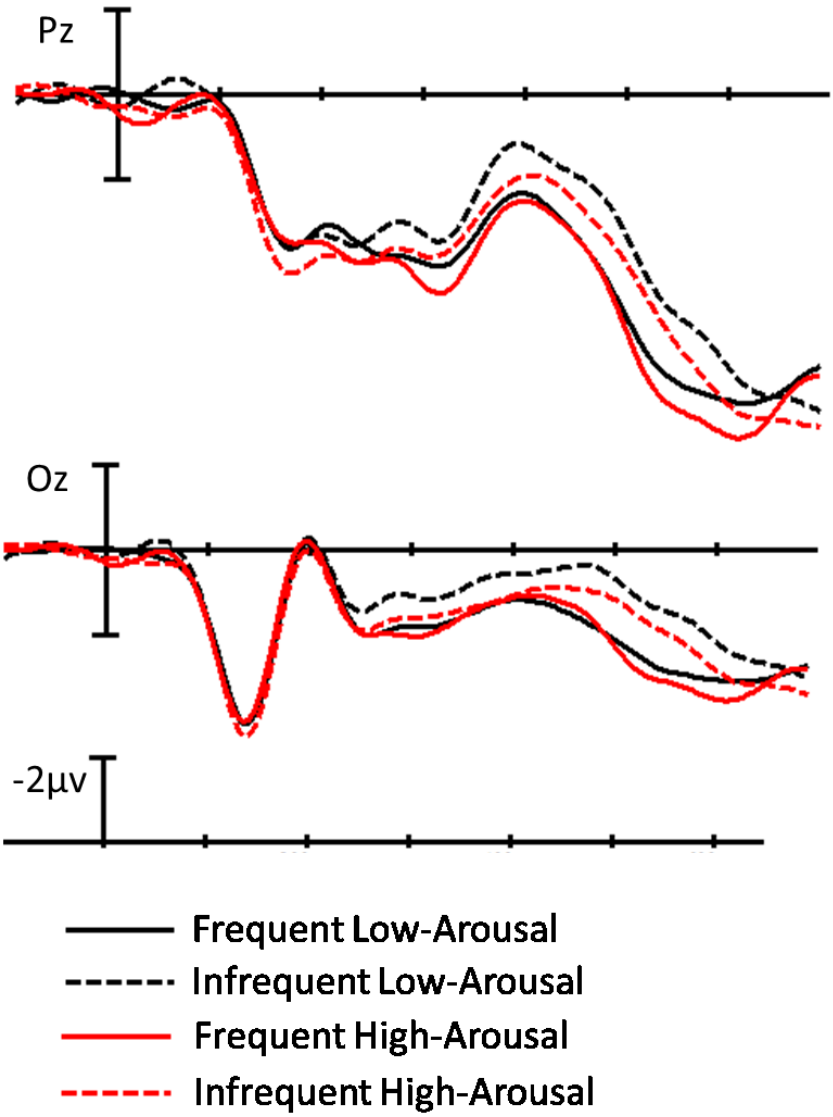

Figure 5: Grand mean waveforms for median-split arousal categories (high, low) and tertiary-split frequency categories (most frequent 1/3, most infrequent 1/3). The attenuation of the frequency effect by arousal is apparent in the early, but not the late, portions of the N400 time window.

To visualize the scalp distribution of the Frequency by Arousal interaction, this prediction model was refit at each individual electrode and the coefficients entered into a

voltage map extrapolation. As shown on Figure 6, the interaction was maximal over posterior electrodes.

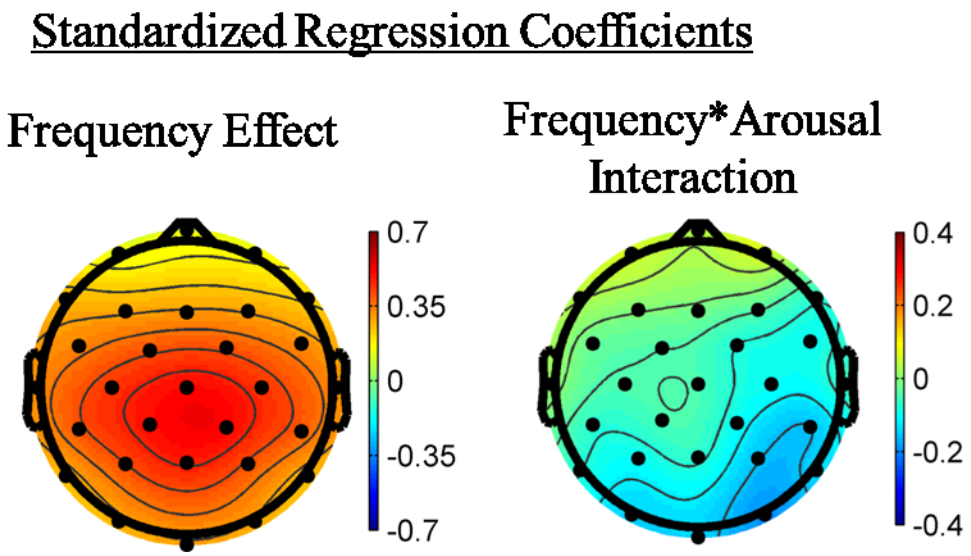

Figure 6: an interpolated distribution of the standardized regression coefficients for frequency and the interaction between frequency and arousal. This was obtained by fitting the prediction model for the N400 time window to each individual electrode. This shows relative effect size and direction, not significance (and thus multiple comparisons concerns are not relevant). A robust frequency effect was observed in the expected direction, and was maximal over centro-parietal electrodes. The influence of arousal on the frequency on effect was maximal bilaterally along the posterior periphery, with a distinct right-lateralized emphasis.

To visualize the time course of the Frequency by Arousal interaction, this prediction model at the occipital region was refit for every 25ms time segment from 0ms to 800ms. As shown in Figure 7, the main effect of Frequency seemed to have two distinct peaks: one relatively early in the N400 time window and one relatively late. As shown in figure 8, the interaction appeared to peak during this early portion of the N400.

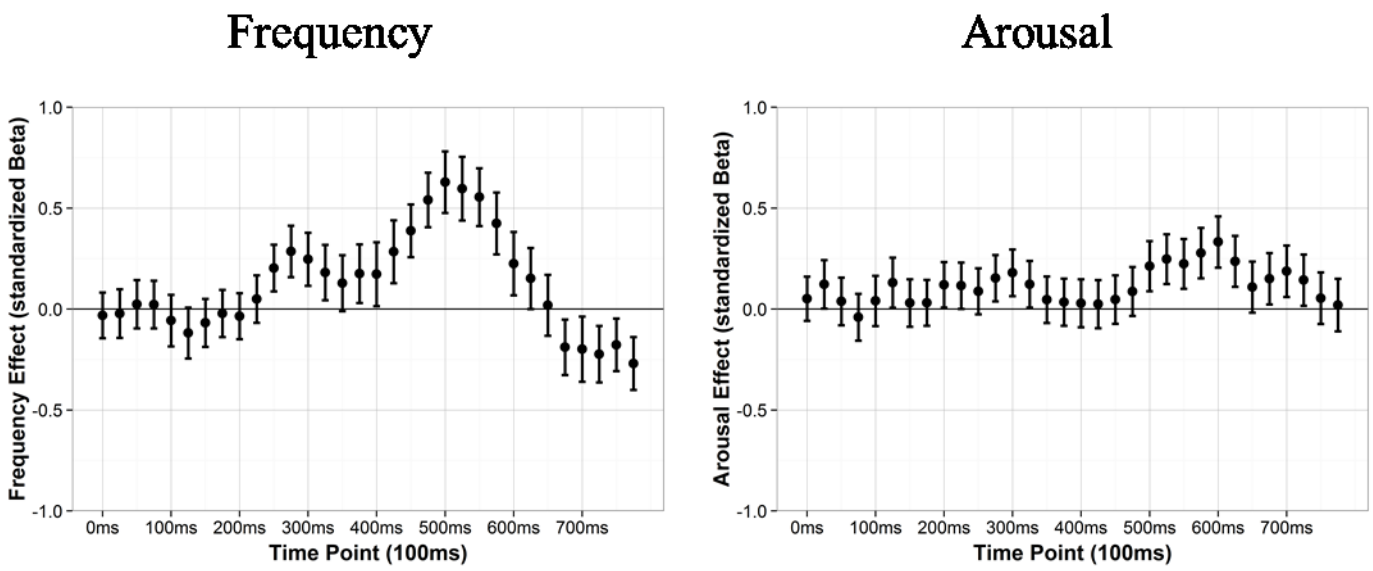

Figure 7: Point estimates and 95% confidence intervals for the standardized regression coefficients for Frequency and Arousal over time. These were determined using the prediction model fit to 25ms increments over the occipital region. The frequency effect showed two distinct peaks pertaining to the N400. The arousal effect showed, predominantly, a robust LPC effect (discussed further in chapter 3).

Exploratory models. To explore the possibility that interaction in the a priori 300-500 time window was driven primarily by this earlier peak, as suggested by Figure 7, we conducted one post-hoc hypothesis test over the occipital region for the earlier portion of the N400 time window, from 275-350. An early Frequency effect was significant in this time window ($\beta = 0.286$, $p < 0.001$), as were small effects of Length, Concreteness, and Word Class. The interaction between Frequency and Arousal was highly significant in this early time window ($\beta = -0.192$, $p = 0.003$). As such, we infer that this interaction peak did not appear to be driven by a nuisance factor that had been omitted from the prediction model.

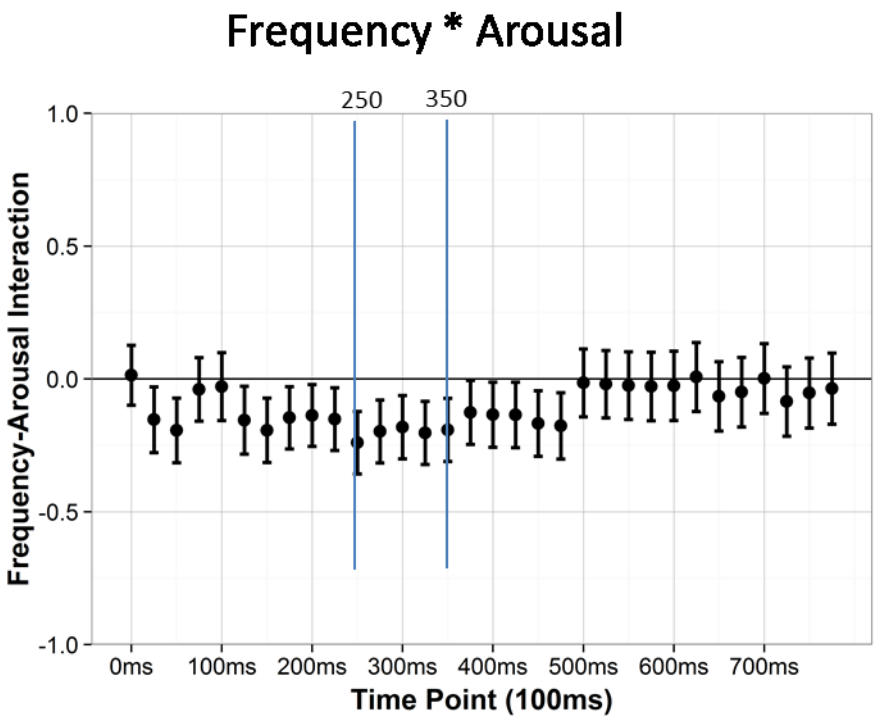

Figure 8: Point estimates and 95% confidence intervals for the standardized regression coefficients for the interaction between Frequency and Arousal over time. These were determined using the prediction model fit to 25ms increments. The maximal interaction occurred during the early peak in the frequency effect, but not the later peak of the frequency effect.

The time-course analysis also seemed to suggest that an interaction between Frequency and Arousal may have also occurred during earlier time windows. To protect against the possibility that the effects visible in the plot of 25ms increments were due to unprotected multiple comparisons, we subjected all 32 models to a Bonferroni correction. Even using this conservative control for multiple comparisons, four of the five 25ms time points between 250ms and 375ms remained significant for the interaction between Frequency and Arousal, and were the only time points to do so.

Discussion

The present investigation was intended to determine whether the N400 word frequency effect might be attenuated by emotion. We found evidence that this was the case. Over posterior electrodes, a robust Frequency effect was significantly reduced for the higher (versus lower) arousal words. As shown in figures 4 and 5, this was primarily driven by a facilitation for infrequent high-arousal words: a one-standard deviation increase in arousal corresponded with roughly a two-thirds reduction in the N400 frequency effect. This interaction was found to hold even after controlling for a large number of potential confounds, including length, concreteness, valence, bigram count, bigram frequency, orthographic neighborhood, word class, and the random variance attributable to the particular words and participants sampled for the study. In contrast, no influence of valence was found on the frequency effect. Further, neither the main effects nor interactions were influenced by task. Overall, these data show that arousal can drive the facilitation of infrequent words during semantic processing.

Notably, the avalanche effect seemed confined to the earliest portions of the N400 time window. Over posterior electrodes, the frequency effect appeared to have one distinct peak around 275-350ms, and another from 425-625ms (figure 7). The earlier frequency effect over posterior electrodes is an almost exact replication of the frequency effect observed by Laszlo and Federmeier (2014), possibly indicating a mapping of form to meaning. The avalanche effect was closely aligned with this early frequency effect (figure 8). In contrast, the interaction between frequency and arousal seemed to diminish rapidly with time, and was not present for most of broader frequency effect occurring

after 350ms, despite the fact that the main effect of frequency was larger.

Implications for the processing of emotional words. These data provide further evidence that emotion can attenuate the influence of word frequency during initial semantic processing. Generally, frequency acts as something of a default prior: for words presented in isolation, participants expect to see frequent words more than in frequent words (i.e. the prior probability of frequent words is higher). But like previous studies, we found that the reliance on this default prior appeared to be diminished for emotional words (Kahan & Hely, 2008; Kuchinke et al., 2007; Mendez-Bertolo et al., 2011). One might conceptualize emotional salience as a broad, pervasive feature of stimuli (Osgood et al., 1967) that was prioritized during semantic processing and dominated over effects of frequency within the N400 time window. Even uncommon words like "avalanche" and "contagion" carry basic emotional implications like "danger". Because N400 is an index of semantic processing, our data suggest that the prioritization of these easily processed emotional features may override the default prior expectancy during the processing of emotional words, as if participants were remaining vigilant for them.

But we extend these previous findings in a number of important ways.

First, we found a continuous and progressive influence of emotion on the N400 frequency effect. The larger the arousal, the smaller the frequency effect over posterior electrodes. The avalanche effect appears, in the very least, to be relatively pervasive across this large sample of English words, making it less likely that a modest subset of aberrant words could be driving the overall effect. It indicates that the prioritization of emotional features during semantic processing may be a difference in scale rather than a

difference in kind. The semantic processing of mildly salient words is still influenced primarily by frequency (i.e. prior expectations), while intensely salient words are processed efficiently regardless of the frequency. However, we caution that the presence of a linear interaction in the present data does not necessarily imply a linear biologic mechanism. Though the linear interaction accounted for a significant amount of the variance in the data, the actual neural generator could be something completely different (Box & Draper, 1987). We simply suggest that a continuous and progressive avalanche effect is a *useful* account of the data, at present.

Second, we implicate arousal specifically as the primary contributor to the avalanche effect. While previous studies have found evidence for an avalanche effect to more emotional (versus less emotional) words using both behavioral measures (Kahan & Hely, 2008; Kuchinke et al., 2007) and ERP measures (Mendez-Bertolo et al., 2011), the present study independently manipulated valence and arousal in order investigate their differential impact. We found that high-arousal words attenuated the N400 frequency effect more than low-arousal words, regardless of whether those words were pleasant, unpleasant, or neutral. Considering that the arousal dimension has also been strongly implicated as an influence for perception, attention, and memory (Mather & Sutherland, 2011), and has been found to elicit some of the earliest ERP effects elicited by emotional words (Bayer, Sommer, & Schacht, 2012; Hofmann, Kuchinke, Tamm, Võ, & Jacobs, 2009), we find it plausible that the features of high-arousal words may be prioritized during semantic processing the most strongly.

And finally, we extend the evidence for the avalanche effect to two novel tasks.

The previous studies to investigate the avalanche effect have primarily used lexical decision tasks, where efficient task performance depends on recognition of lexical units, and the effects of lexical frequency tend to be quite large (Balota & Chumbley, 1984, 1990), since participants can rely on their default expectations (i.e. the probability given prior experience) of encountering any particular word (Fischer-Baum et al., 2014). We asked whether the avalanche effect might also occur when the task emphasizes deep semantic processing instead, such as when categorizing category membership or valence. We found evidence that this was the case.

But contrary to our hypothesis, we did not find that the *particular* task influenced either the main effect of frequency or the interaction between frequency and arousal. Because affective information tends to be prioritized most during tasks where affect is most overtly relevant (Lai et al., 2012; Okon-Singer, Lichtenstein-Vidne, & Cohen, 2013), we anticipated that the valence task could have prompted a larger avalanche effect than the semantic task. This was not the case. All arousing words appeared to be equally facilitated *regardless* of task, just as they were equally facilitated regardless of frequency. Combined with prior research (Mendez-Bertolo et al., 2011), comparable facilitation for infrequent emotional words during semantic processing (as indicated by ERP amplitude) has now been found for tasks that emphasized lexical, semantic, and emotional features, respectively. We suggest that the avalanche effect may be relatively robust to task demands, particularly when compared to other ERP emotion effects (Fischler & Bradley, 2006; Hajcak, MacNamara, Foti, Ferri, & Keil, 2013). This could be consistent with the adaptive significance of the avalanche effect, where facilitation of even infrequent

emotional words could occur regardless of the disposition and attentional deployment of the comprehender at the moment of exposure. However, additional experiments are required across a significantly wider range of tasks before such a conclusion could be safely generalized.

Affective primacy and the N400. The most pressing question raised by these data is *why* infrequent emotional words might be facilitated. As noted above, the ability to rapidly comprehend emotional signals even when they are unexpected or rare may hold adaptive significance. There is other recent evidence of emotional words being facilitated in cases where one would generally expect to see a large N400 amplitude, such as for unexpected and incongruent words in sentences (Delaney-Busch & Kuperberg, 2013) and for contrary moral statements (Leuthold, Kunkel, Mackenzie, & Filik, 2015) which otherwise would be expected to elicit an N400 effect (Van Berkum et al., 2009). This has been attributed to the "affective primacy hypothesis", where the processing of emotional stimuli (including words) is generally facilitated (Storbeck & Clore, 2007) particularly in emotional contexts (Lai et al., 2012).

We suggest that the facilitation for infrequent emotional words may be similarly attributable to affective primacy. Clearly, the default prior expectations (i.e. word frequency) influenced semantic processing more for less arousing words than for more arousing words, as indexed by the N400. Salient words showed a bias: participants were treating infrequent emotional words as more frequent than they actually are. We suggest that this may be because the emotional *features* of even infrequent emotional words are still easily accessible, or perhaps even implicitly monitored for. Emotional features are

pervasive (Osgood et al., 1967), explaining a sizeable portion of the variance in semantic space (Samsonovic & Ascoli, 2010), where affective significance is routinely evaluated during the processing of most stimuli (Hajcak et al., 2010; Hajcak, Weinberg, MacNamara, & Foti, 2012). For the most salient words (i.e. the high-arousal words) these affective features may have been prioritized during the N400 time window, leading to a relatively superficial initial semantic analysis, but a rapid understanding of emotional significance. We think it is telling that arousal *only* influenced the very earliest discernible frequency effect, from 275-350ms, but did *not have any downstream effects* on the much larger Frequency effect that immediately followed (as shown in figures 7 and 8), or even any influence on the frequency effect clearly discernible over the classic spatio-temporal ROI for the N400 components (300-500ms over centro-parietal electrodes). The impact of arousal was isolated to the earliest stages of semantic processing (as also identified by Laszlo & Federmeier, 2014). We suggest that this might indicate that the emotional features of the most salient words were prioritized, and only *after* these affective implications have been activated did this relatively superficial semantic analysis give way to the full semantic analysis.

Such a prioritization is certainly evident in the deployment of attentional focus (Mather & Sutherland, 2011). When a person looks at a complex scene, they are generally going to focus at a gun pointed at them prior to focusing on other elements of the picture. This doesn't preclude them from comprehending the rest of the scene if given sufficient time. But often, the most salient features are evaluated first. We simply suggest that word processing may unfold in a similar manner, where emotionally salient features

of the words are prioritized in an analogous manner to the emotionally salient elements of a scene.

Importantly, we don't intend to argue that the processing of emotional features is *always* prioritized. On the contrary, the influence of emotion during word processing can be strongly task dependent (Delaney-Busch et al., *under review;* Fischler & Bradley, 2006; Holt, Lynn, & Kuperberg, 2009), and whether or not emotional words show behavioral facilitation is often subject to both task and context (Lai et al., 2012). This appears to be because emotional significance is not a static property of the stimulus, but an emergent feature of the stimulus, the context, and the perceiver (Okon-Singer et al., 2013). For example, a spider on your floor might generally dominate attention under most circumstances for most people, but a spider on your floor may even go completely unattended if your house is on fire. And some people actually quite like spiders and keep them as pets. It's not universally true that the spider will always dominate attention. In the present study, the high-arousal words seemed to dominate the early stages of semantic processing *on average.* But we think it is likely that the actual emotional significance for each word depends on the comprehender and the context. Further studies are needed to understand when emotional information seems to be prioritized during language processing and for whom.

Conclusions. In sum, we affirm and extend the existing evidence that the N400 frequency effect may be attenuated for emotional words. Our data suggest that arousal, rather than valence, may be the primary contributor to this effect, influencing an early peak in the N400 frequency effect over posterior electrodes. Further, both the frequency

effect and the interaction with arousal appeared to be approximately linear (as further supported by loess local regressions where linearity was not assumed by the model). Contrary to our expectations, we found that this effect did not differ as a function of task. Instead, our data indicate that this facilitation may be a general tendency for emotional words across the English language, possibly regardless of some changes to task demands. We tentatively suggest that a facilitation of the processing of infrequent emotional words may offer an adaptive significance to the comprehender. Participants appeared to prioritize affective information during initial semantic processing for the most strongly salient words, and only continued with a full semantic processing after.

CHAPTER III

WHEN IT ECHOES: SEMANTIC PRIMING, AFFECTIVE PRIMING, AND TASK EFFECTS DURING WORD PROCESSING

Emotion is central to deriving coherent meaning from a wide universe of possible stimuli and events (Osgood et al., 1967). Even for language, the appropriate comprehension of even simple utterances like "I love you" and "he passed away" requires consideration of the profound social and emotional implications just as much as the knowledge about the properties of love and death. Emotionally significant speech can resonate with our memories and our hopes (Bradley, 1994), our fears and our joys (Bradley & Lang, 2007), and even our most deeply-held values (Van Berkum et al., 2009). However, many questions remain about the exact nature of these affective influences on language comprehension, and the extent to which they are similar to or different from other forms of semantic knowledge. Here, we employ a simple priming paradigm in two experiments to better specify the nature of emotional significance during word processing.

Priming paradigms have offered critical insights into how word meanings are stored and utilized during language comprehension. In the semantic priming effect, the processing of a stimulus, called the target, can be facilitated if preceded by a semantically related or predictive (vs. unrelated or unpredictive) stimulus, called the prime (Meyer & Schvaneveldt, 1971; Neely, 1976). Behaviorally, semantic priming can quicken reaction times in lexical decisions or word categorizations, indicating that retrieval or

manipulation of the target word may be more efficient when primed (Neely, 1991). But similarly, there is also evidence that shared emotional content can facilitate processing, a so-called "affective priming" effect (see Wittenbrink, 2007 for review). For example, "delightful" is responded to more quickly if preceded by "party" then if preceded by "death" (Fazio, Sanbonmatsu, Powell, & Kardes, 1986; Hermans, De Houwer, & Eelen, 1994).

