Evidence accumulation as a model for lexical selection

Royce Anders, Stéphanie Riès, L. Van Maanen, F. -X. Alario

Evidence Accumulation as a Model for Lexical Selection

R. Anders[a,*], S. Riès[a,b], L. van Maanen[c], F.-X. Alario[a]

[a] *LPC UMR 7290, Aix-Marseille Université, CNRS, France*
[b] *Department of Psychology, University of California, Berkeley, USA*
[c] *Department of Psychology, University of Amsterdam, Netherlands*

Abstract

We propose and demonstrate evidence accumulation as a plausible theoretical and/or empirical model for the lexical selection process of lexical retrieval. A number of current psycholinguistic theories consider lexical selection as a process related to selecting a lexical target from a number of alternatives, which each have varying activations (or signal supports), that are largely resultant of an initial stimulus recognition. We thoroughly present a case for how such a process may be theoretically explained by the evidence accumulation paradigm, and we demonstrate how this paradigm can be directly related or combined with conventional psycholinguistic theory and their simulatory instantiations (generally, neural network models). Then with a demonstrative application on a large new real data set, we establish how the empirical evidence accumulation approach is able to provide parameter results that are informative to leading psycholinguistic theory, and that motivate future theoretical development.

Keywords: psychometrics, sequential sampling, neural network models, response time analysis, shifted Wald, lexical retrieval

1. Introduction

Lexical selection may be broadly defined as the process of selecting a lexical target from a number of alternatives, such as when one names an object or a concept. There are a number of conventional psycholinguistic theories aimed to explain the principles that underlie lexical selection (e.g., Chen & Mirman, 2012; Howard et al., 2006; Levelt et al., 1999; Oppenheim et al., 2010), and they notably specify lexico-semantic interactions (semantic interactions on lexical alternatives, as in Figure 1). These theories vary in the nature of the principles they propose, and their levels of specificity. For example, if they also specify details of morphological and phonological element interactions. They may also vary in how far they can be extended, such as if they can account for lexical selection in individuals also with cognitive impairments.

We propose that it may be a disadvantage however, that none of these theories argue it is an evidence accumulation (sequential sampling) kind of process that underlies lexical selection; and/or that none of these theories utilize evidence accumulation models empirically, as a more sophisticated method of analysis to more deeply inform or support their claims. In this paper, we develop a case for both of these statements, and also provide demonstrations. Specifically, we newly

*Principal corresponding author
Email addresses: royce.anders@univ-amu.fr (R. Anders), stephanie.ries@berkeley.edu (S. Riès), l.vanmaanen@uva.nl (L. van Maanen), francois-xavier.alario@univ-amu.fr (F.-X. Alario)

