Running head: GENETIC PROGRAMMING TO MODEL

NUANCE 3.0: Using Genetic Programming to Model Variable Relationships

Geoff Hollis & Chris F. Westbury

Department of Psychology, University of Alberta

P220 Biological Sciences Building

T6G 2E9 Edmonton, Alberta, Canada

Jordan B. Peterson

Department of Psychology, University of Toronto

Sidney Smith Hall, 4th Floor

100 St. George St. M5S 3G3 Toronto, Ontario, Canada

Correspondence on this submission should be addressed to Chris Westbury at the above address or:

Tel: 780-492-8518, Fax: 780-492-1768

E-mail: chrisw@ualberta.ca

Acknowledgements: This work was made possible by a National Engineering And Science Research Council grant from the Government of Canada to Chris Westbury.

Abstract

Previously, we introduced a new computational tool for fitting nonlinear data, and data exploration: The Naturalistic University of Alberta Nonlinear Correlation Explorer (NUANCE) (Hollis & Westbury, in press). When introduced, we demonstrated that NUANCE was capable of providing useful descriptions of data for two toy problems. Since then, we have extended the functionality of NUANCE in a new release (NUANCE 3.0) and fruitfully applied to tool to real psychological problems. Here, we discuss the results of two studies carried out with the aid of NUANCE 3.0. We demonstrate that NUANCE can be a useful tool to aid research in psychology in at least two ways: it is capable of highlighting useful knowledge that might be overlooked by more traditional analytical, factorial approaches. Second, it can be harnessed to simplify complex models of human behavior.

NUANCE 3.0: Using Genetic Programming to Model Variable Relationships

Introduction

Genetic Programming (GP) is a paradigm for automating the process of computer programming. It works in a fashion analogous to selective breeding in biology. The user provides two elements: an operational definition of the goal, and a set of operators and operands that can be used to achieve that goal. In selective breeding, the goal may be coming up with a smaller dog or a cow that produces more milk. In GP, the goal may be developing virtual creatures that maximizes how much virtual food it collects over a finite time interval or discovering an equation that minimizes the error between an actual value and a prediction of it. The important point in all these cases is that the *fitness* of a candidate solution is quantifiable. As long as the dog or error is getting smaller, a solution is getting better.

In selective breeding, the operators are genetically specified and are often (until recently, always) only implicit from the breeder's point of view. A dog breeder can create a smaller dog by selective breeding without ever knowing which genes his directed mating is affecting. Relatives of GP like Genetic Algorithms (Holland, 1992) are analogous, because they use arcane problem-specific binary representations for a solution. However, in GP, the operators are explicitly specified, consisting of well-defined, general computational operations such as addition, subtraction, square root, and log.

When the goal and operators are defined, GP proceeds by creating a large set of computer programs (agents) that combine the operators in random ways. Each agent in the population attempts to solve the problem. The agents that perform best at this task are selected out and mated (duplicated, broken apart and recombined with each other in

random ways) to form new agents. These new agents and the best agents from the previous generation are used to create a new population pool. This process – test, select, and mate – is repeated until a completion criterion specified by the user is met (e.g. until a certain amount of time or number of generations has elapsed, or until one agent gets close enough to the goal).

Recently, we have developed the Naturalistic University of Alberta Nonlinear Correlation Explorer (NUANCE 2.0) (Hollis & Westbury, in press). NUANCE is a platform-independent program written in Java that uses Genetic Programming to model nonlinear variable relationships. When NUANCE was first introduced, it was demonstrated to work on two toy problems. The focus of the following studies is to demonstrate NUANCE can be applied to real problems in psychology. In this paper we introduce and use a new version of the program, NUANCE 3.0. This tool and a manual that includes a description of new features added since the previous version are available as a free download from the Psychonomic Society archive at http://psychonomic.org/archive. Our aim in the present paper is to demonstrate that NUANCE can be applied to real psychological problems, revealing new findings with both utilitarian and theoretical value.

Study 1

The ability to predict human performance can be useful in applied psychology. For instance, Peterson (2005) looked at the domain of the pharmacist. Pharmacists will make errors in the prescriptions they give to customers. As an example of how common these errors are, Peterson cites a survey that found 34% of Texan pharmacists to have an error rate greater than one prescription error per week. Of course, this is extremely undesirable because peoples' health depends on the accuracy of such prescriptions. Most attempts to correct these problems have focused on refining the process of dispensing prescriptions in general, by using better labeling of pharmaceutical products and developing methods to

automate the process. Very little attention has been put towards studying how individual differences among pharmacists might relate to prescription errors. Such research may reveal new methods for dealing with substandard job performance.

Peterson undertook such a study to discover how individual differences might play a role in errors for dispensing drug prescriptions. They administered pharmacists a battery of cognitive tests sensitive to frontal lobe functioning, assessing decision-making, error-monitoring, planning, problem-framing, and novelty-analysis. Using these measures, they were able to correctly classify 77.4% of pharmacists as having been or not having been reprimanded for making prescription errors. Such information is useful, because it may suggest methods for identifying pharmacists at risk for making errors, as well as intervention methods to reduce their error rates.

We wanted to see if the results of Peterson could be improved upon with NUANCE. We were interested in two ways of improving the previous results: increasing the classification accuracy, and developing a simpler classification strategy. The classifications of Peterson were from a logistic regression analysis incorporating performance ratings on seven different tasks that relate to prefrontal cognitive ability. Can we use NUANCE to develop a classification strategy that incorporates fewer parameters? Finding a simpler strategy should make identifying and dealing with pharmacists at risk for making errors more feasible in practice.

Method

Stimuli

The data used by Peterson were used during this study. However, two of the original entries were removed due to missing values. This left us with performance measures for 60 pharmacists across 7 cognitive tasks, 19 of whom had been reprimanded for mis-prescriptions, and 41 who who had not. Data were turned into z-scores before

being used. To prevent the uneven group sizes from allowing base rates to influence how classifications are made, we broke these data into two sets. The first contained the 19 reprimanded pharmacists and 19 randomly selected, unreprimanded pharmacists. The second set contained entries for the remaining 22 pharmacists who were unreprimanded. The first set was used as a *training set* - we supplied it to NUANCE for building a classification model. The second set was used as a *validation set*. After NUANCE had created a classification model, we tested the model on the validation set to ensure it generalized to unseen data.

Procedure

NUANCE was run with default settings on the *training set*, with the exception of 3 parameters: parsimony pressure, minimum constant, and maximum constant.

Parsimony pressure is a parameter that addresses one of GP's major limitations: the fact that functions may get so large that they are either completely incomprehensible, intractable to run, or a combination of the two. The parsimony pressure parameter imposes a user-specifiable fitness penalty on large solutions, which is a percent value equal to the parsimony pressure times the number of nodes (operators and arguments) in the function (Hollis & Westbury, in press). The default parsimony pressure is a 0.2% reduction in fitness for every node in a classification tree. For this problem, parsimony pressure was increased to 1.5%. This was done to encourage the development of simple models.