The extent that affective priming and semantic priming rely on shared or distinct mechanisms is presently unclear (Klauer & Musch, 2001, 2003). In semantic priming, the facilitation is thought to occur because semantic features of the target have already been activated prior to the actual onset of the target. Though strongly elicited by primes that are associated with (i.e. lead to a lexical prediction for) the target (Neely, 1976), the forward association is not strictly necessary (Lucas, 2000; Neely, 1991): it is sufficient simply that the semantic features of the target are pre-activated. As such, the magnitude of the semantic priming effect increases not only when the semantic content of the prime is more strongly related to the target, but also when the semantic content of the prime is more *likely* to be related to the target within the experiment, i.e. when the statistical contingency between the semantic content of the prime and the semantic content of the target is higher (den Heyer, Briand, & Dannenbring, 1983; Groot, 1984; Neely, Keefe, & Ross, 1989; Seidenberg, Waters, Sanders, & Langer, 1984). Overall, semantic priming seems to occur when the semantic content of the prime reliably allows for the expectation of the semantic content of the target (Becker, 1980; Neely, 1991).

However, the mechanisms behind affective priming are less well understood.

Classically, two mechanisms have been posed. First, some have suggested a highly automatic pre-activation of evaluative features when primes are the same valence as the following targets (Hermans, Spruyt, & Eelen, 2003; Rotteveel, de Groot, Geutskens, & Phaf, 2001; Spruyt, Hermans, De Houwer, Vandromme, & Eelen, 2007). This could rely on similar mechanisms as semantic priming (Fazio, 2001), where evaluative features are encoded in semantic memory, but could also rely on a different mechanism, where the ease of *evaluating* the emotional significance of the target is facilitated (rather than the ease of processing the emotional features stored in semantic memory). And second, others have suggested that affective priming may be predominantly a product of response demands, where the decisions necessary to respond to a target are facilitated when the prime would have led to the same decision (or response) as the target (Brouillet & Syssau, 2005; Klauer & Musch, 2003; Klauer & Stern, 1992). Early on, this was developed in part to parsimoniously explain why valence-incongruous trials can sometimes facilitate behavioral responses (vs. valence-congruous trials) when the task involves an incongruous or negative response, such as indicating when something is "false" (Klauer & Stern, 1992). However, both of these views are unable to account for the full pattern of affective priming effects in a recent Meta-analysis (Herring et al., 2013). Other, more nuanced accounts have since been proposed (De Houwer, 2014; Spruyt, De Houwer, & Hermans, 2009), but it remains clear that significantly more evidence is required to elucidate exactly how emotional primes manage to facilitate affectively congruent targets.

Priming and event-related potentials. Though these studies have implicated a

number of possible mechanisms of affective priming, behavioral measures necessarily reflect the combined influence of all processes prior to the response, making mechanistic distinctions more difficult. This limitation is particularly significant in cases where an effect could plausibly occur during the production of a behavioral response itself (or the decision-making process required for the response), as suggested above (Brouillet & Syssau, 2005; Klauer & Stern, 1992). Event-related potentials (ERPs), however, are ideally suited for comparing component processes. ERPs reflect the full time-course of the processes engaged by a stimulus. As such, ERP investigations could help clarify *when* during word processing the affective priming effect occurs, as well as indicating whether the mechanism of action is similar or different to the mechanism behind semantic priming.

There are two ERP components in particular that may be particularly important in understanding the affective priming effect. One ERP component, the N400, appears to reflect the semantic processing: when semantic processing is facilitated by the context, N400 amplitude decreases (Kutas & Federmeier, 2011). As such, semantic priming studies have found robust N400 modulation, with associated or related targets eliciting a significantly smaller N400 amplitude than unassociated or unrelated targets (Bentin, McCarthy, & Wood, 1985). In contrast, a second ERP component, the late positive complex (LPC), tends to reflect evaluative processing: more emotionally significant stimuli tend to elicit a larger LPC amplitude than less emotionally significant stimuli (Hajcak et al., 2010; Hajcak et al., 2012). This has been observed for pictures (Hajcak et al., 2010), faces (Schupp et al., 2004), and words (Citron, 2012; Kissler, Assadollahi, &

Herbert, 2006), among others types of stimuli, indicating that emotional stimuli likely engage general evaluative processes that are somewhat modality independent or downstream from modality-specific processing. The LPC often peaks near and extends beyond the point at which participants in behavioral studies would generally be indicating their response, so it is important to note that the LPC does not reflect the earliest point in processing where the valence of the prime could influence the processing of an emotional target. Rather, the LPC is a robust and reliable indicator of motivational significance and other evaluative processes, and is readily measured using ERPs. It is one of the most likely indicators of changes in evaluative processing. Though the types of processes described in theories of affective priming don't perfectly correspond to these known processes, both of these components could plausibly be expected to be sensitive to affective priming.

To date, a number of studies have identified affective priming effects on the N400 (J. P. Morris, Squires, Taber, & Lodge, 2003; Zhang, Kong, & Jiang, 2012; Zhang, Lawson, Guo, & Jiang, 2006; Zhang, Li, Gold, & Jiang, 2010), with valence-incongruous words eliciting a larger negativity than valence-congruent words. First, Morris and colleagues (2003) presented emotional adjectives that were preceded by noun primes (e.g. Terrorist…cruel). When the emotional adjective was preceded by a noun with a consistent connotation (vs. inconsistent connotation), N400 amplitude was reduced. Zhang and colleagues (2006, 2010) then presented emotional words that were preceded by words and pictures that either had the same valence or the opposite valence. They found negativities peaking at 450ms and 600ms, both of which were interpreted as N400

effects, with smaller amplitudes elicited by words that shared (vs. did not share) the valence of their prime. Finally, Zhang and colleagues (2012) reported N400 affective priming effects predominantly for high-arousal pairs, followed by an additional LPC effect. Together, these studies provide some preliminary evidence that semantic processing (as indexed by the N400 component) may be sensitive to an affective priming manipulation.

In contrast, two studies have provided evidence against N400 modulation by affective priming (Herring, Taylor, White, & Crites Jr, 2011; Hinojosa, Carretié, Mendez-Bertolo, Miguez, & Pozo, 2009), instead finding effects on the LPC. Hinojosa and colleagues (2009) modulated arousal of the primes and targets, rather than valence, and found no evidence for N400 modulation. Instead, they found an LPC effect, where high-arousal words were facilitated by other high-arousal primes (compared to low-arousal primes). However, it is not necessarily the case that the mechanisms behind arousal priming will be the same as the mechanisms behind valence priming, as valence and arousal reflect different aspects of emotion (Abelson & Sermat, 1962; Bradley, 2000). A more recent study (Herring et al., 2011, experiment 3) randomly selected a prime and a target from 20 pleasant and 20 unpleasant words that were matched on arousal until 512 prime-target pairs were obtained. They also found modulation on the LPC, but not on the N400 component. This provides initial evidence that affective priming may act on evaluative processing underlying the LPC rather than the semantic processing underlying the N40 component.

These mixed results are difficult to interpret for several reasons. For instance, the

studies by Zhang and colleagues (2006, 2010) reported unusually long latencies of the negativities (the N400 usually peaks at around 400ms), making it unclear whether this mechanism is directly comparable to semantic priming. Further, they used a cross-modal paradigm that included both words and pictures, while the other studies used stimuli of a single modality. But most pressingly, though Herring and colleagues (2011) systematically included categorical relationships (animals like "shark" vs. person-related nouns like "thief"), only one of these studies controlled for or reported overt measures of semantic relatedness (Hinojosa et al., 2009 reported semantic similarity ratings). As such, it's largely unknown how semantic features in the primes and targets may have influenced the ERP affective priming effects currently reported in the literature.

The present investigation. The present study addresses this gap. Our goal was to understand how the mechanisms behind affective priming relate to the mechanisms behind semantic priming. We implemented a full cross of affective priming and semantic priming in a large, carefully controlled set of stimuli. The effect of affective priming was operationalized using primes that shared or did not share the valence of the target, as determined by ratings studies. The effect of semantic priming was operationalized using associated and unassociated words in an Association factor, as determined by word association norms. Critically, we utilized synonyms and antonyms as a way to systematically and orthogonally manipulate both affective priming and semantic priming conditions at the same time; synonyms and antonyms were both equally associated with the target, but synonyms were the same valence as the target and antonyms were the opposite valence as the target. As a control for any impact that synonyms and antonyms

may have on the target that may occur above and beyond semantic priming and affective priming effects, the same manipulation was repeated for a new set of *neutral* words: neutral targets with neutral primes that also varied by Relationship (synonym, antonym) and Association (associated, unassociated). As a whole, this design gave us a 2 Relationship (synonym, antonym) x 2 Association (associated, unassociated) factorial design for both emotional words and neutral words. Since the hypotheses about how semantic priming and affective priming relate can be distinguished, in part, by task, we conducted the experiment using both a semantic matching task and a valence matching task.

We considered a number of possibilities that could help explain the divergent ERP literature on affective priming. First, emotion could be nonrandomly related to particular semantic features, such that valence correlates with semantic content (Storbeck & Clore, 2007; Storbeck & Robinson, 2004). We call this the "valenced features" hypothesis. Emotional features are powerful indicators of semantic content (Osgood et al., 1967), as words tend to be evaluated as pleasant or unpleasant for a reason (Herring et al., 2013). As such, two randomly chosen unpleasant words may have *semantically* more in common than a random chosen pleasant and unpleasant word pair. In this account, Herring and colleagues (Herring et al., 2011) note that the affective priming manipulation in some previous studies may have simply been a roundabout way of achieving a semantic priming effect, which manifested on the N400. This explanation would predict that after semantic relationships have been carefully controlled, as in the present study, no affective priming effects should be observed on the N400.

Second, previous ERP studies could have found N400 affective priming effects because emotional features (such as valence) are actually stored as part of a word's meaning and activated during semantic processing. We call this the "semanticized valence hypothesis. For example, the fact that a snake is frightening may play a similar role during semantic processing to the fact that a snake is a reptile, where valence might be analogous to category membership or other semantic features. Consequently, the affective priming reported in previous ERP studies could simply be a subtype of semantic priming, with no clear distinction in the primary mechanism. In this view, the affective features of the prime pre-activate and facilitate the processing of targets with the same affective values (Bargh, 1997; Fazio et al., 1986). While previous reviews have attributed this to a spreading activation account (Bargh, 1997), we choose to remain agnostic to the precise mechanism, as recent evidence from other fields has indicated that a prediction mechanism may be at least as plausible as a spreading activation mechanism in explaining semantic (and thus, in this case, affective) priming, as discussed in chapter 4.

This "semanticized valence" hypothesis is in many ways similar to the "valenced features" hypothesis described above; there is not a large distinction in many cases between word valence that is itself stored as a semantic feature and word valence that simply correlates with stored semantic features. But there is one critical difference: once all other semantic relationships have been controlled, the semanticized valence hypothesis suggests that affective priming should still occur on the N400 component (because the prime and target still share a semantic feature - valence) while the valenced features hypothesis suggests that it should not (because the prime and the target no longer

share any semantic features). Unlike previous investigations that manipulated affective priming and category priming (Herring et al., 2011; Storbeck & Robinson, 2004), we utilized a significantly more expansive associative priming manipulation with significantly stronger overt controls on semantic content, where repetition was minimal. As such, an N400 affective priming effect in the present study (particularly one that emulates the semantic priming effect observed in the same participants) would support the semanticized valence hypothesis, but not the valenced features hypothesis.

And third, affective priming could instead be a consequence of facilitation during *evaluative* processing rather than semantic processing. We call this the "evaluative congruity" hypothesis. In this view, the mechanism behind affective priming is completely distinct from the mechanism behind semantic priming. Many affective priming studies have so far endorsed a form of this perspective, where (often post-lexical) evaluative processes are facilitated by affectively congruous (vs. incongruous) trials (Klauer & Musch, 2003). Central to this theory is the understanding that evaluating the valence of the target is necessary when *reporting* the valence of the target as part of the task demands (Herring et al., 2013). Participants simply also apply some of these response demands to the primes (De Houwer, 2014), and the task-relevant evaluation to the target is facilitated if congruous with the evaluation of the preceding prime. Because the evaluative congruity hypothesis posits facilitation during evaluative processing rather than semantic processing, we would expect the facilitation to manifest on the LPC rather than the N400 component. Further, because of the centrality of the task to this explanation, we would expect this LPC affective priming effect to be largest (or perhaps

only present at all) in tasks that emphasized evaluative processing. To address this, we conducted the experiment using two different tasks in two different groups of participants. As evidence of affective priming during tasks other than only evaluative decisions clearly indicates that the actual response planning is insufficient to completely explain affective priming effects (Herring et al., 2013), we instead required participants to overtly attend to the valence of both primes and targets, to maximize the chances of facilitation due to evaluative congruity.

We note that these three mechanisms are not mutually exclusive. Though single-word studies suggest that emotional words tend to elicit differences during evaluative processing rather than semantic processing (Citron, 2012; Kissler et al., 2006), emotional features like valence can still plausibly be reflected in semantic memory enough to induce a priming effect for pairs of words on the N400 as well. Similarly, a correlation between emotion and other semantic content (as in the valenced features hypothesis) can plausibly be substantial even if valence itself is also treated as a semantic feature during processing. Consequently, we assess how the evidence pertains to each of the three hypotheses described above separately, as more than one could be supported by the data.

In addition to providing information about how semantic priming and affective priming relate to one another during word processing, this design also allows for the assessment of how semantic priming is influenced by emotion more generally. Specifically, we collected N400 semantic priming effects for both neutral primes and targets and for emotional primes and targets (averaged over same-valence and opposite-valence primes). It has been suggested that the semantic priming effect may be influenced

by the *salience* of the prime features (Marschark, 1983; Neely, 1991). We asked whether the N400 semantic priming effect might be magnified in cases where the words were *inherently* salient, as for emotional words. We are not aware of any previous ERP studies that have compared semantic priming effects for neutral and emotional words, but the hypothesis is relatively straightforward: the semantic priming effect may be larger for emotional words than for matched neutral words.

Finally, we considered how these priming effects might be influenced by task. First, we instituted a semantically-oriented task, where semantic priming effects were expected to be maximal. Semantic priming effects are generally largest when semantic information is overtly relevant to task demands (Herring et al., 2013) or when predicting aspects of the target (from the given prime) implicitly provides more utility (Lau, Gramfort, Hamalainen, & Kuperberg, 2013; Lau, Holcomb, & Kuperberg, 2013). But the predictions for affective priming differ. The evaluative congruity hypothesis suggests that affective priming effects on the LPC should be minimal (or absent) when evaluation of affective features is not relevant for the task demands, while the semanticized valence hypothesis suggests that affective priming effects on the N400 should still remain robust, and the valenced features hypothesis predicts no affective priming effect on either component (as semantic features that may have correlated with valence have been rigidly controlled). Second, we instituted a valence-oriented task, where semantic priming effects were expected to diminish. Here, the evaluative congruity hypothesis suggests that affective priming effects on the LPC should be maximal, while the semanticized valence hypothesis still predicts robust N400 effects of affective priming and the valenced

features hypothesis still predicts null effects on both components. As such, these two experiments together falsify three plausible explanations of ERP affective priming effects. Data from experiment 1 previously appeared in a masters thesis (Delaney-Busch, 2013).

Experiment 1

Experiment 1 Methods

The present study aimed to isolate the effects of affective priming from the effects of semantic priming, while controlling for the type of relationship between the primes and targets. We therefore implemented a full cross of affective priming (operationalized as a "Relationship" between the prime and the target of either the "same valence" or "opposite valence") and semantic priming (operationalized as an "Association" between the prime and the target of either "associated" or "unassociated"). This was accomplished by using synonyms and antonyms of the target words, where synonyms were the same valence as the target and antonyms were the opposite valence as the target, but both synonyms and antonyms were associated with the targets to the same degree. In this way, we could assess the effect of affective priming for semantically associated primes and targets (e.g. "uptight - tense" vs. "relaxed - tense") and for scrambled (non-associated) primes and targets (e.g. "repulsive - tense" vs. "attractive - tense"), as well as assess the effect of semantic priming on same-valence targets (e.g. "uptight - tense" vs. "repulsive - tense") and opposite-valence targets (e.g. "relaxed - tense" vs. "attractive - tense") separately. A comparable factorial design was implemented for matched neutral prime-target pairs as well, to serve as a control for the nature of the synonym and antonym

primes, and to ensure that only half of the experimental materials were significantly valenced. A portrayal of the full experimental design (with examples) is shown in table 3.

Target Valence	**Association with the Prime**	**Relationship with the Prime**	**Example**
Unpleasant target	Associated	Synonym, Same valence	Uptight - Tense
		Antonym, Opposite valence	Relaxed - Tense
	Unassociated	Same valence	Repulsive - Tense
		Opposite valence	Attractive - Tense
Neutral target	Associated	Synonym, Neutral valence	Beneath - Under
		Antonym, Neutral valence	Over - Under
	Unassociated	Neutral valence	Combine - Under
		Neutral valence	Separate - Under

Table 3: Design and Example Stimuli. Both unpleasant and neutral targets were matched with primes in a 2x2 design that manipulated Association (defined using forward association strength: associated versus unassociated) and Relationship Type (defined as synonym versus antonym). Associated primes were the synonym or antonym of the target itself (e.g. uptight-tense) while unassociated primes where the synonym or antonym of another target used in the experiment (e.g. repulsive-tense, where ""repulsive" was the synonym of "ugly"). All neutral targets had neutral primes, while unpleasant targets had unpleasant synonyms and pleasant antonyms.

Stimulus construction. These stimuli were constructed as follows. First, a total of 40 synonym, antonym, and target triplets (such as "beneath/over-under" or

"uptight/relaxed-tense") were obtained both for neutral and for unpleasant targets from Florida Word Association Norms (Nelson, McEvoy, & Schreiber, 2004). For example, the target "tense" is associated with the cues "uptight" and "relaxed", which were then used in our experiment as the synonym and antonym primes for "tense". Association strength to the target was matched (using a 2-tailed t-test) for each type (synonym, antonym) of prime (neutral: $p = 0.94$; emotional: $p = 0.78$), and association strength of the neutral primes and targets matched the association strength of the emotional primes and targets for both the synonyms ($p = 0.31$) and antonyms ($p = 0.57$).

Lexical properties of the neutral and unpleasant *targets* were matched on word length, bigram count, bigram frequency, and orthographic neighborhood frequency ($ps > 0.2$). Orthographic and bigram properties were obtained from the MCWord database (Medler & Binder, 2005). Frequency was also obtained, using the English lexicon project (Balota et al., 2007), and a slight difference between neutral targets (mean ln frequency = 10.59) and unpleasant targets (mean ln frequency = 9.49) reached significance ($p = 0.002$). See table Xb for characteristics.

Lexical properties of the primes were also matched on word length, bigram count, bigram frequency, and orthographic neighborhood frequency ($ps > 0.1$), both within levels of emotion (synonyms vs. antonyms) and across levels of emotion (emotional primes vs. neutral primes). On average, the frequency of the synonyms (neutral: mean = 9.62; emotional: mean = 8.56) was lower than the frequency of the antonyms (neutral: mean = 10.61; emotional: mean = 10.06), but the size of this difference for neutral primes did not differ from the size of this difference for emotional primes ($p = 0.30$), allowing

the neutral synonyms versus antonym contrast to serve as a control for the affective priming contrast (unpleasant targets preceded by synonyms versus antonyms). See table 4a for characteristics.

In addition to these lexical features, the valence, arousal and concreteness of the primes and targets were assessed and controlled using rating studies (described in "ratings studies" below). All neutral targets had neutral primes, regardless of whether they were synonyms or antonyms (e.g. "beneath…under" and "over…under"). All unpleasant targets had unpleasant synonym primes (e.g. "uptight…tense"), but pleasant antonym primes (e.g. "relaxed…tense"). Primes within each emotion condition were matched on level of arousal; emotional synonyms were matched to emotional antonyms on arousal ($p = 0.52$), and neutral synonyms were matched to neutral antonyms on arousal ($p = 0.16$). Emotional primes and emotional targets were naturally higher in arousal than neutral primes ($p < 0.0001$) and neutral targets ($p < 0.0001$), respectively. As expected, unpleasant synonyms were also more negative than pleasant antonyms ($p < 0.0001$), while the valence of the neutral synonyms and antonyms was matched ($p = 0.55$). All prime types were matched on concreteness ($ps > 0.2$).

Finally, four unassociated conditions were made by scrambling the primes and targets of pairs with the same target valence. For example, the unpleasant target "tense" could be randomly paired with "repulsive" and "attractive", the synonym and antonym primes of a different unpleasant target. Unpleasant targets still had an unpleasant prime or a pleasant prime, but the prime was no longer semantically associated. Similarly, the neutral target "under" could be randomly paired with "combine" and "separate", the

synonym and antonym primes of a different neutral target. All of the randomized primes were confirmed to have no association with their randomized targets, as determined by the Florida Word Association Norms (an association strength of 0 for all random pairs) and a careful manual assessment of the randomly generated pairs.

Ratings Studies. A series of rating studies for valence, arousal, and concreteness was carried out in participants who did not take part in the ERP experiment. In all rating studies, participants were recruited through online postings and were compensated for their time. Informed consent was obtained for all participants, and responses were excluded if there was early language exposure other than English, psychiatric illness, neurological illness, neurological damage including stroke and concussion, or current treatment with psychoactive medication. Further, "catch" questions were used to identify and omit bots. Outliers were also omitted to exclude participants who did not understand the instructions or who appeared to be selecting options randomly: outliers were defined as being an average of two standard deviations or more away from the mean rating for each word. All participants completed a guided practice prior to each survey.

A scale of 1 to 7 was used for both arousal and valence, where 1 is "least arousing" or "most negative", and 7 is "most arousing" or "most positive", with an option to skip unfamiliar words. A scale of 1 to 7 was also used for Concreteness, where 1 is "most abstract" and 7 is "most concrete". Each word was rated by at least 35 participants. For use in constructing prime-target pairs, words were considered "unpleasant" if they had valence ratings between 1 and 3.5, "neutral" if they had valence ratings between 3.5 and 4.5, and "pleasant" if they had valence ratings between 4.5 and 7.

Table 4a: Properties of the Primes

	Word Length	Freq.	Neigh. Freq.	Bigram Freq.	Concreteness	Valence	Arousal	Association Strength*
Unpleasant Synonyms	5.98	8.56	6.96	3547	3.43	2.22	3.94	0.245
Pleasant Antonyms	5.35	10.06	7.62	3831	3.63	5.62	4.04	0.230
Neutral Synonyms	5.48	9.62	7.90	3350	3.77	3.99	3.31	0.197
Neutral Antonyms	5.33	10.61	7.26	3638	3.88	4.05	3.13	0.200

Table 4b: Properties of the Targets

	Word Length	Freq.	Neigh. Freq.	Bigram Freq.	Concreteness	Valence	Arousal
Unpleasant Targets	5.05	9.49	7.16	3630	3.67	2.12	3.97
Neutral Targets	5.3	10.59	7.66	3337	3.91	3.96	3.30

Table 4: Stimulus properties of primes and targets. “Freq.” lists the natural log of the HAL lexical frequency (per million). “Neigh. Freq.” lists the mean ln HAL frequency of the words in the orthographic neighborhood of the words used. “Bigram Freq.” lists the mean frequency (per million) of the bigrams contained in the words used. Concreteness was assessed using a 7-point ratings scale, where 1 was ““most abstract” and 7 was “most concrete”. Valence was assessed using a 7-point ratings scale, where 1 was ““most negative” and 7 was “most positive”. Arousal was assessed using a 7-point ratings scale, where 1 was ““Low” and 7 was “High”. “Association Strength” lists the mean forward association strengths between the associated primes and targets used.
*Unassociated primes all had an association strength of 0.

Counterbalancing. Six experimental lists were created. The first three lists were created by conducting three random pairings between primes and targets (to make the “unassociated” conditions). As a result, while associated primes and targets were the same across all lists, unassociated primes and targets could have up to three distinct

versions (between lists), minimizing the impact of incidental relationships between any two particular unassociated words. Each list contained each target exactly four times: once preceded by its own associated synonym, once preceded by its own associated antonym, once preceded by the random synonym of some other target, and once preceded by the random antonym of some other target. In this fashion, each prime appeared twice in each list: once with its associated target, and once with a random target. Overall, each participant saw exactly the same prime and target words exactly the same number of times (though the order of stimuli and pairings between particular primes and targets could differ for unassociated items).