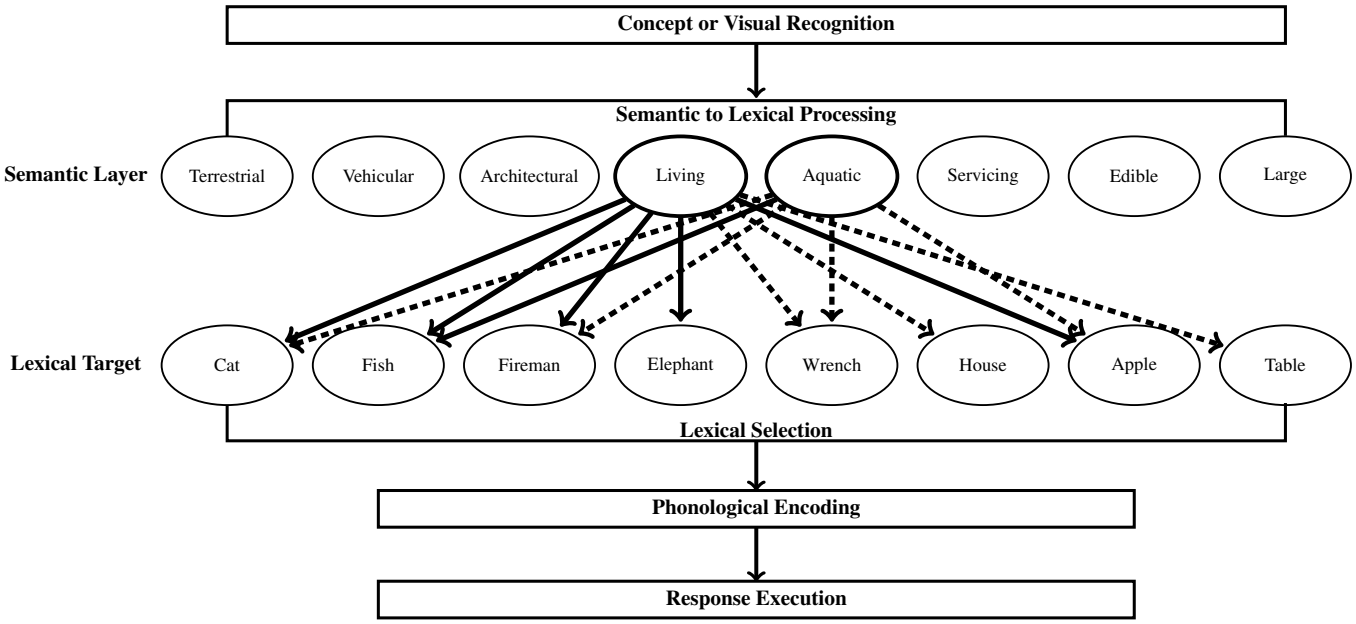

Figure 1: A depiction of the process involved in lexical retrieval, with focus on the details between semantic to lexical processing; in the style of preexisting theories (Dell & Gordon, 2003; Dell & O'Seaghdha, 1992; Dell et al., 1997; Oppenheim et al., 2010). Semantic features 'Living' and 'Aquatic' are activated by the stimulus, which provide excitatory (solid lines) and inhibitory (dashed lines) activations to lexical targets. In a later lexical selection process upstream, it is likely that 'Fish' will win, as earlier it received more excitatory activation inputs than the alternative lexical targets.

propose and demonstrate evidence accumulation as a model for the lexical selection process of lexical retrieval, and further discuss it in the context of conventional psycholinguistic theory and tradition. Secondly and to our knowledge, we provide the first quantitative data fitting of a lexical selection data set (in the picture naming task) with an evidence accumulation model, and we also contribute a new large lexical retrieval data set to this effort. Thirdly, we newly show how this empirical approach with evidence accumulation modeling, can provide parameter results that are informative to a leading psycholinguistic theory, and that motivate future theoretical development.

2. Proposing Evidence Accumulation for Lexical Selection

In this section, we begin with a discussion of how conventional psycholinguistic theories can benefit from using the evidence accumulation approach empirically. Then, we discuss ways in which the evidence accumulation approach can be embedded within current psycholinguistic theory, for modeling the lexical selection process.

2.1. The Empirical Approach

Conventional psycholinguistic theories are generally conceived of by a combination of scholarly interpretations of many prior empirical results, consisting generally of the lexical choice patterns and mean response times (RTs), and simulation exercises of principles, in which simulatory instantiations of the theory are provided in the form of a neural network model (NNM). For the former empirical approach, firstly it is becoming increasingly clear that there is much more to RT data

than the mean (Balota et al., 2008; Balota & Yap, 2011; Luce, 1986). Thus any quantitative or measurement model that can concurrently account for the lexical choices and the full RT distribution, such as an empirical evidence accumulation model, is a more sophisticated analysis tool which can better inform theoretical development. Secondly for the latter, since NNMs are simulations of a psycholinguistic theory, they are not made to quantitatively fit data at high specificity; instead, the general purpose of these NNMs is to show that a simulation of the theory can indeed correspond to specific features of the data.

2.1.1. Simulation versus Fitting

Neural network model simulations can often perform accurately enough to reproduce the lexical choices of a 'canonical' healthy speaker. We utilize canonical in the sense that the NNM is not geared to quantitatively fit specific subjects in a simulation, but that it is instead used to show that its simulation results correspond to a typified speaker, or the average responses pooled across all speakers. In this way, NNMs do generally well to reproduce the response patterns of canonical healthy speakers. Also of note, is that some NNMs allow further calibration to reproduce error patterns of individual patients with aphasia (e.g., see Dell et al., 1997, 2013; Foygel & Dell, 2000; Schwartz et al., 2006).