The minimum and maximum constants are parameters that allow the user to constrain any randomly-generated variables used in evolved equations to a specified range. The default range of constants NUANCE allows for is 0 to 1 which, through division, can emulate all real numbers greater than zero. For this problem, the minimum and maximum constants were set to -3 and 3 respectively, as this was roughly the range over which our predictors varied (-2.7 to 3).

The operator set is the set of operators allowed within evolved functions. For this problem the operator set was limited to the equals operator and the less than operator. This was motivated by the comment that "a linear combination of prefrontal tests was able to correctly classify approximately $\frac{3}{4}$ of pharmacists into the correct groups, reprimanded and unreprimanded. Such classification was particularly accurate in the case of the unreprimanded group, suggesting that executive or prefrontal function scores above a particular cut-off are very infrequently associated with serious performance error" (Peterson, 2005, p. 20). Because of the infrequency of association between high prefrontal functioning scores and performance error, the implication appears to be we can derive a very simple and accurate model of performance by classifying pharmacists based on whether they fall above or below a threshold on some combination of prefrontal functioning tests.

Results

The best equation evolved by NUANCE correctly classified 76% of the pharmacists in the training set (58% accuracy on reprimanded pharmacists and 95% accuracy on unreprimanded pharmacists). It classified 92% of the 22 unreprimanded pharmacists composing the validation set. The pooled accuracy was 82%, which is a 5% improvement in accuracy over the classifier developed by Peterson.

This difference was not statistically significant ($p = 0.66$, by Fisher's exact test). Both classification models performed equally well. However, the solution found by NUANCE was much simpler than the previous solution: after simplifying the model created by NUANCE by removing tautological and contradictory statements, we were left with a single conditional statement incorporating the result of a single test. The final model was:

$$group = \begin{cases} reprimanded & \text{Random letter span task} < \text{-1.26 z-scores} \\ unreprimanded & \text{otherwise} \end{cases}$$

A person's standardized score on a random letter span task can classify pharmacists as well as a strategy using a linear combination of all seven cognitive performance tasks outlined by Peterson can. The random letter span task is described as follows:

> *"For each trial, the participant is asked to input a random sequence of letters for a given letter span (for example, they were asked to 'randomize the letters from L to O' (i.e., a four letter span)). The participant enters letters by using the computer monitor and the mouse. The monitor shows one of the letters from the letter span. If the participant moves the mouse (in any direction) the letter will change. As the participant continues to move the mouse, new letters are displayed. While the mouse is moving, the display will cycle forward through all the letters in the span, beginning again at the first letter after the last letter has been displayed. To select a letter, the participant clicks the mouse button while the letter is on the screen. If the participant produces an acceptable sequence, he is asked to randomize a span one letter longer than the previous span. If the participant makes an error (omission or patterned sequence (e.g. 'L, M, N')), he is given a second chance to randomize a span of the same length. Before beginning the task proper, the participant must successfully randomize two four letter span practice trials. After this, the participant begins the scored trials with a letter span of four. The task terminates when the participant fails to randomize a given length span two trials in a row, or when he correctly completes all trials (span length 4-14)"* (Peterson, 2005, pp. 12-13).

Discussion

Although NUANCE was not able to improve on raw accuracy of classification, it did produce a highly simplified classification model, which is a great improvement practically speaking. Peterson suggested practicing pharmacists be screened with a battery of tests measuring prefrontal cognitive ability. The model derived by NUANCE suggests that results from a single test may be sufficient for accurate screening. The fact that a very specific type of task can predict reprimanded/unreprimanded status of pharmacists so well also gives us insight as to why these pharmacists tend to make errors. The random letter span task taxes working memory, and requires a modest amount of planning based on the contents of working memory. Dispensing errors on a pharmacist's part may be due to a below average capacity of one or both of these faculties. Intervention for reprimanded pharmacists may want to focus on honing a pharmacist's capacity for these aspects of cognition, or on using external aids (such as written notes) for holding information rather than relying on working memory.

When NUANCE was introduced, it was framed as a tool for modeling nonlinear variable relationships (Hollis & Westbury, in press). The results of this study suggest it is not limited to this single type of task. It is demonstrably suitable for simplifying preexisting models. Users have a great deal of control over how important parsimonious solutions are by manipulating the parameters of NUANCE at runtime. In addition to providing us with a great deal of predictive power, NUANCE can reduce complex models to a level of simplicity that give the models practical utility.

Study 2

Lexical access is a complex process influenced by many factors. To complicate matters, these factors will often contribute to the process in complex ways, and interact with other factors in equally complex ways. Empirical research on lexical access typically

follows an analytic approach, with factorial manipulation as the main tool of choice for understanding how factors contribute to the process of lexical access. This approach has almost singlehandedly taken research on lexical access (and psychology in general) to its current standing. But it is not without its shortcomings.

The approach of studying effects through factorial manipulation has flaws when used to study lexical access. Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2005) provide a rigorous coverage of five reasons why factorial manipulation is a limited technique. Briefly, the points are:

– it is difficult to find stimuli that vary only along one categorical dimension

– researchers may have implicit knowledge that biases item selection

– stimuli sets often contain words from opposite ends of a dimension of interest. In some cases, this may change a participants' sensitivity to the factors of interest

– Most variables we study are continuous. Treating them as categorical effects in factorial manipulations decreases reliability and statistical power

– we run into problems concerning whether significant effects are a reflection of lexical processing in general, or an artifact of the selected stimuli. In some cases, it may be hard to differentiate because of point 1.

We would like to elucidate a sixth reason why studying effects factorially should be expected to gloss over critical information. Analysis tools such as ANOVA treat independent variables as if their underlying relationship is *linear* with dependent variables. However, this is a gross oversimplification of things. For example, R. Baayen (2005) examined the relationship between lexical decision reaction times (LDRTs) and 13 predictors. 11 predictors had significant relationships with LDRTs. Of the 11, 6 had nonlinear relationships with LDRTs. Furthermore, 4 of these relationships were nonmonotonic. Nonlinearity is potentially interesting information glossed over by factorial manipulation, and nonmonotonicity is *completely missed*.

Nonlinear relations between stimulus and action (in our case, word properties and lexical decision reaction time) are a *fundamental requirement* for behavior that is sufficiently complex to be worth psychological scrutiny. The reason for this is illustrated by the history of the artificial neural network. Neural networks are biologically inspired computational devices. They have a set of n *nodes* (analagous to neurons) that accept input from an environment or other nodes. When sent input, nodes transform their signal with an *activation function*, multiply it by a weight, and produce the new value as output. Nodes will often be chained together, with one node's output being the input for many other nodes. Through the process of sending signals through a network of such nodes, many psychologically interesting computations can be performed.

Until the late 60's, a specific class of neural networks received a great deal of interest from psychologists: perceptrons. Perceptrons have two input nodes chained to a third, output node. Banks of perceptrons can do tasks like pattern recognition and classification. However, Marvin Minsky and Seymour Papert proved that traditional perceptrons are unable to solve a certain class of problems: linearly nonseparable problems (Minsky & Papert, 1968). This proof rendered perceptrons uninteresting in the context of complex psychological behavior.