Next, the order of these three lists was pseudorandomized into two halves, such that each target only appeared twice in each half of the experiment (once with an associated prime, and once with an unassociated prime; once with a synonym word, and once with an antonym word), and each prime only appeared once in each half. The number of conditions within each half was also held constant. For example, the first half of the experiment for each list contained 10 each of neutral and unpleasant targets preceded by associated synonyms, associated antonyms, unassociated synonyms, and unassociated antonyms, for a total of 80 prime-target pairs within each half of the experiment. Then, three additional lists were created by switching the order of the halves for each of the first three lists.

Participants. Data from 26 young adults (6 men) was collected. One participant was excluded due to computer malfunction during data collection, and another participant was excluded due to excessive artifact, leaving 24 participants for analysis. All

participants were right-handed, native English speakers (having learned no other language before the age of 5) between the ages of 18 and 25. No participants were taking psychiatrically active medications, and none had a history of psychiatric or neurological disorders. All participants had normal or corrected to normal vision, and no history of head trauma. They were compensated for their time, and provided informed consent in accordance with the procedures of the Institutional Review Board of Tufts University. All participants were recruited through online postings at Tuftslife.com (though not all participants were Tufts undergraduates).

Stimulus presentation. Stimulus presentation is summarized in figure 9. Participants sat in a quiet, dimly-lit room, and were instructed to sit still during the experiment with both feet on the floor and both hands on the controller. Each trial started paused at a "blink sign", written as "(- -)", and began when the participant pressed the "advance" button with their index finger. After hitting this button, a fixation cross appeared at the exact center of the screen for 1 second, then disappeared leaving a blank screen for 500ms. Participants were instructed to focus their gaze at the center of the fixation cross, and to continue looking at that spot after the fixation cross disappeared. Then, a prime was presented on the screen for 200ms, followed by 50ms of blank screen. The target then appeared on the screen for 800ms, leaving a total SOA of 250ms from the prime. Participants were instructed to read the words without scanning; the monitor was positioned such that the longest words in the experiment encompassed about 3.5 degrees of the visual field (most words were within 2-2.5 degrees), allowing participants to read the words without moving their eyes. The physical center of each word appeared at the

center of the screen.

Figure 9: Stimulus presentation.

Each target was followed by 300ms of blank screen, then a response prompt ("?"). Participants were given the task of deciding whether the words shown were "related" or "unrelated", using two buttons on a button box. For half of the participants, "related" was on the left, while for the other half, "related" was on the right, counterbalanced such that each of the six lists was seen an equivalent number of times for each response direction. Prior to the experiment, participants were given a 5-minute guided practice session and a

number of examples, in order to ensure that the task and presentation method were understood. None of the practice or example words appeared in the experiment.

EEG recording and processing. Twenty-nine tin electrodes were held in place on the scalp by an elastic cap (Electro-Cap International, Inc., Eaton, OH). Electrodes were also placed below the left eye and at the outer canthus of the right eye to monitor vertical and horizontal eye movements, and on the left and right mastoids. Impedance was kept below 5 kΩ for all scalp and mastoid electrode sites and below 10 kΩ for the two eye channels. The EEG signal was amplified by an Isolated Bioelectric Amplifier System Model HandW-32/BA (SA Instrumentation Co., San Diego, CA) with a bandpass of 0.05 to 30 Hz and was continuously sampled at 200 Hz by an analogue-to-digital converter. The stimuli and behavioral responses were simultaneously monitored by a digitizing computer. Data was collected using a left-mastoid reference, and re-referenced offline to the average of the left and right mastoids. Electrodes were averaged into regions as specified by figure 3.

Experiment 1 Results

Behavioral Results. Mean accuracy was defined as the proportion of times that an associated word pair was identified as "related", or an unassociated word pair as "unrelated". Overall, every participant had a total mean accuracy over 80% (mean 91.5%, SD 3.9%). Participants were significantly more accurate for Neutral words than for Unpleasant words (95% CI = [1.22%, 5.36%] more), for Associated words than for

Unassociated words (95% CI = [3.61%, 12.42%] more), and for Antonyms than for Synonyms (95% CI = [2.65%, 5.18%] more).

Neutral word ERP results. Neutral words were assessed using a 2 Association (associated, unassociated) x 2 Relationship (synonym, antonym) repeated measures ANOVA. One omnibus ANOVA was constructed for the mid-regions that additionally incorporated "Region" (with 5 different regions down the anterior-posterior axis) and another for peripheral regions that additionally incorporated "Region" (anterior, poster) and "Hemisphere" (left, right). All Names and distributions of scalp regions used for analysis are given in figure 3.

N1: 85-165ms. The only significant effect was an interaction between Association and Hemisphere over peripheral regions ($F(1,23) = 7.94$, $p = 0.01$), where associated targets elicited a larger N1 amplitude than unassociated targets over the right hemisphere.

P2: 250-310ms. A main effect of Association was significant for both the mid-regions ($F(1, 23) = 8.67$, $p < 0.01$) and peripheral regions ($F(1,23) = 7.15$, $p = 0.014$) omnibus ANOVA, with unassociated words eliciting a smaller positivity than associated words. Over peripheral regions, this Association effect additionally interacted with Hemisphere ($F(1, 23) = 5.32$, $p = 0.03$), and appeared slightly larger over the right hemisphere. Main effects and interactions involving Relationship did not approach significance in either the mid-regions ($ps > 0.5$) or peripheral regions ($ps > 0.1$) omnibus ANOVA.

N400: 310-510ms. A main effect of Association was highly significant for both

the mid-regions ($F(1, 23) = 81.6$, $p < 0.0001$) and peripheral regions ($F(1, 23) = 51.7$, $p < 0.005$) omnibus ANOVA, with unassociated words eliciting a larger negativity than associated words (see figure 10). Along the mid-regions, Association additionally interacted with Region ($F(4, 92) = 17.2$, $p < 0.0001$), where the Association effect was largest in the central mid-region ($F(1, 23) = 80.4$, $p < 0.0001$) and smallest in the prefrontal ($F(1, 23) = 28.9$, $p < 0.0001$) and occipital ($F(1, 23) = 31.9$, $p < 0.0001$) regions (see voltage maps in figure 10). Along the periphery, the Association effect additionally interacted with hemisphere ($F(1, 23) = 11.4$, $p < 0.005$), and appeared to be slightly larger over the right hemisphere. Main effects and interactions involving relationship type did not approach significance in either the mid-regions (ps > 0.4) or peripheral regions (ps > 0.3) omnibus ANOVA.

LPC: 510-660ms. A modest late positivity was observed following the N400, peaking approximately 570-600ms after stimulus onset. A main effect of Association was highly significant for both mid-regions ($F(1, 23) = 85.8$, $p < 0.0001$) and peripheral regions ($F(1, 23) = 52.0$, $p < 0.0001$) omnibus ANOVA, with associated words eliciting a larger positivity than unassociated words. As shown in figure 10, this is likely a continuation of the Association effect that began in the N400 time window. Along the mid-regions, this Association effect additionally interacted with Region ($F(4, 92) = 15.0$, $p < 0.0001$), where the effect was largest in the central mid-region ($F(1, 23) = 97.1$, $p < 0.0001$) and smallest in the prefrontal ($F(1, 23) = 12.5$, $p = 0.0017$) and occipital ($F(1, 23) = 28.9$, $p < 0.0001$) regions. Along the periphery, an Association x Region x Hemisphere interaction reached significance ($F(1, 23) = 6.45$, $p = 0.018$), with a larger Association

effect along the posterior periphery ($F(1, 23) = 50.4$, $p < 0.0001$) and an Association x Hemisphere x Electrode interaction along the frontal periphery ($F(2, 46) = 5.78$, $p = 0.011$). Main effects and interactions involving relationship type did not approach significance in either the mid-regions (ps > 0.2) or peripheral regions (ps > 0.1) omnibus ANOVA.

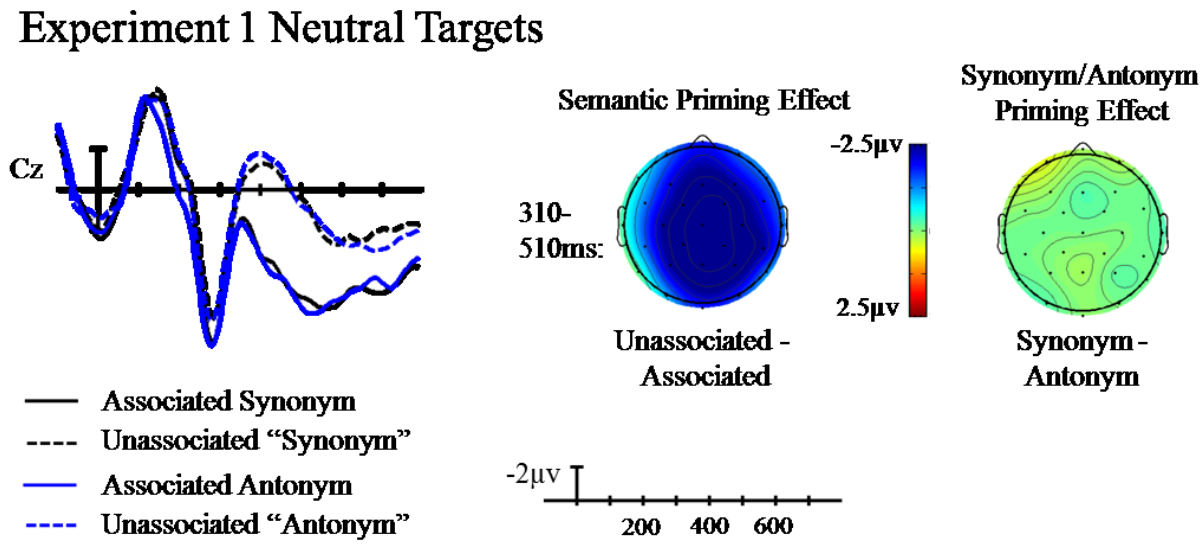

Figure 10: Effect of Association and Relationship for neutral targets in experiment 1. A robust N400 semantic priming effect was observed, maximally over centro-parietal electrodes, and did not differ by relationship type.

Emotional word ERP results. Emotional words were assessed using the same method as neutral words: a 2 Association (associated, unassociated) x 2 Relationship (same-valence synonym, opposite-valence antonym) ANOVA that additionally incorporated spatial factors for mid-regions and peripheral electrodes (see figure 3 for scalp regions).

N1: 85-165ms. A significant main effect of Association was significant over

peripheral regions ($F(1,23) = 9.86$, $p = 0.005$) but was not significant over the mid-regions ($F(1,23) = 4.06$, $p = 0.056$). A four-way interaction between Association, Relationship, Hemisphere, and Region also reached significance over peripheral regions ($F(1,23) = 9.06$, $p = 0.006$), with follow-up tests revealing a larger Relationship effect for associated words ($F(1,23) = 7.8$, $p = 0.01$) than for unassociated words ($ps > 0.6$) over parietal peripheral regions (with no difference over frontal peripheral regions).

P2: 250-310ms. A main effect of Association was significant for both the mid-regions ($F(1, 23) = 14.9$, $p < 0.001$) and peripheral regions ($F(1,23) = 10.95$, $p = 0.003$) omnibus ANOVA, with unassociated words eliciting a smaller positivity than associated words. Over peripheral regions, a Relationship by Region by Hemisphere interaction reached significance ($F(1, 23) = 4.95$, $p = 0.036$). Follow-up tests within each peripheral region revealed a Relationship x Hemisphere interaction that was significant over the posterior ($F(1, 23) = 6.00$, $p = 0.022$) but not the anterior ($p > 0.4$) portions of the scalp, with the Relationship effect contained to the left posterior. No other effects were significant.

N400: 310-510ms. A robust N400 was observed, peaking approximately 410 after stimulus onset. A main effect of Association was highly significant for both the mid-regions ($F(1, 23) = 69.0$, $p < 0.0001$) and peripheral regions ($F(1, 23) = 65.1$, $p < 0.005$) omnibus ANOVA, with unassociated words eliciting a larger negativity than associated words (see figure 11). Association additionally interacted with Region along both the mid-regions ($F(4, 92) = 23.6$, $p < 0.0001$) and periphery ($F(1, 23) = 12.96$, $p = 0.0015$)

where the Association effect was maximal in the central mid-region ($F(1, 23) = 69.45$, $p < 0.0001$) and the posterior periphery ($F(1, 23) = 67.1$, $p < 0.0001$), as shown in the voltage maps in figure 11. Along the periphery, the Association effect additionally interacted with hemisphere ($F(1, 23) = 5.67$, $p = 0.026$), and appeared to be slightly larger over the right hemisphere. Main effects and interactions involving Relationship did not approach significance in either the mid-regions (ps > 0.1) or peripheral regions (ps > 0.1) omnibus ANOVA.

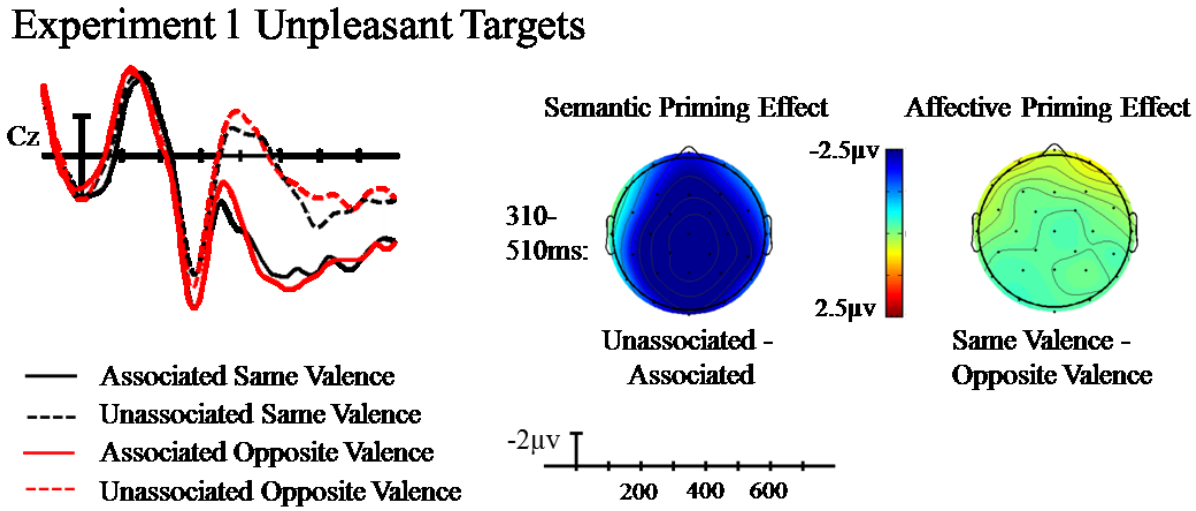

Figure 11: Effects Association and Relationship for emotional targets in experiment 1. A robust N400 semantic priming effect was observed, maximally over centro-parietal electrodes. In contrast, no affective priming effect was observed.

LPC: 510-660ms. A modest late positivity was observed following the N400, peaking approximately 570-600ms after stimulus onset. A main effect of Association was highly significant for both the mid-regions ($F(1, 23) = 66.3$, $p < 0.0001$) and peripheral regions ($F(1, 23) = 58.7$, $p < 0.0001$) omnibus ANOVA, with associated words eliciting a larger positivity than unassociated words. As shown in figure 11, this is likely a

continuation of the Association effect that began in the N400 time window. This Association effect additionally interacted with Region in both the mid-regions ($F(4, 92) = 23.5$, $p < 0.0001$) and peripheral regions ($F(1, 23) = 10.7$, $p < 0.005$) omnibus ANOVA, where the effect was maximal over parietal portions of the scalp (parietal mid-region: $F(1, 23) = 86.3$, $p < 0.0001$; posterior periphery: $F(1, 23) = 93.4$, $p < 0.0001$), but was highly significant everywhere ($ps < 0.005$). Main effects and interactions involving Relationship did not approach significance in either the mid-regions ($ps > 0.1$) or peripheral regions ($ps > 0.1$) omnibus ANOVA.

Strength of the evidence for null effects of affective priming. In the a priori hypothesis tests, we found no evidence of affective priming (operationalized in the "Relationship" factor for unpleasant targets). However, a failure to reject the null hypothesis does not actually characterize the relative strength of the evidence in favor of the null hypothesis. To formally express the relative probability that the null hypothesis may be true (i.e. that affective priming does not influence either the N400 or LPC components under the current experimental conditions), we calculated a Bayes factor for the overall affective priming effect in both the N400 and LPC time windows, providing a ratio of the marginal likelihood of the null hypothesis given the data to the marginal likelihood of the alternative hypothesis given the data (Rouder, Speckman, Sun, Morey, & Iverson, 2009). To minimize the number of tests, we used a single spatio-temporal region of interest for the N400, and another for the LPC, where each component tends to

be maximal. Bayes factors were calculated with a "JBZ" (Jeffreys-Zellner-Siow[3]) prior using version 0.9.8 of the BayesFactor package (described in Rouder et al., 2009).

During the N400 time-window of 310-510ms, the central mid-region yielded a Bayes factor of 5.598. During the LPC time-window of 510-660ms, the parietal mid-regions yielded a Bayes factor of 4.509. In other words, given the data collected within the spatio-temporal regions of interest for the N400 and LPC components, we found that the null hypothesis (that no affective priming occurred within this ROI) was roughly 5.5 and 4.5 times more probable than the alternative hypothesis (that affective priming elicited an effect within this ROI), respectively.

Experiment 1 Discussion

Experiment 1 crossed affective priming and semantic priming in a task that oriented attention towards the semantic features of the primes and targets. This was chosen to reveal the nature of priming effects that occur when emotion is not overtly task-relevant. Associated words elicited a smaller N400 amplitude than unassociated words, with a maximal difference over centro-parietal electrodes. This semantic priming effect was sufficiently robust to surpass even a five-sigma threshold of significance, and exhibited the anticipated direction, distribution, and time course. However, affective priming effects were not discernible in either the N400 or LPC components. This holds a

[3] Here, a Cauchy prior on effect size (a t-distribution with df=1: see Liang, Paulo, Molina, Clyde, & Berger, 2008) is combined with a Jeffreys prior on variance. The scale value for "r" was left at the default 1.0.

number of implications for the different hypotheses of ERP affective priming effects.

In the valenced features hypothesis, strictly controlling for semantic features should eliminate N400 affective priming effects. We found this to be the case. Bayes factors indicated that the null hypothesis (of no affective priming on the N400) was substantially more probable given the data than the alternative hypothesis (that affective priming occurred during the N400 component to a measurable extent). Critically, this found to be true not only for the randomly-paired words, but also for the associated pairs where the relationship between the primes and targets was overtly specified, measured, and matched (using forward association strength and synonym-antonym relationships exclusively). In other words, evidence in favor of the valenced features hypothesis (including positive evidence from the Bayes factor) was found both for words intentionally matched to a minimal degree of semantic relatedness and for words matched to a high degree of semantic relatedness, i.e. where semantic features were controlled across multiple levels.

In the semanticized valence hypothesis, we anticipated that N400 priming effects should be present in a semantic task, where semantic contingencies (including valence) would be most valuable. We did not find evidence that this was the case. No affective priming effects were evident on the N400 component, as described above. However, it's important to note that the valence of the target could be predicted *exactly* on every single trial. Neutral primes always preceded neutral targets, while emotional primes (both pleasant and unpleasant) always preceded unpleasant targets. As such, it's possible that affective priming effects were not discernible because they did not actually manipulate

the *predictability* of target valence, even if they manipulated the *overlap* between the valence of the prime and the valence of the target. Though this stands in contrast to many network models of priming, where the prime activates semantic features necessary for target processing, other theories of semantic priming indicate that what matters isn't the *overlap* in semantic features between the prime and target, but the *contingency* in semantic features (Kuperberg & Jaeger, In Press; Lau, Holcomb, et al., 2013), where the pleasant primes could have been used to predict and pre-activate unpleasant targets (discussed further in chapter 4).

Finally in the evaluative congruity hypothesis (where affective priming is driven primarily by facilitation of congruent task-relevant evaluations of emotional significance), we anticipated that LPC priming effects should be minimal when emotional evaluations were not overtly task-relevant. We found this to be the case: no affective priming effects were visible on the LPC, regardless of whether the words were associated or unassociated. However, the evaluative congruity hypothesis instead predicts that a valence-oriented task should yield robust LPC modulation. We explored this possibility in experiment 2.

Experiment 2 Introduction

Experiment 2 was designed to complete the investigation of affective priming for both the semanticized valence hypothesis and the evaluative congruity hypothesis. First, the perfect statistical contingency between emotional primes and unpleasant targets was

broken by the addition of fillers. The experimental materials analyzed for inference stayed the same, but an additional set of word pairs with pleasant targets was added, such that pleasant and unpleasant primes seen by participants were equally likely to precede pleasant and unpleasant targets. If overlap in semanticized valence is sufficient to drive affective priming, we expect to see affective priming effects on the N400 for the emotional word pairs. Alternatively, if the statistical contingencies along the valence dimension are sufficient to drive affective priming (the alternative explanation offered above), we would expect larger *semantic* priming effects for the *neutral* than the emotional words, as the neutral valence of the target could still be predicted with 100% accuracy while the pleasant or unpleasant valence of the emotional targets could only be predicted with a 50% transition probability.

Second, we asked participants to judge the affective relationship between the primes and targets rather than the semantic relationships. In the evaluative congruity hypothesis, we would now expect an LPC effect of affective priming, as an evaluation of valence is now overtly relevant to task.

Though affective judgments of the target alone are more common (Herring et al., 2013), we instead utilized the affective equivalent of the semantic relatedness task, where both the prime and the target are overtly task-relevant, and there is no time pressure to respond. Specifically, participants were asked to attend to the valence of the primes and the targets, and to indicate via button presses whether the valences of the two words where the same (e.g. both negative) or different (e.g. one was positive and one was negative). A new group of participants were asked to judge all of the same items as were

in experiment 1, meaning that participants also evaluated whether the neutral word pairs shared the same valence or not. Rather than make "neutral" a valence category that could be evaluated, we encouraged participants (through instruction and practice) to judge the *minor* positive and negative connotations of the words, known as microvalences (Lebrecht, Bar, Barrett, & Tarr, 2012). Participants found this task difficult, but doable with practice.

Once again, we considered whether an affective priming effect would be observed on the N400 component (indicating a semanticized valence mechanism), the LPC (indicating an evaluative congruity mechanism), or neither (potentially consistent with semantic features that may have otherwise overlapped if not controlled for).

Experiment 2 Methods

The same stimuli were used as in experiment 1, with the addition of fillers. Fillers consisted of pleasant targets that were preceded by pleasant or unpleasant primes, which were either associated or unassociated. The procedure for generating these pairs was the same as for the unpleasant-target pairs in experiment 1: synonym-antonym-target triplets were obtained from the Florida word association norms (Nelson et al., 2004), triplets were selected for the experiment that most closely matched the primes and targets (and prime-target relationships) of the other stimuli, and unassociated pairs were created by scrambling primes and targets in three different iterations. However, this procedure did not yield a sufficient number of related synonym-antonym-target triplets: we ended up

with half the number we needed in order to balance transition probabilities. We therefore augmented these items with additional synonym-antonym-target triplets that appeared to match the criteria we were interested in, but were not indexed in the Word Association Norms (Nelson et al., 2004). Lexical features, valence, and arousal were matched to the experimental words as closely as possible. Data from fillers were not analyzed.

Data from 27 young adults (11 men) was collected. One participant was excluded due to computer malfunction during data collection, and two other participants were excluded due to excessive artifact, leaving 24 participants for analysis. All participants met the same inclusion criteria as for experiment 1.

EEG recording and stimulus presentation were identical to experiment 1. Presentation is summarized in Figure 9. Trials containing artifact were rejected, constituting 2.1% of trials (with no difference between conditions). Postprocessing and analysis were otherwise also the same as experiment 1.