The largest difficulty for current NNM simulations however, is the ability for these models to concurrently reproduce the RT distributions with the lexical choices, whether it is over all speakers, individual speakers, or experimental design cells. One reason is because many of these NNMs model only a portion of the RT per trial (e.g. the semantic and lexical time portions in Figure 1), and/or simulate a different time statistic than an RT, such as the total number of network iterations or cycles (e.g. Chen & Mirman, 2012; Howard et al., 2006; Oppenheim et al., 2010), which are currently correlated only to some canonical aggregate RT measure (e.g. the mean RT over all speakers, and not the RT distribution itself, or of each speaker). A second reason is that these NNMs are generally too parametrically rich to fit to data (for example, WEAVER++ by Levelt et al. 1999, which aims to model nearly every computational component of the RT). Here we define quantitatively fitting data with a model, as having a principled procedure that allows one to find the most likely and unique set of parameters that appropriately quantify the data. To our knowledge, there is currently no published methodology developed, or rather implemented, for how to quantitatively fit data with one of these NNMs.

In this section we have discussed the challenges for NNM simulations to reproduce response choices, and particularly RT distributions. Capturing both aspects well and concurrently however, is a signature advantage of the empirical evidence accumulation approach, from which psycholinguistic theories can be better informed or even supported (Luce, 1986). This is a notable improvement in contrast to other complementary empirical approaches that can only do one of the two: for example a recent application fits an RT distribution measurement model to the RT data, the ex-Gaussian, which does not handle response choices, and uses the results to inform a better simulation of a psycholinguistic theory (WEAVER++, Roelofs, 2008).

2.1.2. Empirical Implementation

Thus in further detail, the empirical evidence accumulation approach provides at least three functions that current psycholinguistic NNMs cannot yet do: (1) the models can concurrently quantitatively fit the RT distributions and lexical

choices, and at a high resolution (e.g. of every subject, within conditions); (2), they fit the full RT from start to finish, and not only time passed for middling components (e.g. exclusively lexico-semantic interactions, as in Figure 1); and furthermore (3), they can even model participant response biases for certain lexical choices. They are thus one of the most sophisticated empirical approaches currently available for analyzing RT data that consist of a number of choice alternatives (see Ratcliff & McKoon, 2008; Ratcliff & Smith, 2004), from which psycholinguistic theory can be better informed.

The most natural, empirical evidence accumulation approach for lexical selection, which involves a lexical choice among a number of alternatives, is the racing evidence accumulation design, which involves multiple accumulators (e.g. Brown & Heathcote, 2008; LaBerge, 1962; Nakahara et al., 2006; Usher et al., 2002; Vickers, 1970, 1979). In the application to this framework, the activations of each lexical alternative race to a threshold, and the first to arrive wins. This is a departure from what is currently the most conventional evidence accumulation model (e.g. the drift diffusion model DDM, Ratcliff, 1978; Ratcliff & Murdock, 1976), which instead involves a single-accumulator, that is typically used to handle two-alternative forced-choice tasks. This is because in contrast, lexical selection typically involves a $k > 2$ word choice task (choosing a word from more than two options, e.g. selecting 'Fish' as in Figure 1). More specific details on the mechanics of evidence accumulation, such as of Figure 2, will be discussed in Section 3.

A racing accumulation model may quantify the activation of the lexical target that was spoken, as well as the lexical alternatives. There are a number of methods in how one might quantitatively fit data with a race model, however there must be adequate numbers of observations (e.g. along choice alternatives, by subjects and the experimental factors of interest) in a data set to permit an appropriate fit. Otherwise, the empirical racing accumulation model approach instead becomes simulatory, like a number of NNMs, in which it is difficult to know if a preferred parameterization is the most likely and unique set of parameters that appropriately quantify or give rise to the data. Indeed arguably, there is a gap in the current size of most lexical retrieval data sets (e.g. on picture or concept naming) to enable appropriately fitting the most optimal racing models, which quantify the activation of the lexical alternatives, in addition to the lexical target. As a foundational step, we propose that as an interim solution to this (while experiments do not possess sufficient observations along the relevant choice alternatives, crossed with the independent variables of interest), one can first focus on quantitatively fitting simpler models of evidence accumulation, that mainly focus on the activation of the lexical target.