Since Minsky and Papert's proof it has been realized that, although perceptrons are of limited interest to psychology, neural networks in general are powerful enough to offer insights to psychology. Whole perceptrons can be chained together to provide more complex behavior. However, this is contingent on the nodes in each perceptron having *nonlinear activation functions*. Chains of perceptrons with nodes only employing linear activation functions can be reduced to a single bank of perceptrons (Dawson, 2004, pp. 170–173) and, thus, uninteresting by Minsky and Papert's proof.

The lesson is that computational power does not necessarily increase with structural complexity in systems that only perform linear transformations on their inputs. If a

system is to be psychologically interesting - if it is to be more than merely the *sum* of its environment - the system *requires* nonlinear dynamics. As such, psychologists need to pay attention to nonlinearity to get a complete grasp on psychologically interesting behavior. Furthermore, the specific *shapes* of nonlinear relationships are equally important. Minsky and Papert's demonstration that perceptrons are unable to solve linearly nonseparable problems is not true when nonmonotonic activation functions (such as a Gaussian activation function) are used (Dawson & Schopflocher, 1992).

Factorial manipulation does not adequately capture these formal constraints on complex systems. The analytic approach, which is often coupled with factorial manipulation in psychological research, is not without its own shortcomings. This approach- and Popper's hypotheticodeductive approach to science more generally- is theory-driven (Popper, 1959). Research is conducted either to compare the merits of one or more theories, or because a theory makes an unexpected prediction and we are interested in verifying it. The Popperian approach to science is not without its detractors (Neisser, 1997; Feyerabend, 1975). One problem with adopting a strict hypotheticodeductive approach to science is that many topics of psychological scrutiny are complex, with many interacting forces directing how they work. This makes building a complete theory of a psychological topic through a strictly analytical approach difficult. We simply do not have the disposition for thinking in terms of complex, nonlinear interactions. Eventually, we will have to incorporate new methods of analysis into our research programmes.

Van Orden and Paap (1997) give an account of human behavior that – if true – is even more worrisome for investigators who rely on analytic, factorial methods to study lexical access. Their argument suggests that reductive (analytic) approaches to psychology will eventually need to be replaced because human behavior has *reciprocal causality*. To quote, "reciprocal causality implies that each and every component of a system

contributes to every behavior of the whole system" (Van Orden & Paap, 1997, pp. S92). When a system is reciprocally causal, the functioning of its components is context-dependent, and those components are highly interactive. Reciprocal causality calls into question the applicability of an analytic, reductive approach to studying human behavior. Context-dependence implies that a static explanation of the system in question (what a reductive approach aims to provide) will miss critical details. Interactivity also questions the applicability of an analytic approach. Such an approach assumes that the system under question can be broken down to basic components which are the core of what functionally matters. This is at odds with what we would expect in a highly interactive system. In an interactive system, we would more likely expect that individual components mean very little compared to the coordination of those components. Isolating a component may not yield any useful information, since it will ultimately be how that component relates to every other component that matters.

In the current research, we take a synthetic approach to studying lexical access, focusing on mathematical modeling rather than factorial manipulation. At the most, our aim is to demonstrate how this approach can reveal useful information that would otherwise be overlooked by an analytic, factorial approach to studying psychological phenomena. At the very least, we hope to demonstrate that a synthetic approach can aid and inform an analytic approach.

We modeled the relationship between 16 variables of potential relevance to the process of lexical access, and the behavioral measure of lexical decision reaction time (LDRT). LDRTs are a commonly-used measure of how long it takes subjects to decide whether a presented string is a legal word. We studied the individual effects of each variable on LDRTs, as well as all pairwise interaction between our sixteen predictors. This gave us a grand total of 136 "experiments" in this study – an thorough search of the research space that analytic researchers have been exploring experimentally for decades.

We performed this research without entertaining any specific hypotheses. Instead, we are interested in what our search tells us is important.

Method

Stimuli

One advantage of the synthetic approach we have adopted in this study is that it is possible to study many more stimuli than one could realistically use in a single experiment. We used behavioral measures taken from the English Lexicon Project (Balota et al., 2002) – an online database of over 40,000 words and behavioral data collected on participants' response capacities for the words. A total of 4,778 words were used in this study. For a word to be included in the study, it needed to meet the criteria of being 4-6 letters long and have an entry in each of the three repositories from which we drew our predictor and dependent variable values, which are described below.

Predictors

Sixteen predictors were used in this study. Both on an orthographic and phonological dimension, measures of frequency, neighbourhood size, average neighbourhood frequency, positionally controlled bigram/biphone frequencies, and uncontrolled bigram/biphone frequencies were included. Also included were the first and last trigram frequencies for the words. Estimates of these values were retrieved from the Wordmine database (Buchanan & Westbury, 2000) or calculated directly by us using the CELEX database (R. H. Baayen, Piepenbrock, & Gulikers, 1995). In addition to these fourteen predictors, two predictors derived from word co-occurrence frequencies were used: the number of semantic neighbours and average radius of co-occurence (Shaoul & Westbury, 2005). A brief description of each predictor is provided in Table 1.

Procedure

The procedure for this study has two parts. In the first part, each of the predictors are taken alone and used to model LDRTs for half of the 4,778 words- the *training set*- with the other half being defined as the *validation set*. Modeling was performed with NUANCE 3.0. In this portion of the study we had two goals: to understand how much variance in LDRTs these predictors can account for, and to discover and test hypotheses about the shape of the relationship between these predictors and LDRTs.

In the second part of this study, we were interested in studying how well the *interaction* between any two predictors accounts for variance in LDRTs, and in understanding the nature of these interactions. To do this, our 16 predictors are taken two at a time and used to predict LDRTs as in the first portion of the study.

Achieving our goal in the second half of the study is not as straightforward as in the first part. This is because best fit functions provided by NUANCE may contain effects attributable to the individual predictors, in addition to effects attributable their interactions. Decomposing output into its contributing parts can be extremely difficult, as it is not always obvious where the interactions are and where the main effects are in the complex functions provided by NUANCE. We worked around this problem by performing two multiple regressions with the output of the functions supplied by NUANCE, for each predictor pair. The first regression only contains terms for the functions derived when each variable was run alone, as follows:

$$LDRT = \beta_0 + \beta_1 f_1(a) + \beta_2 f_2(b) + error$$

The second regression contains terms for these same functions, plus the function derived in this portion of the study, when both predictors were used together to predict LDRTs, as follows:

$$LDRT = \beta_0 + \beta_1 f_1(a) + \beta_2 f_2(b) + \beta_3 f_3(a, b) + error$$

By subtracting the variance accounted for by the first regression equation from the

variance accounted for by the second regression equation, we can obtain an estimate for how strong the interaction between any two predictors is.