Experiment 2 Results

Behavioral Results. Mean accuracy was defined as the proportion of times that participants correctly identified when an emotional word pair had two words of the "same" or "different" valences. Only word pairs with unpleasant targets were included in this analysis, as there was no "correct" answer for neutral word pairs, and the word pairs with pleasant targets were simply fillers. Overall, mean accuracy was 93.7% (SD 5.9%). Participants did not significantly differ in their ability to identify same-valence pairs and

opposite-valence pairs ($t(23) = 1.64$, $p > 0.1$), but were significantly more accurate in identifying shared valence for semantically associated pairs than for semantically unassociated pairs (95% CI = [3.51%, 12.29%] more).

Neutral word ERP results. As with experiment 1, Neutral words were assessed using 2 Association (associated, unassociated) x 2 Relationship (synonym, antonym) mid-regions and peripheral regions omnibus ANOVAs that additionally incorporated spatial factors (see figure 3 for scalp regions).

N1: 85-165ms. An interaction between Association and Hemisphere was significant over peripheral electrodes ($F(1,23)=7.94$, $p=0.01$), where unassociated words elicited a larger negativity over the right hemisphere than associated words. No further effects or interactions approached significance over either the midline or peripheral regions ($ps>0.2$).

P2: 250-310ms. A main effect of Association reached significance over midline regions ($F(1,23)=5.87$,$p=0.024$), and was likely an early start to the N400 semantic priming effect (where unassociated words elicit a larger negativity than associated words). An Association by Relationship by Region 3-way interaction was also significant over both mid-regions ($F(4,92)=4.26$, $p=0.029$) and peripheral regions ($F(1,23)=5.63$,$p=0.026$) omnibus ANOVAs. Follow-up ANOVAs within each region did not yield any regions with a significant Association by Region interaction. However, when we conducted follow-up ANOVAs separately for both associated and unassociated words, we found a significant Relationship by Region interaction for associated words (mid-regions: $F(4,92)=4.28$, $p=0.024$; peripheral regions: $F(1,23)=4.66$, $p=0.042$), but not

for unassociated words. The effect of Relationship for associated words was only significant in the prefrontal region, where antonyms elicited a larger positivity than synonyms.

N400: 310-510ms. A large main effect of Association was significant over mid-regions ($F(1,23)=26.8$, $p<0.001$) and peripheral regions ($F(1,23)=20$, $p<0.001$) omnibus ANOVAs, with unassociated words eliciting a larger negativity than associated words. This additionally interacted with Region in the mid-regions ANOVA ($F(4,92)=4.85$, $p=0.01$), with follow-up simple-effects ANOVA indicating that the Association effect was largest in central and parietal mid-regions (see figure 12). Finally, an Association by Relationship by Region three-way interaction reached significance over peripheral regions ($F(1,23)=4.9$, $p=0.037$), but follow-up tests within region did not yield any main effects or interactions involving Relationship.

LPC: 510-660ms. A main effect of Association was observed for both mid-regions ($F(1,23)=24.9$, $p<0.001$) and peripheral regions ($F(1,23)=23.6$, $p<0.001$) omnibus ANOVAs. As shown in figure X, this was likely due to a continuation of the N400 semantic priming effect, where unassociated words elicited a larger negativity than associated words. This additionally interacted with Region in the mid-regions ANOVA ($F(4,92)=3.3$, $p=0.038$), with follow-up simple-effects ANOVAs indicating that the Association effect was largest in central and parietal mid-regions (see figure 12). A three-way interaction between Association, Relationship, and Region was significant for both mid-regions ($F(4,92)=3.59$, $p=0.034$) and peripheral regions ($F(1,23)=9.52$, $p=0.05$) omnibus ANOVAs. However, no main effects of Relationship or interactions between

Association and Relationship reached significance within any single region.

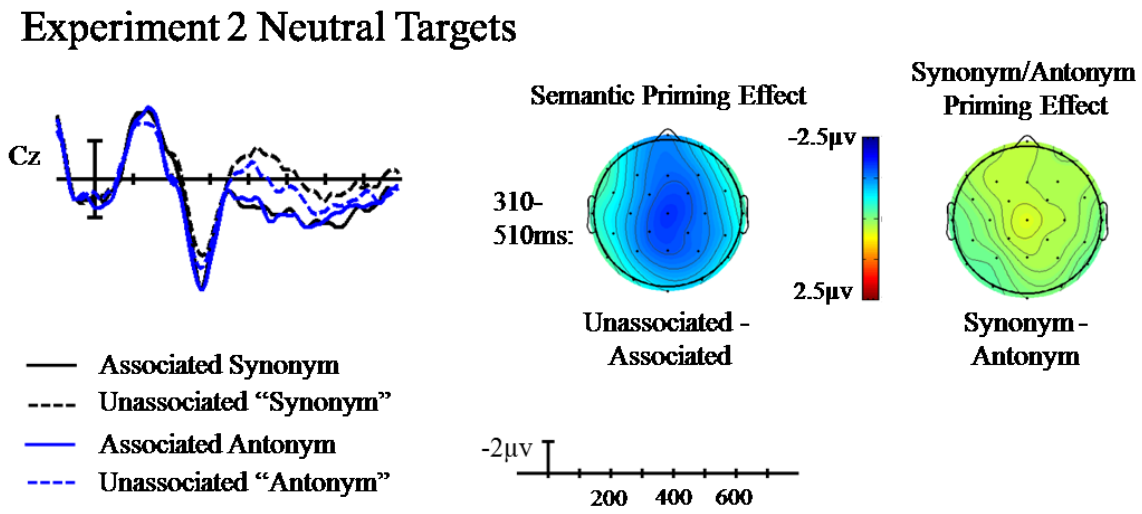

Figure 12: Effect of Association and Relationship for neutral targets in experiment 2. A robust N400 semantic priming effect was observed, maximally over centro-parietal electrodes, and did not differ by relationship type.

Emotional word ERP results. As in experiment 1, unpleasant words were assessed using a 2 Association (associated, unassociated) x 2 Relationship type (same-valence synonym, opposite-valence antonym) ANOVA that additionally incorporated spatial factors for mid-regions and peripheral electrodes (see figure 1 for scalp regions).

N1: 85-165ms. A main effect of Relationship was significant for both mid-regions ($F(1,23)=5.9$, $p=0.023$) and peripheral regions ($F(1,23)=5.89$, $p=0.023$) omnibus ANOVAs. As shown on figure 13, opposite-valence words elicited a larger negativity than same-valence words.

P2: 250-310ms. A main effect of Association was observed in the mid-regions omnibus ANOVA ($F(1,23)=4.73$, $p=0.04$). A three-way interaction between Association, Relationship, and Region was also significant over mid-regions ($F(4,92)=3.29$, $p=0.046$),

though follow-ups within each region did not yield any significant effects or interactions involving Relationship.

N400: 310-510ms. A main effect of Association was observed in both the mid-regions ($F(1,23)=37.9$, $p<0.001$) and peripheral regions ($F(1,23)=20.7$, $p<0.001$) omnibus ANOVAs, with unassociated words eliciting the larger negativity than associated words. This Association effect additionally interacted with Region (mid-regions: $F(4,92)=6.41$, $p=0.003$; peripheral regions: $F(1,23)=4.44$, $p=0.046$), where the prefrontal mid-region and anterior periphery yielded smaller effects than the rest of the scalp. Finally, a Relationship by Region interaction reached significance over peripheral electrodes ($F(1,23)=5.488$, $p=0.028$), but follow-up tests within each peripheral region did not return any significant effects or interactions involving Relationship.

LPC: 510-660ms. Only a main effect of Association was significant in both the mid-regions ($F(1,23)=13.0$, $p=0.001$) and peripheral regions ($F(1,23)=13.9$, $p=0.001$) omnibus ANOVAs. As shown in figure 13, this was likely a continuation of the N400 semantic priming effect. No effects or interaction involving Relationship reached significance.

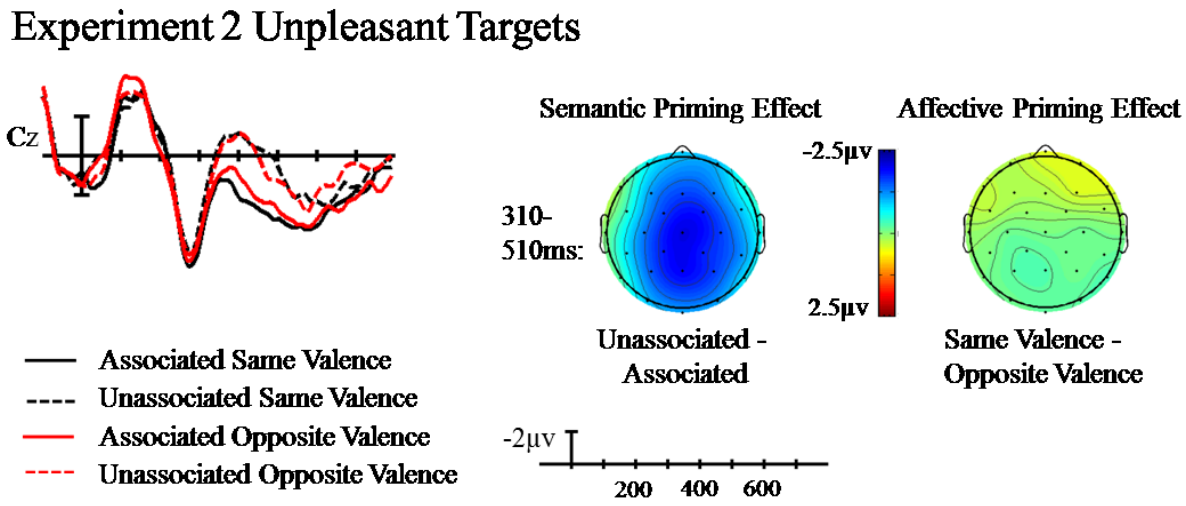

Figure 13: Effects Association and Relationship for emotional targets in experiment 2. A robust N400 semantic priming effect was observed, maximally over centro-parietal electrodes. In contrast, no affective priming effect was observed.

Strength of the evidence for null effects of affective priming. As in experiment 1, we found no evidence of affective priming (operationalized in the "Relationship" factor for unpleasant targets) within the N400 and LPC time windows. To formally express the strength of the evidence in favor of the null hypothesis (i.e. that affective priming does not influence either the N400 or LPC components under the current experimental conditions), we again calculated a Bayes factor for the overall affective priming effect in both the N400 and LPC time windows. The procedure was identical to the one used in experiment 1.

During the N400 time-window of 310-510ms, the central mid-region yielded a Bayes factor of 5.294. During the LPC time-window of 510-660ms, the parietal mid-region yielded a Bayes factor of 6.050. Given the data collected within the spatio-temporal regions of interest for the N400 and LPC components, we found that the null

hypothesis (that no affective priming occurred within this ROI) was roughly 5 and 6 times more probable than the alternative hypothesis (that affective priming occurred within this ROI), respectively.

Between-subject task effects. Visually, the semantic priming effect seemed to be larger in experiment 1 (the semantic task, see figure 14) than in experiment 2 (the valence task, see figure 14). Additionally, we were interested in secondary questions of whether A) the semantic priming effect would be different for neutral words than for emotional words, and B) the effect of target valence (Neutral versus Unpleasant) differed as a function of task (semantic task versus valence task). We conducted additional omnibus ANOVAs in the N400 and late positivity time window to test these three hypothesis. For the mid-regions omnibus ANOVA, "Experiment" was included as a between-subjects factor, with Association, Target Valence, and Region included as within-subjects factors. For the peripheral regions omnibus ANOVA, Experiment was included as a between-subjects factor, with Association, Target Valence, Region (anterior versus posterior), and Hemisphere (left, right) included as within-subjects factors.

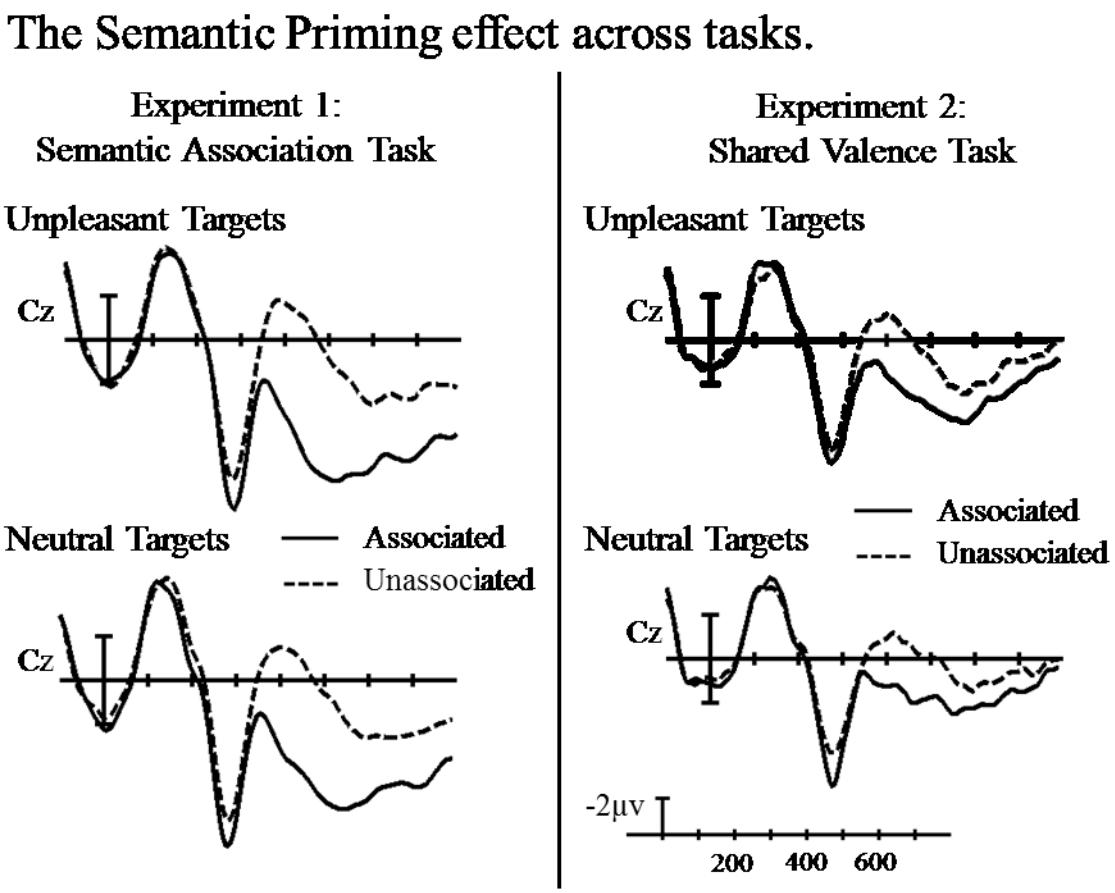

Figure 14: The semantic priming effect plotted separately for emotional words and neutral words in experiments 1 and 2. Two main effects are apparent. First, the overall semantic priming effect in the semantic task was larger than the overall semantic priming effect in the valence task. Second, the overall semantic priming effect was larger for emotional words than for neutral words. No interactions reached significance.

N400: 310-510ms. The main effect of Association was naturally highly significant over both the mid-regions ($F(1,46)=157$, $p<0.0001$) and peripheral regions ($F(1,46)=122$, $p<0.0001$), but Association also significantly interacted with Experiment (mid-regions: $F(1,46)=18.7$, $p<0.001$; periphery: $F(1,46)=17.4$, $p<0.001$), with experiment 1 yielding larger semantic priming effects than experiment 2. A three-way interaction between Association, Experiment, and Region in the mid-regions omnibus ANOVA ($F(4,184)=8.495$, $p<0.001$) further indicated that the distribution of the semantic priming

effect in experiment 1 significantly differed from the distribution of the semantic priming effect in experiment 2 (see figure 14).

A main effect of Target Valence reached significance in both the mid-regions ($F(1,46)=9.77$, $p=0.003$) and peripheral regions ($F(1,46)=11.3$, $p=0.002$) omnibus ANOVA. As shown in figure 15, this was likely due to overlap with the subsequent LPC component, with unpleasant targets eliciting a larger positivity than neutral targets.[4] The Target Valence effect did not interact with Experiment (even when including spatial factors) in either omnibus ANOVA.

Finally, a four-way interaction between Target Valence, Association, Region, and Hemisphere reached significance over peripheral electrodes ($F(1,46)=5.03$, $p=0.03$): in a particular region of the periphery, the semantic priming effect may have differed depending on whether the words were emotional or neutral. Follow-up ANOVAs isolated the Target Valence by Association interaction to the left posterior periphery ($F(1,46)=4.55$, $p=0.038$), where the semantic priming effect for emotional words was larger (partial eta^2=0.585) than the semantic priming effect for neutral words (partial eta^2=0.463).

LPC: 510-660. The main effect of Association was still highly significant over both the mid-regions ($F(1,46)=101$, $p=3.3x10^{-13}$) and peripheral regions ($F(1,46)=76.3$, $p=2.5x10^{-11}$), and still significantly interacted with Experiment (mid-regions: $F(1,46)=15.48$, $p<0.001$; periphery: $F(1,46)=7.14$, $p=0.01$), with experiment 1 yielding larger semantic priming effects than experiment 2. The three-way interaction between

[4] The Target Valence effect additionally interacted with region in both tests.

Association, Experiment, and Region in the mid-regions omnibus ANOVA ($F(4,184)=5.01$, $p=0.007$) was also still significant. All of these effects are likely just a continuation of the N400 semantic priming effects (see N400 results above).

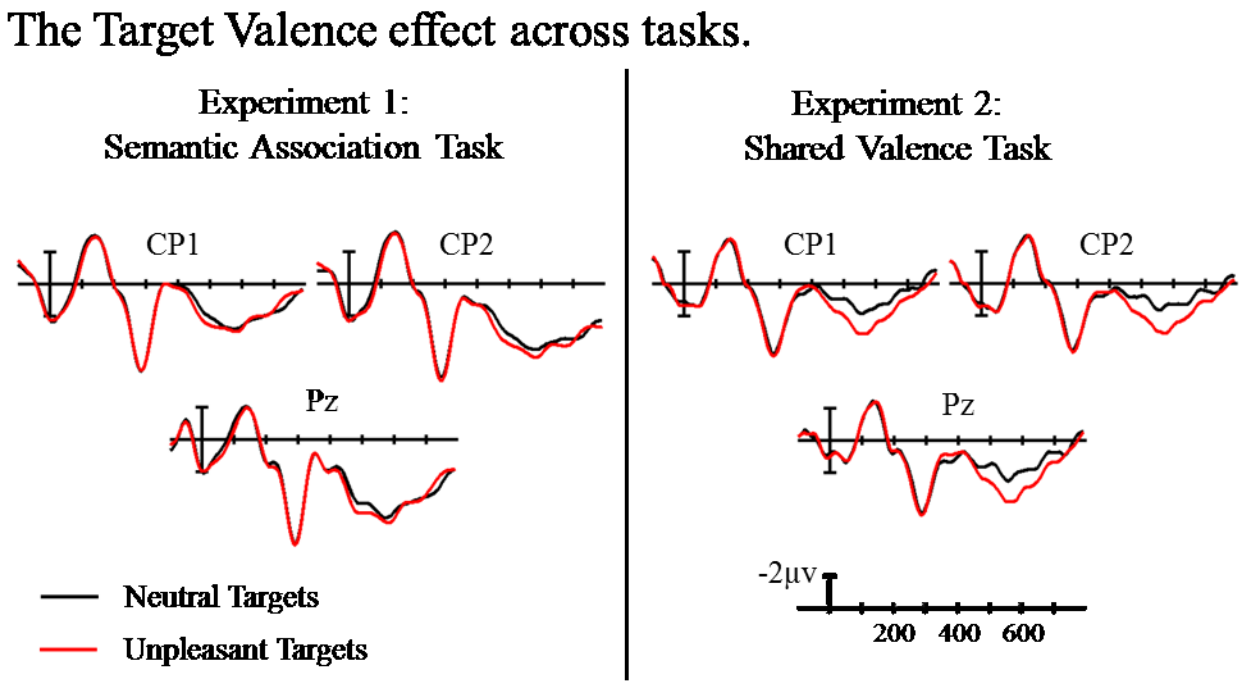

Figure 15: The valence effect for experiment 1 and experiment 2. Overall, the LPC showed a larger positivity for unpleasant targets than for neutral targets. However, this effect did not significantly differ by task ($p = 0.088$).

A main effect of Target Valence reached significance in both the mid-regions ($F(1,46)=19.2$, $p<0.001$) and peripheral regions ($F(1,46)=17.1$, $p<0.001$) omnibus ANOVAs, with unpleasant targets eliciting a larger positivity than neutral targets. As shown in figure 15, the valence-oriented task in experiment 2 appeared to elicit a larger valence effect than the association-oriented task in experiment 1. However, the Target Valence effect did not significantly interact with Experiment, though a Valence by Experiment interaction ($F(1,42)=3.03$, $p = 0.088$) and a Valence by Region by

Experiment interaction ($F(4,184)=2.782$, $p=0.069$) were close to significance.

Finally, a four-way interaction between Target Valence, Association, Region, and Hemisphere reached significance over peripheral electrodes ($F(1,46)=5.125$, $p=0.028$), but follow-up tests within each hemisphere did not yield any significant interactions involving Target Valence and Association.

General Discussion

The present study was designed to assess the manner in which affective priming and semantic priming relate to one another. We implemented a full cross of semantic priming (associated vs. unassociated) and affective priming (same valence, opposite valence) in a large, carefully counterbalanced set of words. Our results were striking.

As anticipated, a robust N400 Association effect was observed for both neutral and emotional words that was widespread, highly significant, and extended until the end of the ERP analysis window. As seen in figures 10-14, unassociated words elicited a larger negativity than associated words starting about 300ms post-stimulus. The N400 semantic priming effect was found to be larger during the semantic task than the valence task, and larger for emotional than for matched neutral pairs in both tasks.

In contrast, no signs of affective priming were present for either the N400 component or the LPC component in either study. Bayes factors indicated that the affective priming manipulation was 4-6 times more probable (given the data) to *not* influence the N400 and LPC components than it was to modulate them. Because

unpleasant targets as a whole elicited greater LPC amplitudes than neutral targets, as shown in figure 15, it is unlikely that the emotional words used in the study were simply not sufficiently emotional. Overall, our results indicate that when semantic associations have been carefully disentangled from valence, affective priming effects were diminished to the point of non-significance, even when participants were asked to attend to the emotional relationships between primes and targets overtly.

Implications for Theories of Affective Priming. The present data are notably only consistent with only one of the three proposed mechanisms of affective priming discussed in the introduction.

First, we considered the "valenced features" hypothesis, where the affective priming effect may primarily be driven by overlap in the semantic or conceptual features that correlate with valence. We hypothesized that after prime-target associations have been controlled for, no effect of affective priming should be apparent on the N400 (as described above). Our data are consistent with this hypothesis. Both experiments provided evidence that the null hypothesis was more probable, given the data, than the alternative hypothesis (of an affective priming effect on the N400 or LPC component). Compellingly, a prior behavioral study comparing semantic to affective priming found a similar pattern: across two different tasks, semantic priming effects were robust, while affective priming manipulations failed to influence reaction times at all (Storbeck & Robinson, 2004). This supports the conjecture from Herring and colleagues (Herring et al., 2011) that some previously reported affective priming effects may have been inadvertent semantic priming effects instead.

Second, we considered the "semanticized valence" hypothesis, where emotional properties like valence are themselves treated as semantic features during word processing, making affective priming simply a subtype of semantic priming. We hypothesized from this that affective priming effects should be apparent on the N400 component, as with some previous studies (J. P. Morris et al., 2003; Zhang et al., 2006; Zhang et al., 2010), even after all other semantic features have been controlled for (Storbeck & Clore, 2007). However, no evidence of an N400 affective priming effect was found in the present study, regardless of which type of relationship (semantic or valence) participants were instructed to monitor. Our data do not support the semanticized valence hypothesis.

This suggests that some previous ERP studies may have found affective priming effects on the N400 component (J. P. Morris et al., 2003; Zhang et al., 2012; Zhang et al., 2006; Zhang et al., 2010) because semantic relationships in previous studies may have inadvertently been left uncontrolled. More generally, picture primes (used in Zhang et al., 2006; Zhang et al., 2010) may just be difficult to relate semantically to word targets in an operationalized manner (Herring et al., 2011). To date, there is now evidence to suggest that controlling for semantic similarity ratings (Hinojosa et al., 2009), category membership (Herring et al., 2011), and association strength (the present experiments) are all sufficient to eliminate N400 affective priming effects. The preponderance of the data supports the hypothesis that some previously reported affective priming effects on the N400 may have actually been inadvertent semantic priming effects. Because of emotion's central role in conceptual similarity and dissimilarity (Samsonovic & Ascoli, 2010), a

manipulation of valence is also a manipulation of meaning, and words that share a valence may be more likely to be conceptually similar than words that do not share a valence.