Specifically in these cases, we propose to model the aggregate activation of the lexical target spoken (and not the lexical alternatives), using arguably the simplest case of a race model, namely the shifted Wald model (SWM, Anders et al., in review; Folks & Chhikara, 1978; Heathcote, 2004; Luce, 1986; Ricciardi, 1977), which involves only one accumulator with one threshold, and positive drift rates. We will discuss the SWM and its relationship to psycholinguistic theory in further detail, in Sections 3 and 4, and then provide a demonstration of the approach on real data in Section 6.

2.2. Placement within Psycholinguistic Theory

While psycholinguistic theory stands to gain a number of potential advantages by the empirical evidence accumulation approach, one might also consider the implications of evidence accumulation theoretically, as an important framework to bring to conventional theory. In this section, we will discuss how appropriately involving such a framework within these lexico-semantic theories, may arguably be a notable advantage, if not an advancement.

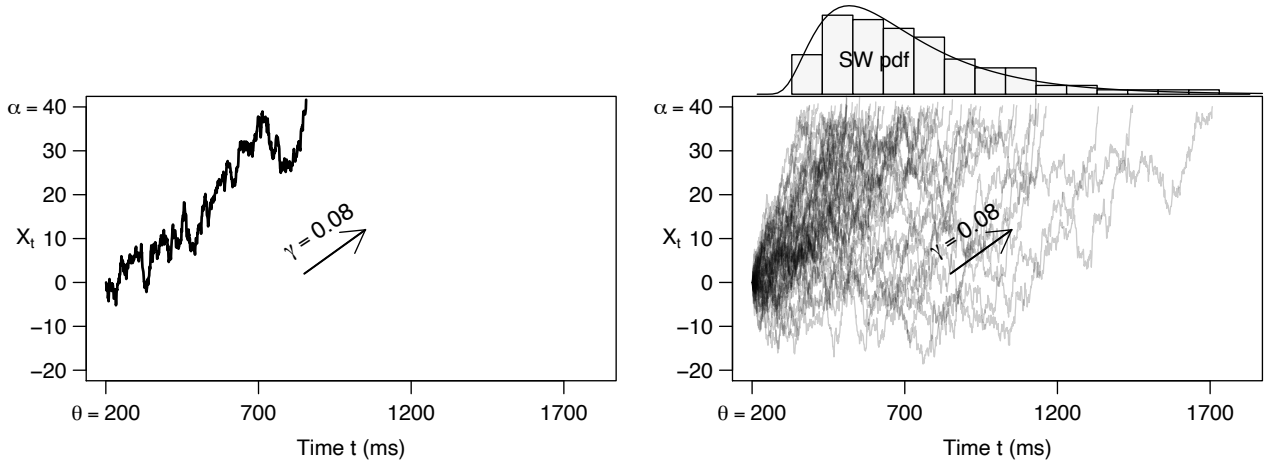

Figure 2: The SWM as a model for lexical retrieval; describing the RT data in the context of the aggregate latent activity of the lexical target accumulating to threshold, α, at rate, γ, where θ accounts for the time lapsed outside of (around) this process (above and below Lexical Selection in Figure 1). Left, a single trial is modeled with the parameters. Right, many trials (e.g. an experimental design cell) are modeled with the same parameter values, and these ultimately form a SW distribution shaped with the same signal accumulation parameters.

The tradition of evidence accumulation modeling in psychology is beginning to have a long history, particularly since the 1960s (Carpenter, 1981; Gerstein & Mandelbrot, 1964; LaBerge, 1962; Ratcliff & Murdock, 1976; Vickers, 1970), and is becoming increasingly prevalent, gaining broadening empirical support for concurrently handling choice behavior and RTs over an increasing number of experimental domains (Anders et al., in review; Donkin & Van Maanen, 2014; Mulder et al., 2013; Ratcliff & Rouder, 2000; Ratcliff & McKoon, 2008; Ratcliff et al., 1999; Trueblood et al., 2014; Van Maanen et al., 2012a; Winkel et al., 2012). This ability to concurrently handle response choices and RTs, is consistent with the notion that while response rates (or correct and error rates) in general may be stochastically quantified by classical signal detection theory (Green & Swets, 1966), response rates stochastically quantified in the context of the time domain is achieved, instead canonically, by the evidence accumulation framework (Pike, 1973).