This method is not without its flaws. There is no guarantee that some *better fit* for each predictor is not embedded within the interaction function of any two predictors: that is, no guarantee that some of the variance our method attributes to the interaction should not properly be attributed to one or the other of the predictors. Insofar as this is the case, our method will incorrectly attribute too much accounting for variance to the pair's interaction. However, no better option for deducing the strength of any predictor pair's interaction presents itself. Decomposing each pair-wise equation by hand is impractical, given how many variable pairs we have and how complex the interactions may be.

The best we can do is try and scour the search space as thoroughly as possible, giving us the maximal probability that we have the most predictive functions in both the individual and pair-wise cases. We did this by running NUANCE on each problem 20 times and taking the best fit across all runs.

Results

The amount of variance in LDRTs accounted for by each individual predictor is displayed in Table 2. All significant interactions are displayed in Table 3. All reported values are from performance on the 2,389 item *validation set*, which NUANCE was not exposed to while modeling LDRTs. With these data, we address three questions: "Which predictors account for the most variance?", "Which predictor are the most interactive?", and "What is the shape of the relationships between predictors and LDRTs?"

Which variables account for the most variance?

It should be noted that the summed variance accounted for by all of the predictors when run individually (Table 2) exceeds 1. This reflects the fact that there is a great deal of overlap between our predictors insofar as how they relate to the process of lexical

access. For instance, we should expect phonological and orthographic frequency to be related to LDRTs in roughly the same manner, since they are strongly correlated ($R = 0.70$ across all 4,778 words; $p < 0.001$). In order to understand which predictors account for unique variance, we performed a linear stepwise, backwards regression on LDRTs with the NUANCE-derived functions of our sixteen predictors as terms in the regression equation. The *validation set* was used to perform the regression. The predictors left in after the backwards, stepwise regression are presented in Table 5. The predictors removed during the model simplification included PFREQ, CONBP, PN, PNFREQ, and ONFREQ – mostly phonological variables whose orthographic counterparts remained in the model.

Of the remaining 11 predictors, the 4 that account for the most variance in LDRTs (OFREQ, LETTERS, ON, and LASTTRI) combine to account for 41% of the total variance in LDRTs. This is 96% of the variance accounted for by all sixteen predictors together. Frequency, length, orthographic neighbourhood size, and body frequency (which is approximated by LASTTRI) are all well-studied variables in lexical access. It did not come as a surprise that orthographic frequency accounts for far more of the variance in LDRTs than any other predictor used in this study. Frequency is well-known to be an important factor in just about every psychological task, including lexical access. What may come as a surprise is how much variance in LDRTs accounted for can be found in only four variables.

Which predictors are most interactive?

As stated earlier, we know language processing is a complex task involving many factors that can interact with each other in complex ways. We will not be able to completely understand the mechanics of language processing in terms of single causes (Van Orden & Paap, 1997); to understand the mechanics of language processing, we will have to understand how different pieces of a language processing system interact with

each another.

We can get an estimate of how *interactive* a variable is by summing across the R^2 values for all significant interactions the predictor is involved in with all other predictors in the study, which are presented in Table 3). The results of this summing are presented in Table 4.

The four predictors whose interactions accounted for the most variance were PFREQ, ONFREQ, UNBP, and LETTERS. Even though ONFREQ was pushed out of the linear stepwise backwards regression of the solitary variables, it was the second most interactive variable out of the sixteen we considered. UNBP was the third most interactive variable, but accounted for the third least amount of variance in LDRTs by itself. These findings suggest there may be some factors in lexical access that make little or no individual contribution to lexical access but are, instead, purely mediating factors.

It is conceivable that the reason why ONFREQ appears so interactive is because of its similarity to ON, which is itself the fifth most interactive variable. The correlation between the best-fit transformations for ON and ONFREQ for predicting LDRTs is very high ($R = 0.74$, $df = 2387$, $p < 0.0001$). Further evidence of their relation is provided by the fact that ON remained in the stepwise backward regression and ONFREQ did not. ONFREQ may simply be getting at the same aspects of lexical access as ON is. However, if we look at the significant interactions, we have good reason to suspect this is not the case. ONFREQ has significant interactions with 5 variables (PFREQ, UNBP, PN, PHONEMES, LETTERS), while ON has interactions with just two variables (PFREQ, OFREQ). There is only one variable with which both ON and ONFREQ interact PFREQ. The *interaction between* ONFREQ and ON was marginally reliable in context of our adjusted α of $\frac{0.05}{120}$ (R = 0.06, df = 2387, p = 0.003). For these reasons, the two variables do not seem to be getting at the same relationships, and ONFREQ appears to be a strictly mediating factor with no individual contribution to lexical access.

Another striking result is that interactions with phonological frequency account for approximately three times more variance in LDRTs than interactions with its orthographic counterpart (Table 3). When the two variables are looked at alone, the ratio flips: phonological frequency accounts for approximately 2.5 times *less* variance in LDRTs than orthographic frequency (Table 2). This does not run counter to the general knowledge that frequency mediates almost every other effect in lexical decision tasks (Cutler, 1981), but it does add an extra layer of complexity to this fact.

What is the shape of the relationships between predictors and LDRTs?

A general pattern stands out between the relationship of frequency variables and LDRTs. The best fit for all frequency measures (excluding uncontrolled bigram/biphone frequency) is a reciprocal relationship. This has a practical implication. By convention, psycholinguistic researchers take the logarithm of variables that have a large range before considering them as predictors of behavioral measures of lexical access (Morrison & Ellis, 2000; Colombo & Burani, 2002; Balota et al., 2005, for instance). However, our findings suggest this is not the best way to handle frequency measures. A reciprocal function seems more applicable in terms of a simple transformation that maximizes the predictive value of most lexical measures with a large range. Table 2 shows how much of a difference taking the reciprocal of a frequency variable makes (NUANCE transformation), compared to logging the measure.

One useful piece of information that can be provided by our method is a principled answer to a question of direct practical importance to experimental psychologists: How large does a variable have to be to be considered high? We know, for example, that word frequency mediates most other variable effects (Cutler, 1981), including the orthographic neighborhood effects that are only seen in low-frequency words (Andrews, 1989). This relationship has been characterized with genetic programming in the past (Westbury,

Buchanan, Anderson, Rhemtulla, & Phillips, 2003). When designing factorial experiments to study the effects of orthographic neighborhood, we must only use low-frequency words. But how low is low frequency? By plotting predicted LDRTs by orthographic frequency, we can see exactly how frequency and LDRTs relate to each other, and get a principled estimate across a large word set of how low low frequency is. Figure 1 suggests there will be a very small effect of ON for words with a frequency above 5 occurrences per million.

As was mentioned earlier, our results suggest that phonological frequency may be more important, with respect to mediating effects. However, the shape of the relationship between it and reaction times was very similar to that of orthographic frequency and reaction times (figure 2).

Discussion

A few points from the above analysis bear further discussion. Our data seem to suggest interactions involving phonological frequency account for more variance than interactions involving orthographic frequency (Table 4). This may be an artifact of the different corpora used to derive these two measures. However, it may also have more important implications for understanding frequency effects in lexical access, and may be worthy of further scrutiny as other phonological frequency values become available.