And finally, we considered the "evaluative congruity" hypothesis, where the affective priming effect relies primarily on evaluative processes. We hypothesized from this that affective priming effects should be apparent on the LPC, as in some previous ERP studies of affective priming (Herring et al., 2011; Hinojosa et al., 2009; Zhang et al., 2012). In particular, given the emphasis on response demands in describing the affective priming effect, we anticipated that such LPC modulation would be larger for the valence task than the semantic task. However, no evidence of a late positivity affective priming effect was found using either task. Our data did not support the evaluative congruity hypothesis.

This raises the natural question of why these previous studies may have elicited LPC affective priming effects while the present two experiments did not. These prior studies differ from the present experiments in a number of ways, which could potentially explain the divergence. The first study (Hinojosa et al., 2009) modulated the arousal instead of the valence of primes and targets, and it's possible that the extended evaluations reflected on the LPC are influenced primarily by the relative salience (rather than the valence) of the stimulus, at least under some conditions (Delaney-Busch, 2013; Delaney-Busch et al., *under review*). This was supported by Zhang and colleagues (2012), where arousal congruity was found to exclusively modulate the LPC. In the final study (Herring et al., 2013), prime and target stimuli were limited to a small set of

animal-related and person-related words that were repeated a considerable number of times. While the present study only showed each target exactly four times and each prime exactly twice, the words in Herring et al. were shown an average of 25.6 times each (some as the prime and others as the target). Higher numbers of repetitions may have attenuated the lexico-semantic processing costs of these words, perhaps making evaluative significance more accessible or salient by comparison.

However, the two experiments presented here yielded one important extension to these previous suggestions: the statistical contingencies between the prime valence and the target valence did not appear to matter. In experiment 1 (the semantic task), every pleasant and unpleasant prime *always* preceded an unpleasant target, without fail. Similarly, neutral primes always preceded neutral targets, without fail. In contrast, in experiment 2 (the valence task), a number of fillers were added, such that pleasant and unpleasant primes were equally likely to precede a pleasant target as an unpleasant target. As described in chapter 4, such a radical decrease in the predictive validity of prime valence might be expected to significantly reduce the processing advantage to the targets if valence is involved in semantic processing. During comprehension, people generally appear to naturally adapt to their expectations to whatever is most felicitous given the context (Chang, Janciauskas, & Fitz, 2012; Dell & Chang, 2014; Hale, 2011; Kleinschmidt & Jaeger, 2015; Kuperberg & Jaeger, In Press; Levy, 2008; Lewis & Bastiaansen, 2015). The present study offers a notable instance in which people did not appear to adapt to a clear contingency. In one sense, this is not so surprising. In chapter 4, we present a case of adaptation where lower predictive validity in the surrounding

context decreased the magnitude of the N400 semantic priming effect (Lau, Holcomb, et al., 2013). We might have similarly expected that lower predictive validity may decrease the magnitude of an N400 affective priming effect. However, no N400 affective priming effect was found in the present experiments, even when predictive validity was high, and thus there was no effect to attenuate. This serves a deeper point, however, that the implicit learning and adaptation mechanisms evidenced in chapter 4 may have important limitations. Despite the fact that the valence of the target could, in experiment 1, be predicted perfectly on every trial, participants did not appear to gain any benefit during word processing from the contingency.

There are a few potential explanations for the lack of adaptation to the predictability of target valence. The concurrent manipulation of task, of course, is important to consider first. By design, the experiment with the lower predictive validity for prime valence (experiment 2) also utilized the task that overtly deployed participants' attention towards valence. As such, it is possible that the apparent lack of affective priming effects across experiments may have instead been due to a task effect (where the valence task increased reliance on valence features) working in opposition to an adaptation effect (where the lower predictive validity in experiment 2 decreased reliance on valence features). Though theoretically plausible, inferring two equal and opposite overlapping effects is significantly less parsimonious than inferring no effect, and this explanation is thus dispreferred.

However, task may have contributed to the lack of adaptation effects in other ways. Particularly, both tasks may have discouraged evaluative *predictions* for the target.

Though the investigation in chapter 4 indicates that participants can adapt to even task-irrelevant contingencies between primes and targets, both of the present tasks involved overtly relating the primes to the targets. This may have encouraged a deliberative *retrospective* comparison of each prime and target, which obviated the demand for predictive cues. In addition, valence could simply be too diffuse of a feature to actually facilitate semantic processing significantly enough to measure with ERPs, even if target valence *was* actually predicted following the presentation of the prime. In other words, even if participants implicitly knew in advance that the target word must be unpleasant, there are so many unpleasant words in the English language that the knowledge might not actually impose sufficient constraints on the target to meaningfully facilitate processing of the target.

But ultimately, the present data could also signify that prediction (and thus adaptation) may have constraints. Specifically, while the current evidence suggests that prediction unfolds in parallel across multiple levels of comprehension (Kuperberg & Jaeger, In Press), from syntax to semantics to phonology, we speculate that the present results could indicate that predictions at the same or similar levels could be subject to some sort of triage or prioritization. Participants may not have adapted to the valence contingency in the present studies because they could already adapt to the *semantic associative* contingencies *instead*, which provide significantly more information about the actual identity of the targets. We return to this idea later in the discussion.

Alternative explanations for the affective priming results. While the "valenced features" hypothesis is our preferred interpretation of the observed pattern of effects,

there are a number of alternative interpretations that are also consistent with the data.

While the present experiments make a number of significant contributions to the understanding of affective priming (discussed above), there are additional considerations and alternative explanations that must be considered. Foremost among these is the difficulty of interpreting a null effect of semantic priming. Though only one of the three proposed mechanisms of affective priming is actually consistent with the present pattern of results, the *positive* evidence is limited to the Bayes factor. The subjects ANOVAs only indicated that the null hypothesis for affective priming effects on the N400 and LPC components should continue to be assumed, meaning that the predictions of the "evaluative congruity" and "semanticized valence" hypotheses were only *unsubstantiated* (rather than rejected by strong inference). The Bayes factors were calculated to characterize the actual strength of the evidence in favor of the null hypothesis more directly. This required the additional assumption of a prior (in this case, the JBZ prior) and still remained strongly dependent on statistical power for interpretation. But overall, our data can formally be considered to a) not substantiate N400 or LPC effects of affective priming, and b) indicate that the null hypothesis may be slightly more probable than the alternative hypothesis in both time windows. Such evidence can still allow for important contributions to understanding, following the example of the (null) affective priming effects reported in the behavioral experiments by Storbeck and Robinson (2004).

It is also possible that the tasks or contexts used for the present investigation were insufficient in some way to elicit an affective priming effect. This would itself be an important finding, but it does lead to a slightly different set of inferences. A recent large

metanalysis of both published and unpublished studies (Herring et al., 2013) systematically determined a number of situational influences on the affective priming effect that are evidenced given the state of the field. This indicated, for example, that affective priming effects tended to be larger or more frequently found for evaluative decision tasks than for lexical decision tasks, for primes with short SOAs than long SOAs, for studies with just noun targets than for studies that also included adjectives and verbs, and also stimuli that had been grouped into a greater number of blocks (to allow participants to take short breaks) than for stimuli that had been grouped into a smaller number of blocks. In every instance save for the target part of speech, we implemented the methodology that should have yielded the largest effect.

However, the exact nature of the tasks used for the present study could have also induced attenuated affective priming effects. Even in the significantly more extensive behavioral literature on affective priming, response demands are typically limited to either the prime position and/or the target position alone (Herring et al., 2013). In the present experiments, participants were overtly asked to provide a response that pertained to the relationship *between* the prime and the target (either associated relationship or emotional relationship). It is possible that these tasks emphasized deliberative *retrospective* processing rather than automatic processing, even with the short SOAs we used. In the minimal sense, the present experiments would then suggest that null effects of affective priming may be a possibility for tasks that emphasize retrospective matching. However, if the present tasks were assumed to allow for the possibility of affective priming given an appropriate stimulus set (such as one that had elicited an affective

priming effect in previous studies), the present data could then potentially inform the understanding of affective priming more generally, as described above.

However, previous behavioral research has also indicated one particularly incisive way that local context can influence the affective priming effect: in essence, affective priming might be strictly subsidiary to semantic priming. In other words, affective priming may only be likely to occur when other semantic features of the stimuli are not only controlled, but *uniform* (for instance, when they are all drawn from the same semantic category). In one set of experiments (Storbeck & Robinson, 2004), an affective priming effect elicited by stimuli of a uniform semantic category was attenuated to non-significance when semantic features were not uniform (also reviewed and discussed in Storbeck & Clore, 2007). Combined with the present data, there are now a number of experiments spanning four different tasks where affective priming effects are not found in the presence of semantic priming effects, in both reaction times and ERP indices. As the such, instead of informing which of several mechanisms could contribute to ERP affective priming effects, the present data could be interpreted as providing further evidence that affective priming may be "washed out" in the presence of a strong semantic priming manipulation.

Storbeck and colleagues attributed the seemingly subsidiary nature of affective priming in the presence of semantic priming to a "semantic categorization [that] precedes the retrieval of affective associations"(Storbeck & Robinson, 2004). But we suggest that semantic relationships in these studies do not necessarily need to always be processed faster than emotional relationships, as such a strict affective primacy does not seem to be

well-supported by the evidence (Lai et al., 2012). Instead, semantic relationships may be *superior sources of information about the target*. Knowing the category membership of the target, as in Storbeck and Robinson (2004), or being able to anticipate the lexical item directly through forward association strength, as in the present studies, is significantly more informative of further semantic features of the target than knowing its valence. It's plausible that the valence of the target word is just not a large enough of a contributor to the work required for lexico-semantic processing for valence to be a meaningful signal.

Compellingly, there is parallel behavioral evidence that *semantic* priming can also be attenuated when a very large proportion of the trials include *repetition* priming, where the prime and target are the same word (Snow & Neely, 1987). In that case, repetition priming is substantially more informative about the semantic features of the target than semantic priming, and semantic priming effects are subsequently diminished.

Generally, these studies may suggest that when more than one statistical contingency is present in an experiment, participants appear to preferentially use whichever single contingency facilitates target processing the most (rather than additively utilize all the contingencies simultaneously). The present experiments are consistent with this view. Together with previous behavioral studies (Storbeck & Robinson, 2004), they may indicate that semantic priming effects may supersede affective priming effects even when valence is overtly task relevant.

If true, this has two primary implications. First, information about target valence and information about target semantics (i.e. stronger information about target identity) may not typically be utilized simultaneously and in parallel when one is significantly

more informative than the other. And second, even if emotion effects can generally occur with inordinate speed during word processing (Hofmann et al., 2009; Scott et al., 2009), the ability to report or make response-relevant decisions about the valence of the target may depend on evaluating the *identity* of the target. This requires just a slight modification of Storbeck and Robinson's original hypothesis: (lexico-)semantic processing may generally precede the evaluative processing *that is specifically required for response criteria* (rather than preceding all evaluative processing, which does not appear to be the case).

This possibility raises a clear prediction: that the semantically unassociated stimuli used for the present study would yield an affective priming effect if presented separately from the semantically associated stimuli. If so, it would suggest that the present null affective priming effects may have been a result of their subsidiary status to semantic priming effects. In contrast, if the same unassociated stimuli used for the present study still failed to produce an affective priming effect when presented alone, it could instead indicate that the null effects observed here were still simply the consequence of having rigorously controlled for semantic features, where word semantics systematically tend to relate to emotional features.

Implications for semantic priming. While the primary focus of these experiments was to shed further insight on the nature of the affective priming effect, these data also hold some compelling implications for the N400 semantic priming effect.

Like many other influences on lexico-semantic processing, the semantic priming effect is sensitive to task demands. For example, normally robust semantic priming

effects seen to related (vs unrelated) targets can be attenuated to non-significance when attention is directed towards particular features of the primes as part of the task (Maxfield, 1997) - the so-called "prime task effect". But even for tasks that still require deploying attention towards the target, the task-relevance of the *relationship* between the prime and the target can significantly change the magnitude of the N400 semantic priming effect. For example, the N400 semantic priming effect is considerably larger when participants are asked to overtly judge the semantic relatedness between the prime and targets as when they are asked to monitor the prime and targets for a particular semantic category (Kreher, Holcomb, & Kuperberg, 2006). The present experiments provide continued support for this pattern: the N400 semantic priming effect was significantly larger for the semantic task than the valence task, as evidenced by an interaction between task and association.

But in addition to this effect of task on the N400 semantic priming effect, these data can also clarify whether the particular *type* of semantic relationship between the primes and targets influences the resulting semantic priming effects. In particular, the present study included both synonyms and antonyms as associated primes (which were then scrambled into the corresponding unassociated prime-target pairs), which were matched to the same degree of forward association strength and other potential confounds. Previous behavioral research has suggested that both synonyms and antonyms yield comparably sized semantic priming effects on reaction times if matched on associative strength (Lucas, 2000; Perea & Rosa, 2002). The present experiments isolate this effect to the N400 ERP component. Specifically, for neutral prime-target pairs

(where emotionality of the primes was not also manipulated), synonym and antonym primes both yielded N400 semantic priming effects of the same size. We note that this is in line with the expectations of a predictive account of the N400 semantic priming effect (Lau, Holcomb, et al., 2013), where it does not appear to matter how exactly the prime relates to the target so long as the target is equally expectable.

Finally, the present study also addresses the question of whether emotional prime-target pairs tend to yield larger N400 semantic priming effects than neutral prime-target pairs. It's plausible that emotional words have sufficiently high arousal to differentially bias attention, where the salient emotional words captured attention and increased focus resulting in generally magnified semantic priming effects. We found this to be the case. Specifically, our analysis revealed that the N400 semantic priming effect was larger for emotional words than for neutral words along the posterior periphery. This occurred irrespective of task. Though we are not aware of previous ERP studies that have directly compared semantic priming effects to matched neutral and emotional words, our results are consistent with data suggesting that induced moods (which may similarly increase arousal and/or attention) also magnify the semantic priming effect (Hänze & Hesse, 1993).

Interestingly, however, the behavioral accuracy data from the present study yielded the opposite effect. In experiment 1, participants were significantly better at correctly identifying whether word pairs were associated when the pair was neutral (vs. when the pair was emotional). In other words, emotion seemed to facilitate the distinction between associated and unassociated words during comprehension, but seemed to

interfere with the distinction between associated and unassociated words during response production. This pattern of effects is evocative of previous work on the influence of semantic priming on the production of picture names (Blackford, Holcomb, Grainger, & Kuperberg, 2012): pictures proceeded by semantically related (vs unrelated) primes showed facilitation during processing (as indicated by the ERP amplitudes), but interference during the production of the picture name (as indicated by longer reaction times). Blackford and colleagues argued that while semantically related primes facilitated processing of the targets during comprehension, they also acted as a close competitor to the correct name during motor planning of the response.

Compellingly, accuracy was numerically the lowest for unrelated words of the same valence, where the "related" valence features may have conflicted with the "unrelated" semantic features pertinent to the actual response. We note however that the interference in reaction times in Blackford et al my not be directly comparable to the interference in accuracy scores for the present studies. Our participants were required to hold their responses through a short delay, so it's unknown whether reduced accuracy to emotional words would also correspond with increased reaction times. As such, this apparent divergence between the ERP semantic priming effect and the behavioral data should be considered as preliminary.

The behavioral data from the present experiments also yielded some other important insights. First, during the semantic task where participants were asked to overtly judge the relationship between the prime and target, a clear asymmetry emerged: participants were significantly better at identifying associated word pairs than

unassociated word pairs. This indicates that participants may have perceived a relationship between some of the randomly paired stimuli, causing them to erroneously identify them as "related" semantically. But because different participants saw different random pairs, it doesn't appear to be the case that some special subset of inadvertently related word pairs was driving this effect (and all unassociated word pairs were confirmed to have a forward association strength of 0). This raises the intriguing possibility that the perceived relationships may have been *subject*-specific rather than item-specific. In other words, each individual may have idiosyncratically perceived relationships particular to their own experience. For example, somebody might perceive the random pair "stop…sell" as "related" instead of "unrelated" if they regularly trade stocks (where a "stop order" involves selling stocks that have dropped in value to a particular limit), or "split…bent" as "related" if they often engage in activities where both properties (split and bent) are likely to occur (such as while chopping wood). Future studies are required to investigate how emotional words are processed not only from item to item, but from subject to subject as well (Delaney-Busch et al., *under review*).

Interestingly, associated words also improved participants' accuracy in identifying the *emotional relationships* between the primes and targets. In experiment 2, participants were asked to indicate whether each prime and target was the same valence (e.g. "both negative") or different valence (e.g. "one positive, one negative"). Though semantic association was orthogonally manipulated from valence relationship, so no contingency could be inferred by the participants, they nonetheless were better at identifying whether or not the word pair shared a valence if those words also happened to

be semantically associated. For example, participants could better identify "uptight…tense" as being the same valence than they could "repulsive…tense", despite the semantic association not being overtly relevant to the task. A closer investigation of behavioral responses indicated that unassociated same-valence primes (like "repulsive…tense") had by far the highest error rate, while even unassociated words with a different valence (e.g. "attractive…tense") yielded similar performance to the associated conditions (e.g. "uptight…tense"). This pattern of results (where unassociated same-valence words in particular provided some difficulty for participants) is somewhat peculiar, particularly since the ERP data did not reflect the facilitation. But as a whole, we speculate that semantically associated words may have been easier for participants to deliberatively evaluate than unassociated words because of our use of synonyms and antonyms, which provided a similarity/opposition cue for semantic relationships that corresponded with the required same valence / opposite valence response demanded by the task (as associated synonyms were always the same valence as the target while associated antonyms were always the opposite valence of the target).

Implications for the processing of emotional words. Finally, the present data also address the understanding of the emotional LPC. Generally, more emotional words elicit a larger LPC than less emotional words over centro-parietal electrodes (Citron, 2012; Kissler et al., 2006). However, this effect is critically dependent on task on context (Fischler & Bradley, 2006; Hajcak et al., 2010; Hajcak et al., 2012). Though emotion effects to words have previously been investigated using semantically-oriented and valence-oriented tasks, we are not aware of any previous studies investigating LPC

emotion effects under task demands that require *comparing* emotional and semantic features to other previously presented words. The present study addressed this gap. Overall, a robust LPC emotion effect was observed with the anticipated direction, time course, and spatial distribution. Although the mean effect was numerically larger in the valence task than the semantic task, as shown in figure 15, we did not find evidence that valence effect was significantly different between the tasks (with $p = 0.088$), despite a relatively high statistical power (with 48 participants viewing about 320 target words each). Thus, our hypothesis that the valence task would yield a larger emotion effect than the semantic task was unsupported. We suggest that the focus on prime-target relationships may require enough difficult deliberative decision-making as to diminish the usual impact of shifting the deployment of attention from semantic feature to emotional features, because the deliberative emphasis remains in both cases (possibly attenuating the sustained evaluation of motivational significance of the targets themselves, which is thought to underlie the LPC).

Conclusions. In sum, the present experiments provide for a number of important conclusions. The primary focus of the study was to determine the relationship between affective and semantic priming in a fully-crossed design. While both studies yielded a large N400 semantic priming effect, neither study yielded evidence of affective priming (or an interaction with semantic priming) on either the N400 or LPC time windows. This pattern is consistent with the hypothesis that some of the previously observed affective priming effects may be attributable to a nonrandom association between valence and meaning, as proposed in a previous investigation of the topic (Herring et al., 2011), but

was not consistent with other proposed mechanisms of affective priming. However, a number of important ways that emotion can influence the N400 *semantic* priming effect were uncovered. First, the N400 semantic priming effect was found to be larger for emotional prime-target pairs than for neutral prime-target pairs, possibly because the emotional prime-target pairs were attended to more intently. And second, the N400 semantic priming effect was found to be smaller when the task required an evaluation of the valence relationships than an evaluation of the semantic associated relationships between the primes and the targets, continuing the trend in the literature where semantically-oriented tasks elicit larger priming effects than other tasks (now including emotional tasks). And finally, we also provide evidence that carefully controlled synonyms and antonyms can both elicit the same N400 semantic priming effect, indicating that the extent that the target is predictable following the prime may be a more important predictor of facilitation than the exact nature of the relationship between the prime and the target. Both the pattern of semantic priming effects and the affective priming effects hold compelling implications for predictive models of language comprehension, and suggest intriguing areas for future research on the topic.

CHAPTER IV

NEVER NOT WRONG: INCREMENTAL AND CONTINUOUS ADAPTION TO THE STATISTICS OF THE CONTEXT IN A SEMANTIC PRIMING PARADIGM

Prediction appears to be a central feature of language comprehension (Kuperberg & Jaeger, In Press). Knowledge about the speaker, the goals and topics of the discourse, and the content under discussion (i.e. real-world knowledge) can all inform our expectations across multiple levels of representation about upcoming words. Particularly in noisy environments, these expectations about the nature of the speech signal can reduce comprehension errors and allow for understanding at the rapid rate of conversation (Kleinschmidt & Jaeger, 2015). When incoming speech is similar to previously encountered speech, whether at phonological (Bradlow & Bent, 2008), lexical (DeLong, Urbach, & Kutas, 2005; Smith & Levy, 2013), syntactic (Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; Levy, 2008), or semantic (Kuperberg, 2013; Yildirim, Degen, Tanenhaus, & Jaeger, 2013) levels, the known statistical contingencies can be used to facilitate accurate comprehension, and when incoming speech is appreciably different from previously encountered speech, the implicit prediction error allows for learning and adaptation to account for the novel statistical contingencies (Chang et al., 2012; Fine et al., 2013; Kleinschmidt & Jaeger, 2015; Qian et al., 2012). In the present investigation, we explored the nature of this adaptation during semantic processing by manipulating the contingencies between pairs of words (primes and targets) across blocks of an experiment.

When *semantic* information is activated prior to bottom-up input (i.e. predicted), the semantic processing for the incoming word is typically facilitated. Event-related potentials (ERPs), a direct online measure of neural activity over the time course of stimulus processing, indicate that this facilitation of semantic processing tends to occur between 300-500ms post-onset, on a negativity called the N400 component (Kutas & Hillyard, 1980, 1984). When a word in a particular context (such as a sentence or discourse) is more predictable, indexed by cloze probability[5] (Taylor, 1953), the N400 amplitude tends to be lower (i.e. less negative) than when a word is less predictable given the context (Kutas & Federmeier, 2011). But the context need not be elaborate for semantic pre-activation (and the corresponding facilitation) to occur. For example, the processing of a single word (called the "target") is facilitated when preceded by just a single other associated (Meyer & Schvaneveldt, 1971; Neely, 1976) or semantically related (Lucas, 2000) word (called the "prime"), corresponding to a reduction in N400 amplitude (Bentin et al., 1985; Holcomb, 1988; Rugg, 1985). This is called the "semantic priming effect".

The strength of the semantic priming effect depends on the extent to which the prime can pre-activate the semantic features of the target. Naturally, this includes primes that relate to the targets to a greater or lesser extent, but it also includes the *reliability* with which primes relate to the targets. Specifically, the semantic priming effect is

[5] Though this is a measure of lexical probability, anticipation of a lexical item can be expected to correlate strongly with anticipation of semantic, phonological, and orthographic features. As such, cloze probability strongly corresponds with N400 amplitudes, despite the fact the the N400 component itself may more precisely reflect semantic processing (Federmeier & Kutas, 1999; Kuperberg & Jaeger, In Press)

sensitive to the *proportion* of related pairs seen by the participant across a given contextual environment (den Heyer et al., 1983; Groot, 1984; Neely et al., 1989; Seidenberg et al., 1984). A recent ERP study isolated this proportion effect on the N400 component (Lau, Holcomb, et al., 2013): when the proportion of related prime-target pairs in a particular experimental block was 10%, participants showed a very small N400 semantic priming effect, but then when the proportion of the same related prime-target pairs was 50% in a subsequent block, the same participants showed a significantly enhanced N400 semantic priming effect. This pattern of effects was replicated using MEG and ERP measures (supplemented by fMRI) in a new set of participants (Lau, Gramfort, et al., 2013). These data indicate that participants may have predicted semantic features of the target more strongly when the predictive validity of the primes was higher.