Our observation that some variables appear to have interactions but account for little or no variance in LDRTs individually (most notably, ONFREQ and UNBP) seem in line with an account of psychological systems as reciprocally causal, as laid out by Van Orden and Paap (1997). However, we also note that almost all of the variance accounted for in LDRTs derives from 4 main effects. Ultimately, the practical concerns to an analytic approach suggested by reciprocal causality are left in limbo.

Fifteen of our sixteen predictors entered into significant nonlinear relations with lexical decision reaction times. Furthermore, all of these relations were simple (as far as

nonlinear relations are concerned), being monotonically increasing or decreasing functions. As we suggested earlier in our introduction, such relationships are not coincidental. They are fundamental to psychological entities. How fundamental nonlinearity is to psychology is suggested by the fact that, on average, our untransformed predictors account for a mere 35% of the variance in LDRTs that our transformed predictors account for (Table 2).

Inasmuch as one goal of investigations such as ours is to maximize our ability to predict some dependent measure, our finding that most variables measuring the frequency of some event have an inverse relationship with LDRTs is an important one. Previous research that has looked at frequency as a continuous variable have looked at log frequency (Morrison & Ellis, 2000; Colombo & Burani, 2002; Balota et al., 2005, for example). However, a logarithmic transformation does not appear to be the best transformation for frequency measures. At least for orthographic frequency, a reciprocal function of frequency accounts for 4% more variance in LDRTs than a log function. This is a substantial gain in our ability to predict LDRTs when compared to the amount of unique variance accounted for in LDRTs by most predictors (Table 5), constituting as it does 36% of the total variance accounted for.

The reason why a reciprocal transformation of frequency measures is a better fit for LDRTs than a logarithmic transformation may be explained by physiological constraints. We can only respond so fast to a stimuli. At some point, it becomes a physiological impossibility for us to respond any faster. This real-world constraint is captured by the asymptotic nature of a reciprocal function, but not by the continually increasing nature of a logarithmic function. An appeal to physiological constraints would seem to suggest any predictor with a large range should have a reciprocal-like relationship with measures of human performance. Generally this was true in our study. All but two of our frequency-related variables had reciprocal relationships with LDRTs. It is curious, then, why the relationship between our measures of uncontrolled bigram and biphone

frequencies and LDRTs was not best described by a reciprocal function, yet both accounted for unique variance in LDRTs (Table 5).

Conclusion

We have demonstrated that NUANCE is helpful for making sense of real problems in psychology. The studies previously described elucidate two ways which NUANCE can aid research in psychology. First, it can help simplify complex models by pruning factors that do not matter. Second, it can discover new relationships that were not previously thought to exist. These two abilities can aid in theory development as well as theory simplification. They can also be of utility in applied situations where human behavior is a critical factor. The importance of such tools is accentuated by our earlier assertion that nonlinearity is of fundamental importance to psychological behavior and our inability to easily reason in terms of complex, nonlinear relationships.

These results will hopefully encourage researchers to employ the use of NUANCE in their own work. Automating the discovery of new knowledge in the manner we have described here has very little overhead in terms of resources, and may bring to light information that would otherwise be overlooked by a traditional, analytic approach to psychology.

References

Andrews, S. (1989). Frequency and neighorhood effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 802–814.

Baayen, R. (2005). Twenty-first century psycholinguistics: Four cornerstones. In A. Cutler (Ed.), (chap. Data mining at the intersection of psychology and linguistics). Hillsdale, NJ: Lawrence Erlbaum Press.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The celex lexical database. release 2 (cd-rom).* Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Balota, D., Cortese, M., Hutchison, K., Neely, J., Nelson, D., Simpson, G., et al. (2002). *The english lexicon project: A web-based repository of descriptive and behavioral measures for 40,481 english words and nonwords. http://elexicon.wustl.edu.*

Balota, D., Cortese, M., Sergent-Marshall, S., Spieler, D., & Yap, M. (2005). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

Blalock, H. (1972). *Social statistics.* NY: McGraw-Hill.

Buchanan, L., & Westbury, C. (2000). *Wordmine database: Probabilistic values for all four to seven letter words in the english language. http://www.wordmine.org.*

Colombo, L., & Burani, C. (2002). The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain and Language*, *81*, 398–411.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*, 65–70.

Dawson, M. (2004). *Minds and machines.* Oxford, UK: Blackwell Publishing.

Dawson, M., & Schopflocher, D.(1992). Modifying the generalized delta rule to train networks of nonmonotonic processors for pattern classification. *Connection Science*, *4*, 19–31.

Feyerabend, P.(1975). *Against method*. London, England: New Left Books.

Holland, J.(1992). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. Cambridge, Ma, USA: MIT Press.

Hollis, G., & Westbury, C.(in press). Nuance: Naturalistic university of alberta nonlinear correlation explorer. *Behavioral Research Methods. Behavioral Research Methods*. in press.

Minsky, M., & Papert, S.(1968). *Perceptrons: an introduction to computational geometry*. Cambridge, MA, USA: MIT Press.

Morrison, C. M., & Ellis, A. W.(2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, *91*, 167–180.

Neisser, U.(1997). The future of cognitive science: An ecological analysis. In D. M. Johnson & C. E. Erneling (Eds.), (pp. 247–260). New York: Oxford University Press.

Peterson, J.(2005). *To err is human; to predict, divine: Neuropsychological-cognitive profiles of error-prone pharmacists.* University of Toronto, Toronto, Ontario, Canada. (Unpublished manuscript)

Popper, K.(1959). *The logic of scientific discovery*. New York: Harper and Row.

Shaoul, C., & Westbury, C.(2005, april). *Word frequency effects in high-dimensional co-occurence models: A new approach.* University of Alberta, Edmonton, Alberta, Canada. (Unpublished manuscript)

Van Orden, G., & Paap, K. (1997). Functional neuroimaging fails to discover pieces of mind in the parts of the brain. *Philosophy of Science*, *64*, 85–94.

Westbury, C., Buchanan, L., Anderson, M., Rhemtulla, M., & Phillips, L. (2003). Using genetic programming to discover nonlinear variable interactions. *Behavioral Research Methods, Instruments, and Computers*, *28*, 202–216.

Table 1

Descriptions for the 16 predictors used in study 2.

Variable	Descrption
LETTERS	The word's length, in letters
PHONEMES	The word's length, in phonemes
OFREQ	The orthographic frequency (per million) of the word
ON	The number of orthographic neighbours of the word
ONFREQ	The average OFREQ of the word's orthographic neighbours
PFREQ	The phonological frequency (per million) of the word
PN	The number of phonological neighbours of the word
PNFREQ	The PFREQ of the word's phonological neighbours
CONBG	The summed frequency that any two letter-pairs in the word occur together in the place they are in for the current word. Only counted across words of the same length.
UNBG	The summed frequency that any two letter-pairs in the word occur. Position in the word and word length do not matter.
CONBP	The summed frequency that all two phoneme-pairs in the word occur together in the place they are in for the current word. Only counted across words with equal number of phonemes.
UNBP	The summed frequency that any two phoneme-pair in the word occur. Position in the word and phoneme count do not matter.
FIRSTTRI	The frequency with which the first three letters of the word occur as the first three letters for all words.
LASTTRI	The frequency with which the last three letters of the word occur as the last three letters for all words.
ARC	The average distance between a word and all of its semantic neighbours.
NN	The number of semantic neighbours the word has.

Table 2

*Variance in LDRTs accounted for by each predictor, its log transformation, and it's best-fit NUANCE transformation. All values are for performance on the validation set. All log and NUANCE-transformed effects significant at $p < 0.001$. For untransformed variables, p-values of 0.05, 0.01 and 0.001 denoted by *, **, and ***, respectively. Differences in predictive power between NUANCE-derived fits and best maximum of the other two fits are marked: p-values of 0.05, 0.01, and 0.001 denoted by †, ††, and †††, respectively (for the methodology used to determine significance values for correlational differences, see Blalock (1972)).*

Variable	Untransformed	Log Transformed	NUANCE
OFREQ	0.015***	0.331	0.363††
PFREQ	0.002**	0.121	0.141†
LASTTRI	0.003***	0.072	0.131†††
FIRSTTRI	0.004***	0.092	0.115††
ON	0.078***	0.093	0.093
NN	0.065***	0.096	0.085
ONFREQ	0.000	0.045	0.076†††
PN	0.054***	0.072	0.066
ARC	0.039***	0.047	0.059
LETTERS	0.053***	0.048	0.059
PHONEMES	0.042***	0.039	0.048
PNFREQ	0.001*	0.027	0.048†††
CONBG	0.004***	0.008	0.025†††
UNBP	0.011***	0.011	0.018†
UNBG	0.006***	0.007	0.006
CONBP	0.001*	0.005	0.003

Table 3

Significant pairwise interactions. p-values of $p < \frac{0.05}{120}$*,* $p < 0.01$*, and* $p < 0.05$ *denoted by ***, **, and *, respectively.*

Variable 1	Variable 2	R^2
LETTERS	PFREQ	0.015***
FIRSTTRI	LASTTRI	0.010***
ON	PFREQ	0.009***
CONBP	UNBP	0.008***
LETTERS	OFREQ	0.007***
CONBG	UNBG	0.006***
ONFREQ	PFREQ	0.006***
PHONEMES	ONFREQ	0.006***
ONFREQ	UNBP	0.005***
ONFREQ	PN	0.005***
LETTERS	ONFREQ	0.005***
OFREQ	ON	0.005***
PFREQ	NN	0.004**
ONFREQ	UNBG	0.004**

Variable 1	Variable 2	R^2
PHONEMES	PFREQ	0.004**
ON	ONFREQ	0.004**
OFREQ	PN	0.003**
PN	PNFREQ	0.003*
UNBG	UNBP	0.003*
PHONEMES	ON	0.003*
LETTERS	CONBP	0.002*
PN	UNBP	0.002*
PN	UNBG	0.002*
ON	NN	0.002*
PFREQ	UNBP	0.002*
PNFREQ	UNBP	0.002*
PFREQ	PNFREQ	0.002*

Table 4

For each variable, this Table presents the summed R^2 for all significant ($p < \frac{0.05}{120}$) interactions the variable is involved in.

Variable	$\sum R^2$
PFREQ	0.030
ONFREQ	0.028
LETTERS	0.027
UNBP	0.013
ON	0.013
OFREQ	0.011
FIRSTTRI	0.010
LASTTRI	0.010
CONBP	0.008
CONBG	0.006
UNBG	0.006
PHONEMES	0.006
PN	0.005
PNFREQ	0.000
ARC	0.000
NN	0.000

Table 5

*Variables left in after stepwise, backward regression of the 16 individual variables, and the amount of unique variance each accounts for. p-values of 0.05, 0.01 and 0.001 denoted by *, **, and ***, respectively.*

Variable	R^2
OFREQ	0.321***
LETTERS	0.059***
ON	0.018***
LASTTRI	0.008***
FIRSTTRI	0.004***
PHONEMES	0.004***
ARC	0.004***
UNBG	0.003***
NN	0.002**
CONBG	0.001*
UNBP	0.001*