The semantic priming effect is likely sensitive to predictive validity because participants adapt to changes in the statistical contingencies over time (Lau, Holcomb, et al., 2013). Lau and colleagues write that "once participants pick up on the fact that many of the word pairs form an associative unit, they began to try to predict the pair itself as a representation in working memory… different responses across relatedness proportion are only because of participants implicitly noticing the change in predictive validity *across time*" [emphasis ours]. We don't intend to imply that this adaptation is "active" or strategic. Rather, implicit adaptation is the natural consequence of updating prior expectations with new information (Chang et al., 2012; Fine et al., 2013; Kleinschmidt & Jaeger, 2015; Qian et al., 2012). This iterative cycle of updating prior expectations allows comprehenders to hone in on representing the pattern of statistical regularities that

minimizes prediction error (Kuperberg & Jaeger, In Press) and maximizes the utility of predictions (H. Feldman & Friston, 2010; Friston et al., 2015). Such minimization of prediction error is an important computational principal across many domains of cognition and perception (A. Clark, 2013) beyond language.

However, adaptation of the N400 semantic priming effect to changes in predictive validity has not to our knowledge been directly observed. The effect of relatedness proportion (e.g. 10% vs 50% related) across blocks on the semantic priming effect is thought to be a consequence of the adaptation that occurs *within* blocks (den Heyer et al., 1983; Groot, 1984; Lau, Holcomb, et al., 2013; Neely et al., 1989; Seidenberg et al., 1984). And the adaptation of prediction error should manifest as adaptation of the N400 component, which clearly shows block-level change (Lau, Holcomb, et al., 2013) and is thought to reflect the extent to which prediction error has been minimized (Kuperberg & Jaeger, In Press). As such, how exactly the N400 semantic priming effect evolves over *trials* as participants adapt is largely an open question. Recently, relatively novel regression techniques have been used to model the trial-by-trial change in N400 amplitudes that occur over words in a sentence (Payne et al., 2015), where N400 amplitudes decrease as the growing context informs semantic predictions to a greater and greater extent (see also Van Petten & Kutas, 1990a, 1991). For the present investigation, we apply related techniques to the data from the original Lau et al experiment (2013) in order to characterize how exactly the N400 adapted to the predictive validity of the experimental context.

We directly assessed the trial-by-trial adaptation in two ways. First, we fit a local

regression to the N400 amplitudes for related and unrelated words separately over trials for both the high-proportion block and the low-proportion block, in order to investigate how the N400 amplitudes evolved over time. And second, we fit a mixed-effects model including block, condition, and trial in order to test the hypothesis that the learning curve for the semantic priming effect in the low-proportion block differed from the learning curve for the semantic priming effect in the high-proportion block.

We expect that participants should show a period of rapid adaptation when the relatedness proportion shifts from 10% related to 50% related. During this period, participants should begin to learn the new statistical contingencies and update their predictions, gradually and incrementally honing in on the more informative prior.

Methods

The data utilized for the present analysis are described in detail in a previous publication (Lau, Holcomb, et al., 2013). Briefly, each participant saw 80 related and 80 unrelated prime-target pairs that were evenly distributed into a low-proportion block (alongside 280 fillers that were all unrelated) and a high-proportion block (alongside 120 unrelated fillers and 160 related fillers). The low-proportion block was always first, to minimize carry-over effects. These items were drawn from a broader set of 320 associated prime-target pairs, such that every target appeared with a related prime for some participants and an unrelated prime for some other participants, controlling features of the target across participants by design. All related prime-target pairs had at least a 0.5

forward association score from the University of South Florida Association Norms (Nelson et al., 2004). Specific prime-target pairs appeared in both blocks, counterbalanced across participants. No pair was presented more than once to each participant. 40 task-relevant animal probes were also added to each block, and could appear in either the prime or the target position. Stimuli were presented in yellow against a black background with an SOA of 600ms and a target duration of 900ms (see figure 16). Participants were instructed to press a button as quickly as possible when they saw an animal word.

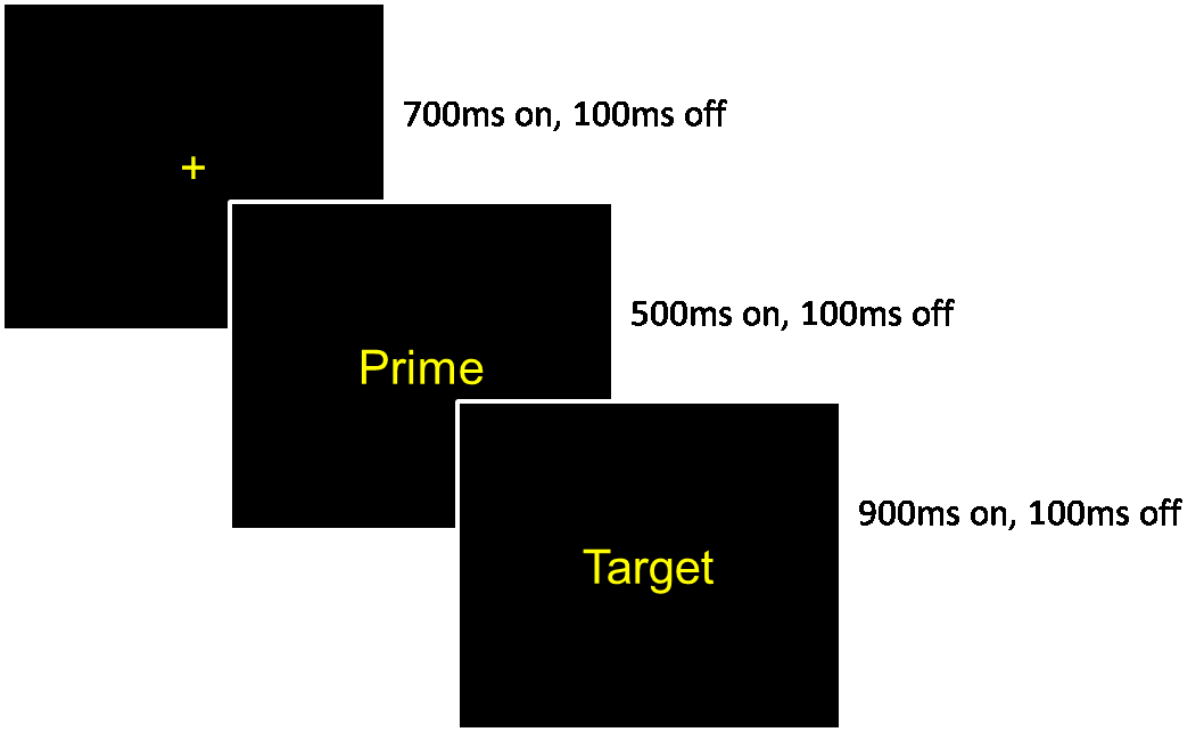

Figure 16: Stimulus presentation.

Data was collected from 32 right-handed participants (13 men) between age 19-24

and no prior history of reading disability or neurological disorder. All participants were native English speakers, who had not learned another language prior to age 5. EEG data from each participant was collected from 29 electrodes distributed across the scalp with a left-mastoid reference. Data was continuously sampled at a rate of 200Hz, with a bandpass filter of 0.01-40Hz. ERPs were time-locked to the presentation of the target and segmented. Trials containing artifacts were rejected (10% in total), and the remaining trials were adjusted to a -100ms-0ms baseline and subjected to an additional 15hz low-pass filter prior to analysis. The mean trial-by-trial amplitudes were extracted for a priori time windows (particularly the N400 time window from 300ms to 500ms), and data from sets of three electrodes at a time were averaged into regions (5 down the anterior-poster midline axis, and four along the periphery of the left and right anterior and posterior portions of the scalp). No further averaging was done beyond this a priori spatio-temporal selection (e.g. individual trials were not averaged into conditions).

Statistical analysis. Initial evidence for a block influence on the trial-by-trial adaptation of the semantic priming effect was obtained using loess local regression. Specifically, single-trial N400 amplitudes were regressed over trials for both related and unrelated targets within each block. Loess regressions are nonparametric, and the shape of the learning curve in each block did not need to be assumed (and instead could be fit naturalistically). We hypothesized that the semantic priming effect should stabilize at a constant, relatively low mean difference in the first block, before rapidly growing over successive trials early in the second (high-proportion) block. Loess regressions used second-order polynomials locally fit by ordinary least-squares over neighborhoods

defined by a tricubic weighting with a span of 0.65. 95% confidence bands were estimated using a smoothed t-based approximation (Wickham, 2009), but are likely slightly anticonservative due to the assumption of local Gaussian error, and are therefore utilized here only for informal inference (e.g. visualization).

Regressions for formal hypothesis testing modeled the "learning curve" for the semantic priming effect, where the semantic priming effect could change over trials (expressed by ordinal position within the experiment). As the exact shape of the learning curve was unknown (but was not likely to be linear), it was approximated using a polynomial, as per the Weierstrass Approximation Theorem (Weierstrass, 1885). This model utilized a large number of parameters and was likely to be overfit, but provided a relative balance between flexibility and interpretability. Overall, Block and Prime were included as fixed categorical factors, and Ordinal Position was included as a fixed third-order polynomial, along with all available interactions. Sets of three electrodes each were averaged into two centro-parietal regions where the N400 is typically maximal (see figure 3), and Region (Central, Parietal) was additionally added to the model, interacting with all terms. The maximal random effects structure was initially included (Barr et al., 2013), but was reduced where the model did not converge or showed evidence of converging at the margins (Bates et al., 2015).

Results

Nonparametric loess local regression models were fit independently for related and unrelated targets in both the high and the low proportion blocks, shown in figure 17. In the low-proportion blocks, related and unrelated words did not yield different N400

amplitudes until a period of time near the middle of the block (roughly between trials 150-300 in a block of 400 trials). In contrast, the related and unrelated words in the high-proportion block quickly diverged during the N400 time window to an increasing extent, showing initial separation before trial 100 and peaking around the middle of the block (with a larger and more sustained peak semantic priming effect than the low-proportion block), before converging again at the very end of the experiment. This pattern of change over trials is consistent with our a priori expectations.

The learning curves were numerically estimated using a polynomial approximation, where a third-order polynomial for Ordinal Position was added to a mixed-effects regression model also containing Block (low proportion, high proportion), Prime (related, unrelated), and Region (Central and Parietal mid-regions, where the N400 effect tends to be maximal). The three-way interaction between Block, Prime, and a linear Ordinal Position term reached significance ($t(4047) = 2.078$, $p = 0.0378$), and did not differ by particular centro-parietal region. A two-way interaction between Ordinal Position and Prime ($t(4002) = -2.361$, $p = 0.0183$) also reached significance, along with a main third-order effect of Ordinal Position ($t(4018) = -2.191$, $p = 0.0285$). With the dramatic loss of power that occurred from approximating the learning curves using a polynomial approximation (where 32 fixed effects are derived), the overall Block by Prime interaction reported by the original study was not significant ($t(3999) = -1.840$, $p = 0.0659$), though we note that in the present analysis, this effect is interpreted marginal to all other (dozens of) terms of the model (whereas the original study expressed the overall main effect averaged over these other influences to amplitude).

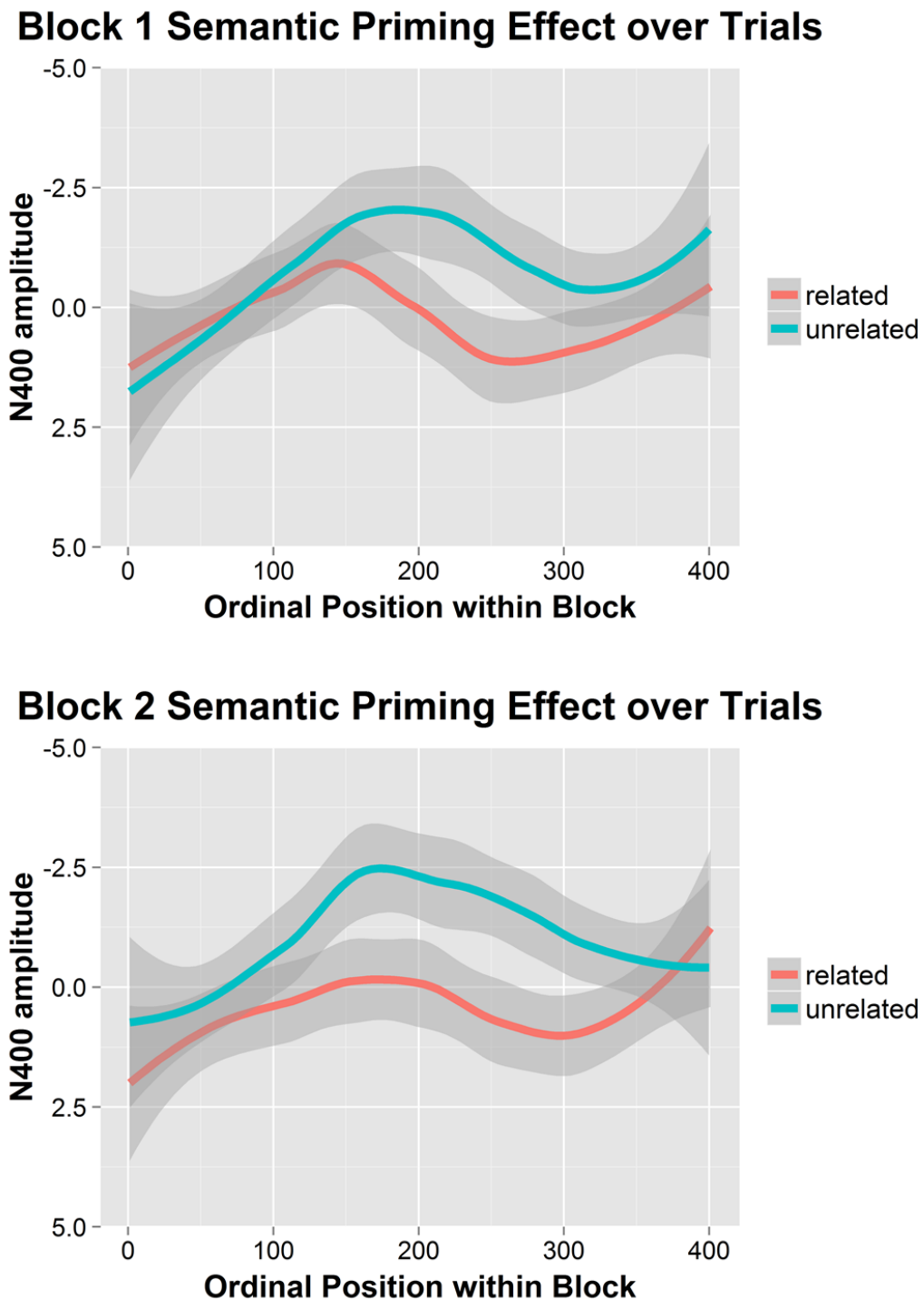

Figure 17: Centro-parietal N400 amplitudes over trials within each block. A semantic priming effect is visible in both blocks after some period of adaptation, but in the high-proportion block, it starts faster, grows bigger, and is sustained longer.

Additionally, we asked whether there was evidence of a reliable change in the priming effect over time that differs between blocks, regardless of the significance of the individual polynomial terms. This allows for the possibility that the overall approximation generally approached a learning curve present in the data regardless of whether specific polynomial terms accounted for significant variation on their own. To test this, we constructed a model where the N400 amplitude for all electrodes was estimated using Block, Prime, and a fifth-order polynomial for Ordinal Position, along with every two-way and three-way interaction (model 1), then compared the overall fit of this model to one where the three-way interactions (and only the three-way interactions) were omitted (model 2). The maximal random effects structure for both models was the same, and included random intercepts for Subject, Item (specifically, the identity of the target), and Channel, by-Channel random slopes for the Prime effect, by-Subject random slopes for the Prime and Block effects, and by-Subject random slopes for the interaction between Prime and Block, with no evidence of convergence at the boundary.

A Chi-squared test comparing the overall fit of these two models indicated that model 1 (the full model) had a significantly better fit than model 2 ($Chi^2(5)=199.8$, $p < 0.001$), providing evidence that the interaction between Block, Prime, and Ordinal Position accounted for significant variation in N400 amplitudes. AIC, BIC, and log-likelihood estimates of the two models all suggested similar improvements in overall fit.

Discussion

Previously, Lau and colleagues demonstrated that the N400 semantic priming effect was sensitive to the predictive validity of the local context (Lau, Holcomb, et al., 2013), which was attributed to adaptation within the blocks. We provide direct evidence for this adaptation with a trial-by-trial re-analysis of the data from Lau et al. In the low proportion block, participants appeared to adapt to the 10% relatedness contingency near the middle of the block, showing a modest semantic priming effect after approximately 150 trials (shown in figure 17). In the high proportion block, participants showed a more rapid adaptation, where the semantic priming effect began earlier (as shown in figure 17), and appeared to reach a larger magnitude by the middle of the block. A regression model including block, condition, and ordinal position indicated that the change in the semantic priming effect over trials was different from the low-proportion block to the high-proportion block.

Together, these data tentatively support our hypothesis that the change in relatedness proportion would result in rapid initial adaptation near the beginning of the high-proportion block. Though block and order were confounded in the original design, this pattern of results is not consistent with a fatigue effect (which would have manifested as a smaller priming effect in block 2, rather than a larger effect), a practice effect (which would have manifested as generally increased semantic priming effects over trials regardless of the location of the proportion manipulation), or a categorical shift in explicit strategy (where a contingency is consciously identified and then utilized at a particular point in time). Instead, our data is consistent with a predictive mechanism that adapts

incrementally and continuously to the statistics of the stimuli (Chang et al., 2012; Fine et al., 2013; Kleinschmidt & Jaeger, 2015; Qian et al., 2012) at the level of semantic features. Specifically, both blocks showed evidence of adaptation *within* the block, but the block with higher predictive validity appeared to show adaptation towards a higher degree of prediction.

The evolution of the absolute amplitudes of the N400 also yielded some interesting results. First, near the beginning of each block (but particularly block 1), the absolute amplitudes for both related and unrelated words (shown in figure 17) appeared to become larger (more negative) over trials. We speculate that participants may have required some time to grow accustomed to the manner of stimulus presentation in order to engage deeply with the semantic content of each stimulus. And second, near the end of block 2 (and to some extent the end of block 1), the semantic priming effect appeared to attenuate. This could indicate that participants adapted to something we did not originally anticipate: the length of the blocks. In other words, our participants appeared to show rapidly diminishing effects when they knew the block (and thus the experiment) was near the end.

Implications for language adaptation. These data carry a number of implications.First, these analyses provide continued support for the consensus interpretation that the size of the N400 semantic priming effect depends in part on the relatively frequency with which related primes and targets occur. But we extend these previous results to directly show how participants learned about local probability over successive trials, which largely drove the semantic priming effects (and their relative size) in both the low-proportion and

the high-proportion blocks.

This adaptive prediction is central to "Bayesian updating" models of language processing (Kleinschmidt & Jaeger, 2015; Kuperberg & Jaeger, In Press). When new information is gained, the comprehender updates their beliefs about the inferred determinants of the speech production (i.e. the generative model of the language being produced). For instance, comprehenders use previous exposure to anticipate the auditory features of speech such as accent (Bradlow & Bent, 2008; Clarke & Garrett, 2004) or speech cadence (Brown, Dilley, & Tanenhaus, 2012), but can also adapt to the structure (Chang et al., 2012; Dell & Chang, 2014; Demberg, Keller, & Koller, 2013; Fine et al., 2013; Jaeger & Snider, 2013) and semantic content (as with the present investigation). Across many levels of language processing, prior beliefs are iteratively updated, incrementally converging on the generative model *explaining* the incoming speech for which prediction error (i.e. Bayesian surprise) is minimal. This adaptation need not be active nor strategic: in such models, language comprehension *is* language learning, where the difference between prior expectations and the actual obtained input (the prediction error) necessarily informs the posterior probability expected for subsequent input (i.e. the new prior).

This sort of adaptive prediction mechanism serves a number of purposes. Prediction allows for accurate communication across a "noisy channel" (e.g. in a loud room or during a windy day) or low stimulus intensities (e.g. at a whisper) and breaks the resource bottleneck that would otherwise occur at word onset (because some of the "work" of comprehension can be accomplished prior to word onset). This sort of

incremental learning not only drives adaptation of existing knowledge in adult speakers (Kleinschmidt & Jaeger, 2015), but also can allow for the initial acquisition of phonemes (N. H. Feldman, Griffiths, Goldwater, & Morgan, 2013), words (Rabovsky & McRae, 2014), and syntax (Dell & Chang, 2014; Fine et al., 2013) in language learners.

However, an adaptive learning mechanism carries risks. Specifically, a prediction mechanism that adapts too quickly is liable to be "overfit", where it is so dependent on the most recent events (as opposed to a longer history of events) that predictions are less successful. For example, a single prediction error could be due to real differences in the context or the speaker etc., or it could be due to the intrinsic variability of speech production (Kleinschmidt & Jaeger, 2015). Similarly, a prediction mechanism that adapts too slowly is also likely to make poor predictions, as prior expectations that are poorly-suited for the context continue to be held despite evidence to the contrary. In other words, an adaptive prediction mechanism is faced with the classic bias-variance trade-off (James et al., 2013). An ideally efficient comprehender is a non-trivial assumption (Hale, 2011): it remains unknown how exactly people balance bias against variance during sentence comprehension.

But one recent proposal helps address the issue of handling prior expectations that are very ill-suited for the context. Specifically, comprehenders might switch *between* different predictive models or priors depending on the circumstances (Kleinschmidt & Jaeger, 2015; Qian et al., 2012). For example, rather than modifying an all-purpose generative model to adapt to the phonemic mappings of a particular speaker with a pronounced accent, a comprehender might switch to different generative model specific

to that speaker or accent. Variance can be minimized by adapting relatively conservatively in most instances, while bias can be minimized by selecting the appropriate model or prior. There is some evidence that this may be the case (Qian et al., 2012). Further, it also may be the case that the likelihood of model-switching given some prediction error might not be fixed. For instance, simultaneous learning of two different artificial grammars is improved when cues such as speaker identity indicate when the "language" is switching (Weiss, Gerfen, & Mitchel, 2009), even if the statistical contingencies of the actual artificial grammars are the same. In general, a rational comprehender may switch generative models when surprisal is very high (Kuperberg & Jaeger, In Press; Qian et al., 2012), adapt slowly to the familiar, and adapt rapidly to the novel (Kleinschmidt & Jaeger, 2015).

However, while most of the interpretation of Lau and colleagues were consistent with our results, we note one possible novel effect. In the original analysis, the proportion manipulation appeared to mainly influence the N400 amplitude of the related targets, and not the unrelated targets (see analysis on page 492 of Lau et al). This was interpreted as an increased semantic facilitation for expected words in the high (versus low) proportion block. Logically, the semantic processing of an incoming word is facilitated if some of the semantic features of that word have already been predicted ahead of time (Federmeier, 2007). This indicated that the consequence of high prediction validity (as in the block with a high proportion of related targets) was more prediction.

But because the distribution of prior probabilities must sum to one, assigning more probability mass to the expected (i.e. related) words also necessarily entails

assigning less probability mass to the unexpected (i.e. unrelated) words. This implies that when predictive validity is high, the facilitation for related targets should also correspond with a very slight *interference* for the processing of nearly all other words, such as the unrelated targets. The present analysis appears to indicate that this may have been the case. As shown in figure 17, the change in the semantic priming effect over trials appeared to be driven by both a smaller N400 amplitude for related words *and* a larger (and more sustained) N400 amplitude for unrelated words. The consequence of predicting the related targets *more* may have been to predict everything else *less,* and this could be reflected in the evolution of N400 amplitudes over time.