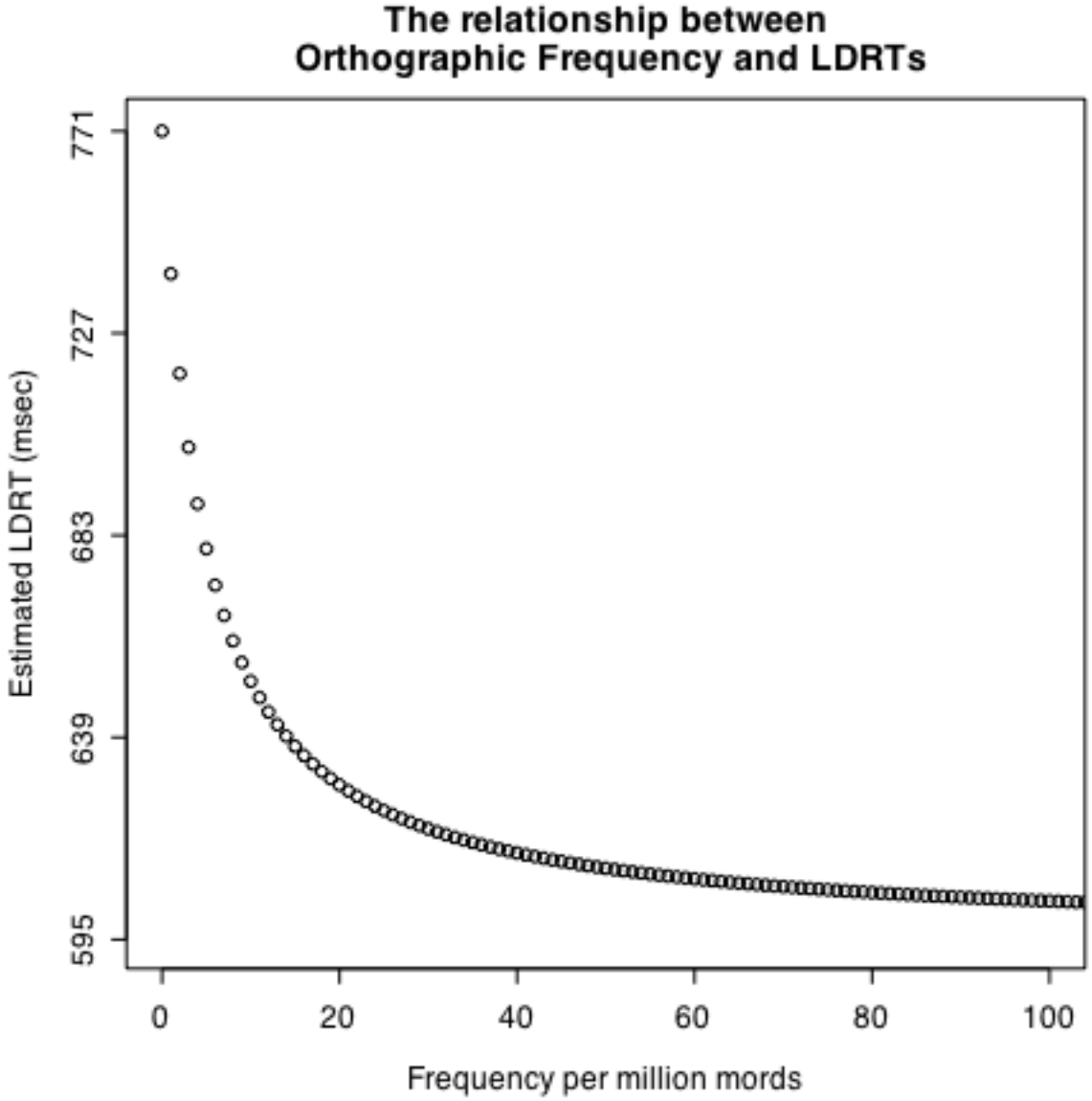
Figure Captions

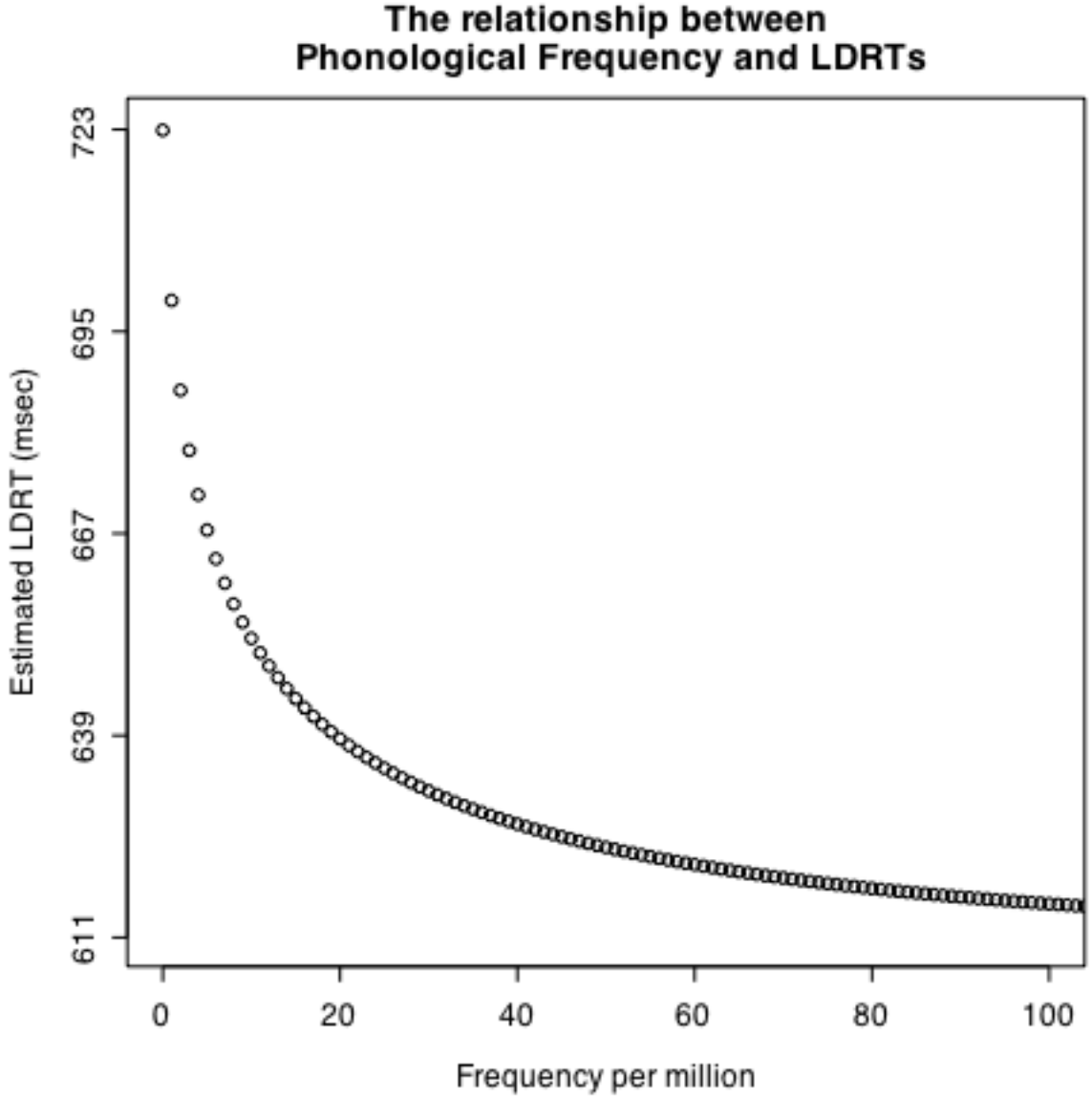
Figure 1. Orthographic frequency plotted against estimated lexical decision reaction times. The relationship is identical for Phonological frequency.

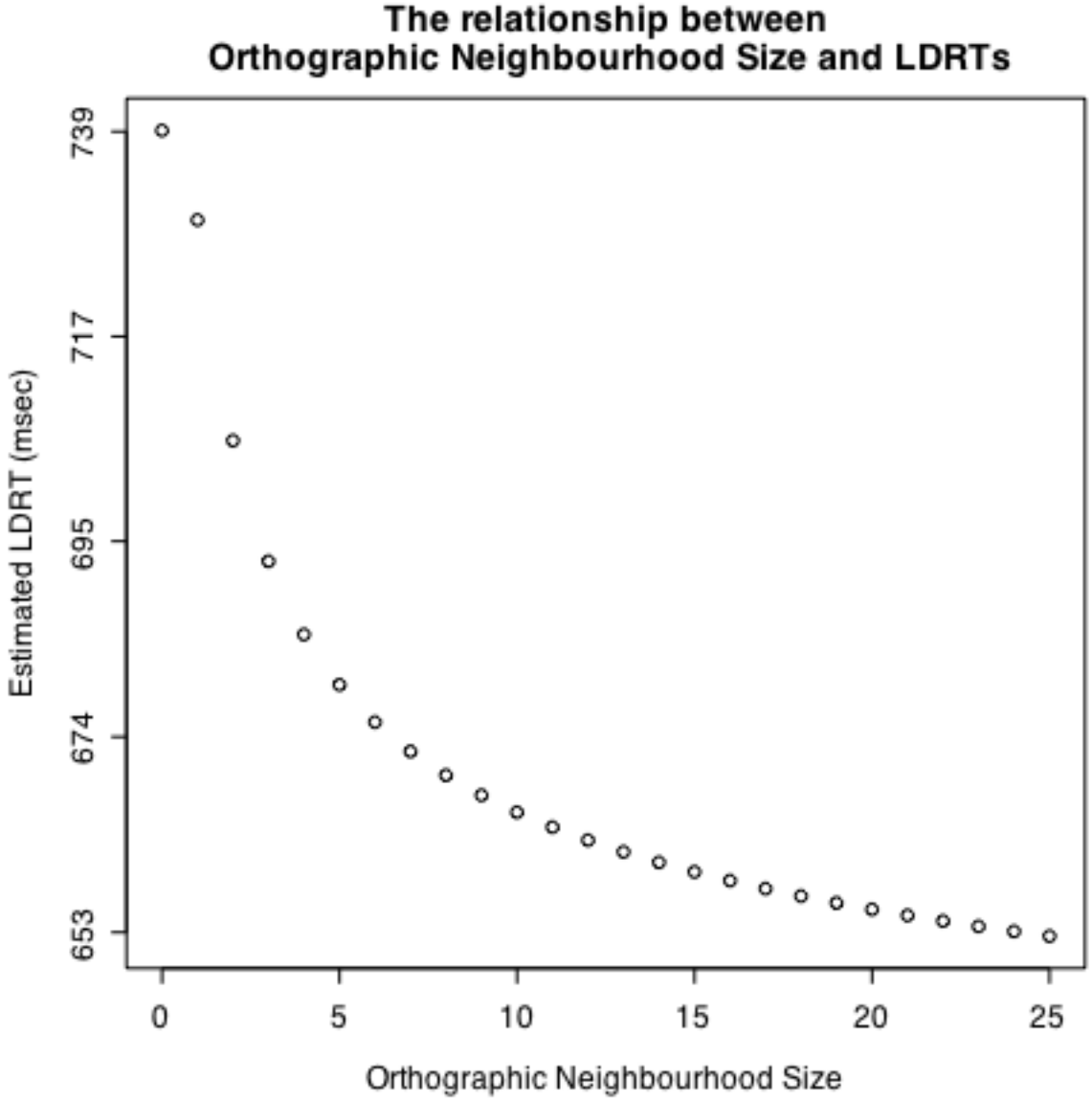
Figure 2. Phonological frequency plotted against estimated lexical decision reaction times.

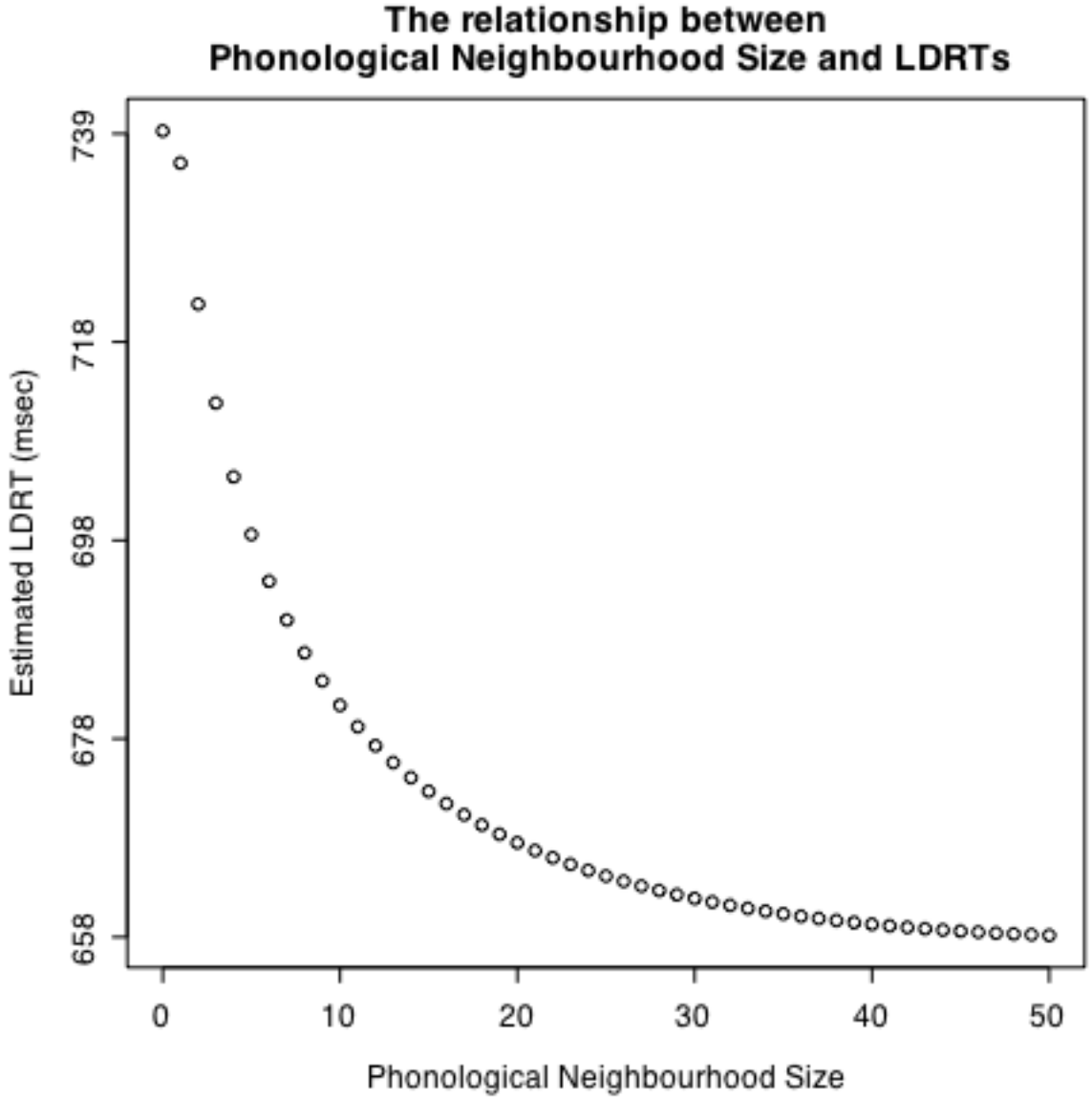
Figure 3. Orthographic neighbourhood size plotted against estimated lexical decision reaction times.

Figure 4. Phonological neighbourhood size plotted against estimated lexical decision reaction times.


The relationship between
Orthographic Frequency and LDRTs
Estimated LDRT (msec)
595
639
683
727
771
0
20
40
60
80
100
Frequency per million mords


The relationship between
Phonological Frequency and LDRTs
Estimated LDRT (msec)
723
695
667
639
611
0
20
40
60
80
100
Frequency per million


The relationship between
Orthographic Neighbourhood Size and LDRTs
Estimated LDRT (msec)
739
717
695
674
653
0
5
10
15
20
25
Orthographic Neighbourhood Size


The relationship between
Phonological Neighbourhood Size and LDRTs
Estimated LDRT (msec)
739
718
698
678
658
0
10
20
30
40
50
Phonological Neighbourhood Size