This raises the interesting question of whether the processing of the task-relevant *animal* words could have also been slightly impeded in a similar way, as they were also technically unrelated targets (simply task-relevant unrelated words). The original paper provides some indication that this may have been the case. Participants were significantly slower at responding to animal words in the high-proportion block than the low proportion block (Lau, Holcomb, et al., 2013). Further, the animal words, as well as the other task-irrelevant unrelated targets, elicited a larger anterior negativity in the high-proportion block than the low-proportion block, which was interpreted as higher working memory load for words other than the predicted one.

At face value, these data could be interpreted as suggesting that participants adapted in way that made them *worse* at the actual task requested of them. Such a result would hold interesting implications about the nature of what the adaptation accomplishes for the comprehender. But while the possibility is suggestive given the

present data, we caution against drawing firm conclusions at present. The task response (of identifying animal words with a rapid button press) required making a decision about whether each word was animal. Even though the high-proportion block appeared to slow the button-presses for animal words, it's still plausible that the high-proportion block could have also *facilitated* the decision *not* to press the button for the related targets, which showed significant facilitation in the ERPs. Because the related targets did not require a button press, it's not actually known whether adaptation facilitated or impeded the task-imposed *decision* about when to press the button on average across all trials. Future studies are required to specifically address how adaptation to stimulus statistics might relate to task demands. Until then, it remains only an interesting proposition.

Implications for the evolution of N400 amplitudes over the course of an experiment. Finally, the present study qualitatively appeared to yield some apparent trial-by-trial changes to the N400 semantic priming effect beyond the adaptation that was expected to occur to the proportion manipulation. Specifically, as shown in figure 17, the priming effect appeared to diminish rapidly as the experiment neared completion at the end of block 2 (and to some extent, at the end of block 1 as well). There are a number of plausible effects that could have overlapped with the adaptation effect to contribute to this pattern. The primary possibility is simply a fatigue effect, where participants' attention may have started to wane after an extended length of time in the experiment. Sleepy or distracted participants would have yielded generally attenuated effects. Other possible consequences of fatigue include additional noise from alpha activity in the EEG that could obscure the ERP effect of interest, attenuating the semantic priming effect as a

result of diminished power and/or a lower signal to noise ratio. The impact of fatigue on priming was recently illustrated in a meta-analysis of *affective* priming studies (discussed in chapter 3), where the priming effect was larger for studies who broke their stimuli up into more blocks (versus fewer, longer blocks), which allowed participants to take a short break (Herring et al., 2013). Further adaptation studies are required to explore the circumstances under which fatigue effects are observed, and how best to combat them.

However, this observed pattern of block-final effects is also consistent with some possible strategic explanations, and cannot be attributable entirely to fatigue without further study. For example, it may be the case that participants were able to develop expectations about the length of the second block based on their experience with the first block, since the two blocks were identical in length. As such, participant effort could have diminished not necessarily from fatigue alone, but also from the anticipation of finishing. It is not currently known whether the anticipation of task cessation can generally attenuate task performance in this manner (e.g. like a runner slowing down as they approach the finish line), but we speculate that it may be a plausible contributor.

Future directions. The present analysis is just a small step in understanding how language processing adapts over time. The present study suggests that comprehenders may implicitly adapt to predictive validity during a semantic priming paradigm, but a number of open questions remain. It's currently unknown how adaptation might occur to other proportion manipulations (e.g. 50% vs 90% related pairs), where predictions can be made with near-certainty. Disconfirmation of near-certain predictions (such as for unrelated targets in a 90% proportion block) may prompt a unique type of updating, such

as model-switching (Kuperberg & Jaeger, In Press; Qian et al., 2012). It's also currently unknown whether these ERP adaptation effects would also occur for types of priming (e.g. identity priming, affective priming, or simultaneous priming) or at different stimulus onset asynchronies.

But more broadly, there are a wide range of "levels" of language processing beyond word semantics where prediction, and thus adaptation, could also plausibly occur (Kuperberg & Jaeger, In Press), including message-level, pragmatic, syntactic, lexical, orthographic and phonological, and even basic audio-visual information. Both prediction and adaptation within many of these levels remain minimally characterized in the neuroscience literature at present, though preliminary data are promising (Kuperberg & Jaeger, In Press; Lewis & Bastiaansen, 2015).

Further, the present investigation is agnostic to the precise pattern of adaptation that occurred during the experiment. Both of the analytic techniques utilized here are not meant to offer direct evidence for any particular quantification of learning or adaptation. The LOESS regression does not provide interpretable quantitative information about how the N400 semantic priming effect changed over trials, as the model is not described by a singular equation. And the polynomial approximation used for the mixed models was intended to encompass many potential learning curves (along with other overlapping effects of time, including fatigue and electrode drift), with coefficients that are not likely to (and not intended to) generalize to other experiments. Though the LOESS regression qualitatively suggests that the block transition was followed by a period of rapid learning (where the related and unrelated targets yielded N400 amplitudes that increasingly

diverged), there is no applicable decision criteria with which to discern different rates of learning.

However, the N400 semantic priming effect at any particular trial should be a direct consequence, in part, of the information yielded by the preceding trials. Here, we utilized ordinal position within a block to reflect the amount of available information (about the statistics of the context) that participants had been exposed to. But the actual priming effect at a particular point in time is not literally a consequence of ordinal position; it is instead the consequence of the learning that actually occurred on a trial-by-trial basis. What the preceding trials actually are (and in what order they are in) matters. Future studies should move from investigating adaptation as the *average* effect that occurs at a certain *position* within a block (as with the present study) to investigating the effect for *particular stimuli* that occur given the *specific information* that has been provided to participants during the previous trials.

In addition, while conducting our investigation, we encountered a number of challenges that future ERP studies will need to address. Foremost among these is that, unlike the previous behavioral studies, the effect of interest in an ERP measure is embedded within trial-level noise (i.e. amplitude variation unrelated to stimulus processing) that is roughly an order of magnitude larger. Though aggressive signal extraction procedures could be used to improve the signal-to-noise ratio, conducting appropriate inferential procedures on such heavily manipulated outcomes may be difficult. Importantly, increasing the number of trials does not necessarily improve statistical power in an adaptation paradigm as it would other analyses using trial-level

data (see chapter I). For example, in the present investigation, increasing the number of trials within each block does not actually add additional measurements of the period where the actual learning occurs. Instead, the added trials simply extend the end of the blocks, where the adaptation between trials is expected to gradually converge to nothing in most adaptive models (Kleinschmidt & Jaeger, 2015). However, increasing the number of participants does, in fact, increase the number of measurements during the period of maximal adaptation. This suggests that effective design of adaptation experiments may include relatively short paradigms with a relatively large number of subjects (see also Kurumada, Brown, & Tanenhaus, 2012; Yildirim, Degen, Tanenhaus, & Jaeger, 2016).

Second, we noted above that trial-by-trial ERP amplitudes may include multiple overlapping processes, including fatigue and strategic effects that may have intermixed with the adaptation effect. Such complex data require significant model flexibility in order to account for the potential evolution of effects through time. We approximated the present adaptation using third-order polynomials, but the use of polynomial approximations can rapidly reduce power and make comprehension difficult. We suggest that generalized additive models may be useful for identifying the signal for some experiments, to capture the trial-by-trial change marginal to some *a priori* learning effect.

Conclusions. In sum, the present investigation was intended to explore the adaptation over trials hypothesized to contribute to the between-block effect observed in the previous Lau et al study (2013). Using novel ERP analytic techniques, we found that both blocks showed evidence that the N400 semantic priming effect grew in size as participants learned about the local context (figure 17). As hypothesized by Lau and

colleagues, the period of maximum adaptation appeared to occur following the onset of the high-proportion block, when predictive validity shifted from low to high. This is consistent with several theories of language adaptation, where the brain works to minimize prediction error over successive iterations (Kleinschmidt & Jaeger, 2015). We also present a number of additional findings about how the amplitude of the N400 component evolved over the course of the experiment, including possible fatigue and strategic effects (where participants may have anticipated the end of the second block based on their experience with the length of the first block). Overall, our investigation is broadly consistent with an adaptive probabilistic prediction mechanism that works during language comprehension to passively minimize semantic prediction error over time.

CHAPTER V:

CONCLUDING REMARKS

The meaning of a word is more than definition. The comprehension of the words "love" or "death" entails a deep understanding of the social and emotional implications, as wells as contextual and motivational significance. In this dissertation, we explored the neuroscience of how emotional content, local context, and task demands influence word processing. In chapter I, we described analytic approaches to these questions that model word processing as the convergence of the particulars of the item, the comprehender, and context. In chapter II, we applied these techniques to investigate how the semantic processing of infrequent words is facilitated if that word is also highly emotional, regardless of whether attention is oriented towards semantic features or emotional features. In chapter III, we presented two experiments that investigated the role of valence in semantic processing by implementing a full cross of semantic priming and affective priming, finding radically different patterns of effects that suggest the two may rely on distinct mechanisms. And in chapter IV, we investigated how the semantic priming effect adapts to the local context as participants implicitly learn the statistical contingencies, using a novel trial-by-trial adaptation analysis that shows the evolution of the semantic priming effect through time. Overall, these studies suggest that semantic processing is broadly supported by the rich experiences and patterns available to the comprehender, including knowledge of emotional significance and implicit contextual expectations. Language comprehension is as richly textured as the comprehenders.

Chapter II presented an investigation that was intended to determine whether the N400 word frequency effect might be attenuated by emotion. We found evidence that this was the case. The processing of infrequent words was facilitated if that word was highly arousing. Notably, this avalanche effect seemed confined to the earliest portions of the N400 time window, where frequency effects first reliably onset (Laszlo & Federmeier, 2014). These data provide further evidence that emotion can attenuate the influence of word frequency during initial semantic processing. But we extend these previous findings in a number of important ways. First, we found a continuous and progressive influence of emotion on the N400 frequency effect. Second, we implicate arousal specifically as the primary contributor to the avalanche effect. And finally, we extend the evidence for the avalanche effect to two novel tasks. But contrary to our hypothesis, we did not find that the particular task appeared to matter. We suggest that it may be advantageous in some circumstances to maintain a vigilance (or expectation) for highly emotionally significant stimuli (including words), particularly if that stimulus tends to be rare.

Chapter III presented a pair of experiments that were designed to assess the manner in which affective priming and semantic priming relate to one another. As anticipated, a robust N400 Association effect was observed for both neutral and emotional words. Across the two studies, this effect appeared in the a priori spatio-temporal ROI to a significance of eight sigma. In contrast, no signs of affective priming were present for either the N400 component or the LPC component in either experiment. Bayes factors indicated that the affective priming manipulation was 4-6 times more probable (given the data) to not influence the N400 and LPC components than it was to

modulate them. This held a number of implications for the possible mechanisms driving ERP affective priming effects.

While the primary focus of these experiments was to shed further insight on the nature of the affective priming effect, these data also hold some compelling implications for the N400 semantic priming effect. Like many other influences on lexico-semantic processing, the semantic priming effect was found to be sensitive to task demands, with larger effects observed for the semantic task than the valence task. Further, the present study also addresses the question of whether emotional prime-target pairs might yield larger N400 semantic priming effects than neutral prime-target pairs. We found that this was the case, regardless of task.

And finally, chapter IV presented an investigation of how N400 semantic priming effects adapt to the features of the local context over time. Previously, Lau and colleagues demonstrated that the N400 semantic priming effect was sensitive to the predictive validity of the local context (Lau, Holcomb, et al., 2013), which was attributed to adaptation within the blocks. We provide direct evidence for this adaptation with a trial-by-trial re-analysis of the data from Lau et al. Together, these data tentatively support the hypothesis that the change in relatedness proportion resulted in rapid initial adaptation near the beginning of the high-proportion block. These data carry a number of implications. First, these analyses provide continued support for Lau et al's interpretation that the size of the N400 semantic priming effect depends in part on the relatively frequency with which related primes and targets occur. This adaptive prediction is central to "Bayesian updating" models of language processing (Kleinschmidt & Jaeger, 2015;

Kuperberg & Jaeger, In Press). Consistent with this model, however, we find novel preliminary evidence that unrelated words may have been more difficult to process after adapting to the high-proportion block compared to the low-proportion block. The present analysis is just the small step in understanding how language processing adapts over time. But more broadly, there are a wide range of "levels" of language processing beyond word semantics where prediction, and thus adaptation, could also plausibly occur (Kuperberg & Jaeger, In Press), including message-level, pragmatic, syntactic, lexical, orthographic and phonological, and even basic audio-visual information.

The evolution of the absolute amplitudes of the N400 also yielded some interesting results. First, near the beginning of each block (but particularly block 1), the absolute amplitudes for both related and unrelated words appeared to become larger (more negative) over trials. And second, near the end of block 2 (and to some extent the end of block 1), the semantic priming effect appeared to attenuate. This could be due to learning and fatigue effects (also discussed in chapter I). However, this observed pattern of block-final effects is also consistent with some possible strategic explanations, and cannot be attributable entirely to fatigue without further study. Specifically, we speculate that this convergence could be due to an expectation for the end of the experiment, as the end of the second block could be predicted based on participants' experience with the first block.

Overall, these studies suggest that semantic processing is broadly supported by the rich experiences and patterns available to the comprehender, including knowledge of emotional significance and implicit contextual expectations. Comprehension through

definition is as limited as paint by numbers. Language comprehension is as richly textured as the comprehenders.

APPENDIX A

OVERVIEW OF SIMPLE REGRESSION

The construction of a linear regression model should be intuitive for most researchers. Variance in some predictor(s), such as word frequency, is use to model the variance in some outcome, such as N400 amplitudes. In the case of a single predictor (plus an intercept) and a single outcome, the model is simply a line in the form $y \sim bx + i + e$, where any particular outcome y is the sum of the intercept i, the predictor x multiplied by some effect b, and the residual error e. Given a new x (e.g. the frequency of a novel word), we can use this line to predict what we expect y (e.g. N400 amplitude) to be, on average. Because of the additive nature of the model, the intercept i is simply the value of y when x equals 0. Easily interpretable intercepts can be obtained by centering the predictor(s) (i.e. subtracting the mean), such that the mean of x is zero. Because regression lines pass through the point mean(x), mean(y), the value of the intercept is known when mean(x)=0. Specifically, the intercept i is simply mean(y), or the average outcome. Centering predictors in this way does not change the estimation of b.

The regression coefficient b is determined through some process of fitting a model (in this case, a line) to relate the change in x to the change in y (Hastie et al., 2009). In the ordinary least squares (OLS) approach, the regression line will minimize the sum of the squared residuals (i.e. the overall squared distance between the predicted y-value on the regression line and the actually observed y at that point). This procedure assumes that the residual error e is normally distributed. For large data sets where both predictors and outcomes are normally distributed and the model is correctly specified, OLS is generally pretty reliable and unbiased. But in instances of missing data or more

modest samples sizes, OLS tends to be less reliable, in part because the normality assumption may not be exactly met.

One alternative to OLS is to find whichever model maximizes the likelihood that the observed data could have been obtained. In other words, which regression coefficient b yields the largest likelihood of the data given the model containing that parameter b? This is called Maximum Likelihood Estimation (MLE). If all of the assumptions of OLS are exactly met, it can be proved that the OLS model is actually also the model with the maximum likelihood (Hald, 1999): OLS regression is simply a specific case of MLE. Because MLE naturally yields the OLS model in cases where the OLS is licensed, and otherwise yields more robustly reliable models for most real experiments, MLE is typically used in psycholinguistics and other fields, and a restricted MLE (REML) is the default for the lme4 program in R.

Often, regression models will include more than one predictor. These are all fit simultaneously - there is not "order" to deriving the parameters in a MLE procedure. Where the regression model for a single predictor is simply a line, the regression model for two parameters is a plane: the value of y will depend on both the value of $x1$ and the value of $x2$. In short, the regression model describes a hyperplane across the parameters. Each parameter reflects the relationship between a change in x and the corresponding change in y holding all else in the model equal.

The parameters indicate how the outcome changes as a function of the predictors. They can be interpreted as the "slope" of the hyperplane drawn through the feature space.

But importantly, this slope will necessarily depend on both the units of x and the units of y. For instance, in modeling the price of diamonds based on their carats, the slope will be very high, as a very small change in carats will correspond with a very large increase in price. If we expressed the mass of diamonds in milligrams instead of carats (where 1 carat = 200mg), the slope would be two hundred times smaller, and if we expressed the value of diamonds in cents rather than dollars, the slope would be a hundred times larger. However, the actual relationship between mass and value is fixed across all of these different methods of measurement, even while the metric regression coefficients increase or decrease with the particular units of measurement. Because the metric regression coefficient can be thought of in units y/x (to reflect the unit change y expected per unit change in$x$), a unitless, standardized coefficient can be obtained by multiplying b by sd(x)/sd(y), canceling these units out and scaling x and y to equivalently-spread normal distributions. This standardized regression coefficient β, or "Beta", will be the same for any equivalent measure of mass and price. This allows effects of various predictors on the same outcome to be compared to one another directly.

The standardized regression coefficient instead reflects in part the strength of the relationship between the predictor and the outcome. If x and y correspond perfectly and exactly, where any 1SD change in x corresponds with a 1SD change in y, β would be 1. In the case where x and y are completely unrelated, β would be zero, as the MLE best-fit line should not estimate different y values for different values of x. The similarity between standardized regression coefficients and correlation coefficients is not accidental. In fact, for a simple OLS regression with one predictor, β for x will actually

be identical to Pearson's correlation coefficient for x and y. For multiple regression, the standardized parameter estimate is slightly different from the partial correlations between x and y, but still generally indicates the marginal impact of x on y.

A framework for working with regression coefficients of ERP data has recently been developed by Smith and Kutas (Smith & Kutas, 2015a, 2015b).

Bibliography

Abelson, R. P., & Sermat, V. (1962). Multidimensional scaling of facial expressions. *Journal of experimental psychology, 63*(6), 546-554. doi: 10.1037/h0042280

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods, 16*(2), 270-301. doi: 10.1177/1094428112470848

Alnosaier, W. S. (2007). *Kenward-Roger approximate F test for fixed effects in mixed linear models.* (PhD Thesis), Oregon State University. Retrieved from https://ir.library.oregonstate.edu/xmlui/handle/1957/5262

Aquino, J. M., & Arnell, K. M. (2007). Attention and the processing of emotional words: Dissociating effects of arousal. *Psychonomic bulletin & review, 14*(3), 430-435. doi: 10.3758/BF03194084

Arnell, K. M., Killman, K. V., & Fijavz, D. (2007). Blinded by emotion: target misses follow attention capture by arousing distractors in RSVP. *Emotion, 7*(3), 465-477. doi: 10.1037/1528-3542.7.3.465

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

Baayen, H., Tremblay, A., & Hendrix, P. (2010a). Exploring Linguistic Components of Evoked Response Potentials with Generalized Additive Models (Gams). *Psychophysiology, 47*, S16-S16.

Baayen, H., Tremblay, A., & Hendrix, P. (2010b). Modeling linguistic components of evoked response potentials with generalized additive models. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale, 64*(4), 297-297.

Baayen, H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and language, 81*(1-3), 55-65. doi: DOI 10.1006/brln.2001.2506

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *J Exp Psychol Hum Percept Perform, 10*(3), 340-357.

Balota, D. A., & Chumbley, J. I. (1990). Where are the effects of frequency in visual word recognition tasks? Right where we said they were! Comment on Monsell, Doyle, and Haggard (1989). *J Exp Psychol Gen, 119*(2), 231-237.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459. doi: 10.3758/BF03193014

Barachant, A., & Cycon, R. (2015). Grasp-and-lift-EEG-challenge (GitHub repository). https://github.com/alexandrebarachant/Grasp-and-lift-EEG-challenge

Bargh, J. (1997). The automaticity of everyday life. In R. Wyer (Ed.), *The automaticity of everyday life: Advances in social cognition* (Vol. 10, pp. 1-61). Mahwah, NJ: Lawrence Erlbaum.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang, 68*(3). doi: 10.1016/j.jml.2012.11.001

Bates, D. M. (2010). lme4: Mixed-effects modeling with R. *URL http://lme4. r-forge. r-project. org/book*.

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. Retrieved from: http://cran.r-project.org/package=lme4

Bayer, M., Sommer, W., & Schacht, A. (2012). P1 and beyond: Functional separation of multiple emotion effects in word recognition. *Psychophysiology, 49*(7), 959-969. doi: 10.1111/j.1469-8986.2012.01381.x

Becker, C. A. (1980). Semantic Context Effects in Visual Word Recognition - an Analysis of Semantic Strategies. *Memory & cognition, 8*(6), 493-512. doi: Doi 10.3758/Bf03213769

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical neurophysiology, 60*(4), 343-355.

Blackford, T., Holcomb, P. J., Grainger, J., & Kuperberg, G. R. (2012). A funny thing happened on the way to articulation: N400 attenuation despite behavioral interference in picture naming. *Cognition, 123*(1), 84-99. doi: 10.1016/j.cognition.2011.12.007

Blais, C., Stefanidi, A., & Brewer, G. A. (2014). The Gratton effect remains after controlling for contingencies and stimulus repetitions. *Frontiers in psychology, 5*, 1-11. doi: ARTN 1207

10.3389/fpsyg.2014.01207

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research, 2*(1).

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley.

Bradley, M. M. (1994). Emotional memory: A dimensional analysis (pp. 97-134).

Bradley, M. M. (2000). Emotion and Motivation. In J. T. Cacioppo, L. G. Tassinary, & G. G. Bernston (Eds.), *Handbook of Psychophysiology* (Vol. 1, pp. 602-642).

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings (Vol. C-1). Gainesville, FL. The Center for Research in Psychophysiology, University of Florida.

Bradley, M. M., & Lang, P. J. (2007). Emotion and Motivation. In J. T. Cacioppo, L. G. Tassinary, & G. G. Bernston (Eds.), *Handbook of Psychophysiology* (Vol. 3rd, pp. 581).

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729. doi: DOI 10.1016/j.cognition.2007.04.005

Briesemeister, B. B., Kuchinke, L., & Jacobs, A. M. (2011). Discrete emotion effects on lexical decision response times. *PLoS ONE, 6*(8), e23743. doi: 10.1371/journal.pone.0023743

Briesemeister, B. B., Kuchinke, L., & Jacobs, A. M. (2014). Emotion word recognition: discrete information effects first, continuous later? *Brain research, 1564*, 62-71. doi: 10.1016/j.brainres.2014.03.045

Brouillet, T., & Syssau, A. (2005). Connection between the evaluation of positive or negative valence and verbal responses to a lexical decision making task. *Canadian journal of experimental psychology, 59*(4), 255-261.

Brown, M., Dilley, L., & Tanenhaus, M. (2012). *Real-time expectations based on context speech rate can cause words to appear or disappear*. Paper presented at the 34th annual conference of the Cognitive Science Society.

Carretie, L. (2014). Exogenous (automatic) attention to emotional stimuli: a review. *Cognitive, affective & behavioral neuroscience, 14*(4), 1228-1258. doi: 10.3758/s13415-014-0270-2

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics …*. doi: 10.1002/lnc3.337

Chawla, A., Maiti, T., & Sinha, S. Kenward-Roger approximation for linear mixed models with missing covariates (D. o. S. a. Probability, Trans.): Michigan State University.

Chumbley, J. I., & Balota, D. A. (1984). A word's meaning affects the decision in lexical decision. *Mem Cognit, 12*(6), 590-606.

Citron, F. M. (2012). Neural correlates of written emotion word processing: a review of recent electrophysiological and hemodynamic neuroimaging studies. *Brain and language, 122*(3), 211-226. doi: 10.1016/j.bandl.2011.12.007

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci, 36*(3), 181-204. doi: 10.1017/S0140525X12000477

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of verbal learning and verbal behavior, 12*(4), 335-359.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*(6), 3647-3658. doi: 10.1121/1.1815131

De Houwer, J. (2014). A propositional perspective on context effects in human associative learning. *Behav Processes, 104*, 20-25. doi: 10.1016/j.beproc.2014.02.002

Delaney-Busch, N. (2013). *The processing of emotional features in single and primed words.* (Master's of Science), Tufts University.

Delaney-Busch, N., & Kuperberg, G. R. (2013). Friendly drug-dealers and terrifying puppies: affective primacy can attenuate the N400 effect in emotional discourse contexts. *Cognitive, affective & behavioral neuroscience, 13*(3), 473-490. doi: 10.3758/s13415-013-0159-5

Delaney-Busch, N., Wilkie, G., & Kuperberg, G. (*under review*). Vivid: How valence and arousal influence word processing under different task demands.

Dell, G. S., & Chang, F. (2014). The P-chain: relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B-Biological Sciences, 369*(1634). doi: ARTN 20120394

10.1098/rstb.2012.0394

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience, 8*(8), 1117-1121. doi: 10.1038/nn1504

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210. doi: 10.1016/j.cognition.2008.07.008

Demberg, V., Keller, F., & Koller, A. (2013). Incremental, Predictive Parsing with Psycholinguistically Motivated Tree-Adjoining Grammar. *Computational Linguistics, 39*(4), 1025-1066. doi: DOI 10.1162/COLI_a_00160

den Heyer, K., Briand, K., & Dannenbring, G. L. (1983). Strategic factors in a lexical-decision task: evidence for automatic and attention-driven processes. *Mem Cognit, 11*(4), 374-381.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion, 15*(2), 115-141. doi: 10.1080/0269993004200024

Fazio, R. H., Sanbonmatsu, D. M., Powell, M. C., & Kardes, F. R. (1986). On the automatic activation of attitudes. *Journal of personality and social psychology, 50*(2), 229-238.
Federmeier, K. D. (2007). Thinking ahead: the role and roots of prediction in language comprehension. *Psychophysiology, 44*(4), 491-505. doi: 10.1111/j.1469-8986.2007.00531.x
Federmeier, K. D., & Kutas, M. (1999). A Rose by Any Other Name: Long-Term Memory Structure and Sentence Processing. *Journal of Memory and Language, 41*(4), 469-495. doi: 10.1006/jmla.1999.2660
Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in human neuroscience, 4*, 215. doi: 10.3389/fnhum.2010.00215
Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A Role for the Developing Lexicon in Phonetic Category Acquisition. *Psychological review, 120*(4), 751-778. doi: 10.1037/a0034245
Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS ONE, 8*(10). doi: ARTN e77661

10.1371/journal.pone.0077661
Fischer-Baum, S., Dickson, D. S., & Federmeier, K. D. (2014). Frequency and regularity effects in reading are task dependent: Evidence from ERPs. *Lang Cogn Neurosci, 29*(10), 1342-1355. doi: 10.1080/23273798.2014.927067
Fischler, I., & Bradley, M. M. (2006). Event-related potential studies of language and emotion: words, phrases, and task effects. *Progress in brain research, 156*, 185-203. doi: 10.1016/S0079-6123(06)56009-1
Fogel, A., Midgley, K., Delaney-Busch, N., & Holcomb, P. J. (2013). *Processing emotion and tabooness in a native vs. a second language: an ERP study.* Paper presented at the 20th Annual Meeting of the Cognitive Neuroscience Society.
Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6*(4), 187-214. doi: 10.1080/17588928.2015.1020053
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. doi: Doi 10.1037//0096-3445.121.4.480
Groot, A. M. B. (1984). Primed lexical decision: Combined effects of the proportion of related prime-target pairs and the stimulus-onset asynchrony of prime and target. *The Quarterly Journal of Experimental Psychology, 36*(2), 253-280. doi: 10.1080/14640748408402158
Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*(12), 1711-1725. doi: 10.1111/j.1469-8986.2011.01273.x
Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology, 48*(12), 1726-1737. doi: 10.1111/j.1469-8986.2011.01272.x
Guillet, R., & Arndt, J. (2009). Taboo words: the effect of emotion on memory for peripheral information. *Mem Cognit, 37*(6), 866-879. doi: 10.3758/MC.37.6.866

Hajcak, G., MacNamara, A., Foti, D., Ferri, J., & Keil, A. (2013). The dynamic allocation of attention to emotion: simultaneous and independent evidence from the late positive potential and steady state visual evoked potentials. *Biol Psychol, 92*(3), 447-455. doi: 10.1016/j.biopsycho.2011.11.012

Hajcak, G., MacNamara, A., & Olvet, D. M. (2010). Event-related potentials, emotion, and emotion regulation: an integrative review. *Developmental neuropsychology, 35*(2), 129-155. doi: 10.1080/87565640903526504

Hajcak, G., Weinberg, A., MacNamara, A., & Foti, D. (2012). ERPs and the study of emotion. In S. J. Luck & E. S. Kappenman (Eds.), *Oxford Handbook of ERP components*. New York: Oxford University Press.

Hald, A. (1999). On the history of maximum likelihood in relation to inverse probability and least squares. 214-222. doi: 10.1214/ss/1009212248

Hale, J. T. (2011). What a Rational Parser Would Do. *Cogn Sci, 35*(3), 399-443. doi: 10.1111/j.1551-6709.2010.01145.x

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models - The R Package pbkrtest. *Journal of Statistical Software, 59*(9), 1-30.

Hänze, M., & Hesse, F. W. (1993). Emotional influences on semantic priming. *Cognition & Emotion, 7*(2), 195205. doi: 10.1080/02699939308409184

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association, 72*(358), 320-338. doi: 10.2307/2286796

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer New York.

Hauk, O., Davis, M. H., Ford, M., Pulvermuller, F., & Marslen-Wilson, W. D. (2006). The time course of visual word recognition as revealed by linear regression analysis of ERP data. *NeuroImage, 30*(4), 1383-1400. doi: 10.1016/j.neuroimage.2005.11.048

Hauk, O., & Pulvermuller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clin Neurophysiol, 115*(5), 1090-1103. doi: 10.1016/j.clinph.2003.12.020

Hauk, O., Pulvermuller, F., Ford, M., Marslen-Wilson, W. D., & Davis, M. H. (2009). Can I have a quick word? Early electrophysiological manifestations of psycholinguistic processes revealed by event-related regression analysis of the EEG. *Biol Psychol, 80*(1), 64-74. doi: 10.1016/j.biopsycho.2008.04.015

Herbert, C., Kissler, J., Junghöfer, M., Peyk, P., & Rockstroh, B. (2006). Processing of emotional adjectives: Evidence from startle EMG and ERPs. *Psychophysiology, 43*(2), 197-206. doi: 10.1111/j.1469-8986.2006.00385.x

Hermans, D., De Houwer, J., & Eelen, P. (1994). The affective priming effect: Automatic activation of evaluative information in memory. *Cognition & Emotion, 8*(6), 515 <last_page> 533. doi: 10.1080/02699939408408957

Hermans, D., Spruyt, A., & Eelen, P. (2003). Automatic affective priming of recently acquired stimulus valence: Priming at SOA 300 but not at SOA 1000. *Cognition & Emotion, 17*(1), 83 <last_page> 99. doi: 10.1080/02699930302276

Herring, D. R., Taylor, J. H., White, K. R., & Crites Jr, S. L. (2011). Electrophysiological responses to evaluative priming: the LPP is sensitive to incongruity. *Emotion, 11*(4), 794-806. doi: 10.1037/a0022804

Herring, D. R., White, K. R., Jabeen, L. N., Hinojos, M., Terrazas, G., Reyes, S. M., . . . Crites Jr, S. L. (2013). On the Automatic Activation of Attitudes: A Quarter Century of Evaluative Priming Research. *Psychological Bulletin*. doi: 10.1037/a0031309

Hinojosa, J. A., Carretié, L., Mendez-Bertolo, C., Miguez, A., & Pozo, M. A. (2009). Arousal contributions to affective priming: electrophysiological correlates. *Emotion (Washington, D.C.), 9*(2), 164-171. doi: 10.1037/a0014680

Hofmann, M. J., Kuchinke, L., Tamm, S., Võ, M. L., & Jacobs, A. M. (2009). Affective processing within 1/10th of a second: High arousal is necessary for early facilitative processing of negative but not positive words. *Cognitive, affective & behavioral neuroscience, 9*(4), 389-397. doi: 10.3758/9.4.389

Holcomb, P. J. (1988). Automatic and attentional processing: an event-related brain potential analysis of semantic priming. *Brain Lang, 35*(1), 66-85.

Holt, D. J., Lynn, S. K., & Kuperberg, G. R. (2009). Neurophysiological correlates of comprehending emotional meaning in context. *Journal of cognitive neuroscience, 21*(11), 2245-2262. doi: 10.1162/jocn.2008.21151

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: effects of word frequency. *Percept Psychophys, 40*(6), 431-439.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59*(4), 434-446. doi: 10.1016/j.jml.2007.11.007

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*(1), 57-83. doi: 10.1016/j.cognition.2012.10.013

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (4th ed.). New York: Springer.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *Journal of the American Statistical Association, 93*(442), 720-729. doi: Doi 10.2307/2670122

Kahan, T. A., & Hely, C. D. (2008). The role of valence and frequency in the emotional Stroop task. *Psychonomic bulletin & review, 15*(5), 956-960. doi: 10.3758/PBR.15.5.956

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*(3), 983-997.

Kissler, J., Assadollahi, R., & Herbert, C. (2006). Emotional and semantic networks in visual word processing: insights from ERP studies. *Progress in brain research, 156*, 147-183. doi: 10.1016/S0079-6123(06)56008-X

Klauer, K. C., & Musch, J. (2001). Does sunshine prime loyal? Affective priming in the naming task. *The Quarterly journal of experimental psychology.A, Human experimental psychology, 54*(3), 727-751.

Klauer, K. C., & Musch, J. (2003). Affective Priming: Findings and Theories. In J. Musch & K. C. Klauer (Eds.), *The Psychology of Evaluation: Affective Processes in Cognition and Emotion* (Vol. 1, pp. 9-33): Psychology Press.

Klauer, K. C., & Stern, E. (1992). How attitudes guide memory-based judgments: A two-process model. *Journal of experimental social psychology, 28*(2), 186-206. doi: 10.1016/0022-1031(92)90038-L

Kleinschmidt, D. F., & Jaeger, F. T. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review, 122*(2), 148. doi: 10.1037/a0038695

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of experimental psychology.Learning, memory, and cognition, 20*(4), 804-823. doi: 10.1037/0278-7393.20.4.804

Kousta, S. T., Vinson, D. P., & Vigliocco, G. (2009). Emotion words, regardless of polarity, have a processing advantage over neutral words. *Cognition, 112*(3), 473-481. doi: 10.1016/j.cognition.2009.06.007

Kreher, D. A., Holcomb, P. J., & Kuperberg, G. R. (2006). An electrophysiological investigation of indirect semantic priming. *Psychophysiology, 43*(6), 550-563. doi: 10.1111/j.1469-8986.2006.00460.x

Kuchinke, L., Vo, M. L., Hofmann, M., & Jacobs, A. M. (2007). Pupillary responses during lexical decisions vary with word frequency but not emotional valence. *Int J Psychophysiol, 65*(2), 132-140. doi: 10.1016/j.ijpsycho.2007.04.004

Kunst-Wilson, W. R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. *Science (New York, N.Y.), 207*(4430), 557-558.

Kuperberg, G. R. (2013). The proactive comprehender: What event-related potentials tell us about the dynamics of reading comprehension. In B. Miller, L. Cutting, & P. McCardle (Eds.), *Unraveling the Behavioral, Neurobiological, and Genetic Components of Reading Comprehension* (pp. 176-192). Baltimore: Paul Brookes Publishing.

Kuperberg, G. R., & Jaeger, F. T. (In Press). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). *Pragmatic interpretation of contrastive prosody: It looks like speech adaptation.* Paper presented at the Proceedings of the 34th annual conference of the Cognitive Science Society.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621-647. doi: 10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science (New York, N.Y.), 207*(4427), 203-205. doi: 10.1126/science.7350657

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161-163. doi: 10.1038/307161a0

Kutas, M., Van Petten, C., & Kluender, R. (2006). Psycholinguistics electrified II: 1994-2005. In M. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (Vol. 2nd Edition, pp. 659). New York, NY: Elsevier.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmerTest'. from https://cran.r-project.org/web/packages/lmerTest/

Lai, V. T., Hagoort, P., & Casasanto, D. (2012). Affective Primacy vs. Cognitive Primacy: Dissolving the Debate. *Frontiers in psychology, 3*, 243. doi: 10.3389/fpsyg.2012.00243

Lang, P. J., & Bradley, M. M. (2009). Emotion and the motivational brain. *Biological psychology*. doi: 10.1016/j.biopsycho.2009.10.007

Laszlo, S., & Federmeier, K. D. (2010). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*. doi: 10.1111/j.1469-8986.2010.01058.x

Laszlo, S., & Federmeier, K. D. (2014). Never seem to find the time: evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience, 29*(5), 642-661. doi: 10.1080/01690965.2013.866259

Lau, E. F., Gramfort, A., Hamalainen, M. S., & Kuperberg, G. R. (2013). Automatic semantic facilitation in anterior temporal cortex revealed through multimodal neuroimaging. *J Neurosci, 33*(43), 17174-17181. doi: 10.1523/JNEUROSCI.1018-13.2013

Lau, E. F., Holcomb, P. J., & Kuperberg, G. R. (2013). Dissociating N400 effects of prediction from association in single-word contexts. *J Cogn Neurosci, 25*(3), 484-502. doi: 10.1162/jocn_a_00328

Lebrecht, S., Bar, M., Barrett, L. F., & Tarr, M. J. (2012). Micro-valences: perceiving affective valence in everyday objects. *Frontiers in psychology, 3*, 107. doi: 10.3389/fpsyg.2012.00107

LeDoux, J. E. (2012). Evolution of human emotion: a view through fear. *Progress in brain research, 195*, 431-442. doi: 10.1016/B978-0-444-53860-4.00021-0

Leuthold, H., Kunkel, A., Mackenzie, I. G., & Filik, R. (2015). Online processing of moral transgressions: ERP evidence for spontaneous evaluation. *Soc Cogn Affect Neurosci, 10*(8), 1021-1029. doi: 10.1093/scan/nsu151

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177. doi: 10.1016/j.cognition.2007.05.006

Lewis, A. G., & Bastiaansen, M. (2015). A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex, 68*, 155-168. doi: 10.1016/j.cortex.2015.02.014

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410-423. doi: Doi 10.1198/016214507000001337

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic bulletin & review, 7*(4), 618-630. doi: Doi 10.3758/Bf03212999

Luck, S. J. (2014). An Introduction to the Event-Related Potential Technique Preface. *Introduction to the Event-Related Potential Technique, 2nd Edition*, Vii-+.

Marschark, M. (1983). Semantic Congruity in Symbolic Comparisons - Salience, Expectancy, and Associative Priming. *Memory & cognition, 11*(2), 192-199. doi: Doi 10.3758/Bf03213474

Mather, M., & Sutherland, M. R. (2011). Arousal-biased competition in perception and memory. *Perspectives on psychological science, 6*(2), 114-133. doi: 10.1177/1745691611400234

Maxfield, L. (1997). Attention and semantic priming: a review of prime task effects. *Consciousness and cognition, 6*(2-3), 204-218. doi: 10.1006/ccog.1997.0311

Medler, D. A., & Binder, J. R. (2005). [MCWord: An On-Line Orthographic Database of the English Language]. Web Page.

Mendez-Bertolo, C., Pozo, M. A., & Hinojosa, J. A. (2011). Word frequency modulates the processing of emotional words: convergent behavioral and electrophysiological data. *Neuroscience letters, 494*(3), 250-254. doi: 10.1016/j.neulet.2011.03.026

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in Recognizing Pairs of Words - Evidence of a Dependence between Retrieval Operations. *Journal of experimental psychology, 90*(2), 227-&. doi: DOI 10.1037/h0031564

Monahan, J. L., Murphy, S. T., & Zajonc, R. B. (2000). Subliminal mere exposure: specific, general, and diffuse effects. *Psychological Science, 11*(6), 462-466.

Morris, J. P., Squires, N. K., Taber, C. S., & Lodge, M. (2003). Activation of Political Attitudes: A Psychophysiological Examination of the Hot Cognition Hypothesis. *Political Psychology, 24*(4), 727-745. doi: 10.1046/j.1467-9221.2003.00349.x

Morris, J. S. (2002). The BLUPs are not "best" when it comes to bootstrapping. *Statistics and Probability Letters, 56*(4), 425-430. doi: 10.1016/S0167-7152(02)00041-X

Neely, J. H. (1976). Semantic priming and retrieval from lexical memory: Evidence for facilitatory and inhibitory processes. *Memory & cognition, 4*(5), 648-654. doi: 10.3758/BF03213230

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes in reading: Visual word recognition, 11*, 264-336.

Neely, J. H., Keefe, D. E., & Ross, K. L. (1989). Semantic priming in the lexical decision task: roles of prospective prime-generated expectancies and retrospective semantic matching. *J Exp Psychol Learn Mem Cogn, 15*(6), 1003-1019.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc, 36*(3), 402-407.

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychol Rev, 113*(2), 327-357. doi: 10.1037/0033-295X.113.2.327

Okon-Singer, H., Lichtenstein-Vidne, L., & Cohen, N. (2013). Dynamic modulation of emotional processing. *Biological psychology, 92*(3), 480-491. doi: 10.1016/j.biopsycho.2012.05.010

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1967). *The Measurement of Meaning*: University of Illinois Press.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika, 58*(3), 545-554. doi: 10.1093/biomet/58.3.545

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52*(11), 1456-1469. doi: 10.1111/psyp.12515

Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychol Res, 66*(3), 180-194. doi: 10.1007/s00426-002-0086-5

Qian, T., Jaeger, T. F., & Aslin, R. N. (2012). Learning to represent a multi-context environment: more than detecting changes. *Frontiers in psychology*. doi: 10.3389/fpsyg.2012.00228

R Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to Deal with “The Language-as-Fixed-Effect Fallacy”: Common Misconceptions and Alternative Solutions. *Journal of Memory and Language, 41*(3), 416-426. doi: http://dx.doi.org/10.1006/jmla.1999.2650

Rabovsky, M., & McRae, K. (2014). Simulating the N400 ERP component as semantic network error: Insights from a feature-based connectionist attractor model of word meaning. *Cognition, 132*(1), 68-89. doi: 10.1016/j.cognition.2014.03.010

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: effects of word frequency, verb complexity, and lexical ambiguity. *Mem Cognit, 14*(3), 191-201.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science, 6*(1), 15-32.

Rotteveel, M., de Groot, P., Geutskens, A., & Phaf, R. H. (2001). Stronger suboptimal than optimal affective priming? *Emotion (Washington, D.C.), 1*(4), 348-364.

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review, 16*(2), 225-237. doi: Doi 10.3758/Pbr.16.2.225

Rugg, M. D. (1985). The effects of semantic priming and work repetition on event-related potentials. *Psychophysiology, 22*(6), 642-647.

Rugg, M. D. (1990). Event-related brain potentials dissociate repetition effects of high- and low-frequency words. *Mem Cognit, 18*(4), 367-379.

Russell, J. A. (1980). A Cicumplex Model of Affect. *Journal of personality and social psychology, 39*(6), 1161-1178. doi: 10.1037/h0077714

Samsonovic, A. V., & Ascoli, G. A. (2010). Principal Semantic Components of Language and the Measurement of Meaning. *PLoS ONE, 5*(6), e10921. doi: 10.1371/journal.pone.0010921

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin, 2*(6), 110-114. doi: 10.2307/3002019

Schupp, H. T., Ohman, A., Junghöfer, M., Weike, A. I., Stockburger, J., & Hamm, A. O. (2004). The facilitated processing of threatening faces: an ERP analysis. *Emotion (Washington, D.C.), 4*(2), 189-200. doi: 10.1037/1528-3542.4.2.189

Scott, G. G., O'Donnell, P. J., Leuthold, H., & Sereno, S. C. (2009). Early emotion word processing: evidence from event-related potentials. *Biological psychology, 80*(1), 95-104. doi: 10.1016/j.biopsycho.2008.03.010

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre-and postlexical loci of contextual effects on word recognition. *Memory & cognition*.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science, 25*(3), 289-310. doi: Doi 10.1214/10-Sts330

Smith, N. J., & Kutas, M. (2015a). Regression-based estimation of ERP waveforms: I. The rERP framework. *Psychophysiology, 52*(2), 157-168. doi: DOI 10.1111/psyp.12317

Smith, N. J., & Kutas, M. (2015b). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology, 52*(2), 169-181. doi: DOI 10.1111/psyp.12320

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319. doi: DOI 10.1016/j.cognition.2013.02.013

Snow, N., & Neely, J. H. (1987). Reduction of Semantic Priming from Inclusion of Physically or Nominally Related Prime-Target Pairs. *Bulletin of the Psychonomic Society, 25*(5), 335-335.

Solomon, R. L., & Howes, D. H. (1951). Word frequency, personal values, and visual duration thresholds. *Psychol Rev, 58*(4), 256-270.

Spruyt, A., De Houwer, J., & Hermans, D. (2009). Modulation of automatic semantic priming by feature-specific attention allocation. *Journal of Memory and Language, 61*(1), 37-54. doi: 10.1016/j.jml.2009.03.004

Spruyt, A., Hermans, D., De Houwer, J., Vandromme, H., & Eelen, P. (2007). On the nature of the affective priming effect: effects of stimulus onset asynchrony and congruency proportion in naming and evaluative categorization. *Memory & cognition, 35*(1), 95-106.

Storbeck, J., & Clore, G. L. (2007). On the interdependence of cognition and emotion. *Cognition & Emotion, 21*(6), 1212-1237. doi: 10.1080/02699930701438020

Storbeck, J., & Robinson, M. D. (2004). Preferences and inferences in encoding visual objects: a systematic comparison of semantic and affective priming. *Personality & social psychology bulletin, 30*(1), 81-93. doi: 10.1177/0146167203258855

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology, 18*(6), 643 <last_page> 662. doi: 10.1037/h0054651

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Thode, H. (2002). *Testing for Normality*: CRC Press.

Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological science : a journal of the American Psychological Society / APS, 20*(9), 1092-1099. doi: 10.1111/j.1467-9280.2009.02411.x

van Casteren, M., & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods, 38*(4), 584-589.

Van Petten, C., & Kutas, M. (1990a). Interactions between Sentence Context and Word-Frequency in Event-Related Brain Potentials. *Memory & cognition, 18*(4), 380-393.

Van Petten, C., & Kutas, M. (1990b). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & cognition, 18*(4), 380-393. doi: 10.3758/BF03197127

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Mem Cognit, 19*(1), 95-112.

Vinson, D., Ponari, M., & Vigliocco, G. (2014). How does emotional content affect lexical processing? *Cogn Emot, 28*(4), 737-746. doi: 10.1080/02699931.2013.851068

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191-1207. doi: 10.3758/s13428-012-0314-x

Weierstrass, K. (1885). Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen. *Sitzungsberichte der Akademie zu Berlin, 2*, 633-639 and 789-805.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). SPEECH SEGMENTATION IN A SIMULATED BILINGUAL ENVIRONMENT: A CHALLENGE FOR STATISTICAL LEARNING? *Language learning and development : the official journal of the Society for Language Development, 5*(1), 30-49. doi: 10.1080/15475440802340101

Welch, B. L. (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika, 34*(1/2), 28-35. doi: 10.2307/2332510

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis.*: Springer New York.

Wittenbrink, B. (2007). Measuring Attitudes Through Priming. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit Measures of Attitudes* (Vol. 1st, pp. 17-58): Guilford Press.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language, 72*, 37-48. doi: 10.1016/j.jml.2013.12.003

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2013). Linguistic variability and adaptation in quantifier meanings. *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society*, 3835-3840.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language, 87*, 128-143. doi: http://dx.doi.org/10.1016/j.jml.2015.08.003

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American psychologist, 35*(2), 151.

Zhang, Q., Kong, L., & Jiang, Y. (2012). The interaction of arousal and valence in affective priming: behavioral and electrophysiological evidence. *Brain research, 1474*, 60-72. doi: 10.1016/j.brainres.2012.07.023

Zhang, Q., Lawson, A., Guo, C., & Jiang, Y. (2006). Electrophysiological correlates of visual affective priming. *Brain research bulletin, 71*(1-3), 316-323. doi: 10.1016/j.brainresbull.2006.09.023

Zhang, Q., Li, X., Gold, B. T., & Jiang, Y. (2010). Neural correlates of cross-domain affective priming. *Brain research, 1329*, 142-151. doi: 10.1016/j.brainres.2010.03.021