Beyond the previous empirical and theoretical applications of evidence accumulation however, there are also continuing developments of the framework in the neurosciences literature; particularly in support of the presence of upstream neuronal units that compute accumulated evidence in such similar fashions, across a variety of experimental domains. Specifically, in research across monkeys (de Lafuente et al., 2015; Shadlen & Newsome, 2001), rats (Brunton et al., 2013; Erlich et al., 2015; Hanks et al., 2015), and humans (Donner et al., 2009; Fitzgerald et al., 2015; Kelly & O'Connell, 2013; O'Connell et al., 2012) performing various tasks; and typically the specific frontoparietal areas that are active in the computation of evidence accumulation, depend on the kind of task that is performed (Mulder et al., 2014). Moreover, some of these authors show that the informational activity downstream is strongly corresponding to the respective upstream frontoparietal activity, which is computing the accumulated evidence (or favor) for a particular choice behavior. From these kinds of increasingly prevalent findings, it is not unreasonable to consider that earlier processing (e.g. in semantic levels and before, in Figure 1) is feeding activity into an upstream lexical selection, evidence accumulation process. Then as for lexical selection, where

such an activation accumulation, racing process may occur in the brain is still subject to debate. However studies in naming latencies (Moss et al., 2005; Schnur et al., 2009) and errors due to lesions (Schnur et al., 2005, 2006, 2009) in lexical selection tasks, are correlated with variations of the activity and/or integrity of the left inferior frontal gyrus (LIFG).

Finally, there are some related previous applications. For example, the evidence accumulation framework has been applied to a more "artificial" language domain involving word-discrimination, known as lexical decision (Taft & Forster, 1975), in which stimuli generally consist of a single word or pseudo-word, and responses consist of a two-alternative choice task for word identification (indicate 'yes' for is a word, or 'no' for is not a word; for examples see Dufau et al. 2012; Ratcliff et al. 2004; Wagenmakers et al. 2008). Secondly, there is notably a formal argument of evidence accumulation for lexical retrieval made prior (RACE/A, Van Maanen & Van Rijn, 2007; Van Maanen et al., 2009, 2012b), in the context of picture-word interference (PWI); but this evidence accumulation instantiation of standard ACT-R theory (Anderson, 1996; Anderson et al., 2004) is different from the conventional psycholinguistic theories that we focus on here, which instead specify lexico-semantic interactions in the style of Figure 1. In common with the conventional theories however, RACE/A provides a theoretical, simulatory model which cannot quantitatively fit data in the way we have defined it.

Thus given all of these considerations previously discussed, it may very well be reasonable to consider the possible connection (and advantages) that evidence accumulation may bring to describing other processes that have not been previously considered in conventional theory, such as lexical selection. Our work builds upon the previous works mentioned. Notably, our contribution is to develop the empirical and theoretical cases for evidence accumulation in the more complex setting of conventional lexico-semantic theories of lexical retrieval, where in the experimental domain, stimuli consist of concepts or pictures, and the responses consist of a lexical selection among a number of alternatives ('cat', 'apple', 'lawyer', 'mathematics'). In the following section, we develop the empirical application of evidence accumulation to lexical selection data, and then proceed with theoretical connections.

3. The Shifted Wald Evidence Accumulation Model

In this section, we introduce and describe the SWM for the empirical evidence accumulation approach, and then in the following section, we link the SWM process and parameters to conventional psycholinguistic theory. As mentioned previously, the SWM is one method in which the evidence accumulation approach may be empirically applied to lexical selection data. It could be considered a canonical, simplest case of racing accumulation that models the activation of the lexical target spoken, and not the alternatives. Then as observation numbers permit in lexical retrieval data, one would do well to also consider more complex racing models for the empirical approach, as they could reveal more thorough information, certainly about the lexical alternatives.

The SWM is an evidence accumulation model involving the same kind of random-walk process as the popular DDM, except there is only one absorbing threshold, and the drift rates are positive. While the DDM is appropriate for two-alternative forced choice paradigms, the SWM is appropriate for paradigms in which one characteristic response is observed across varying latencies. The race model expansion of the SWM occurs when many SWM accumulators are involved in a single trial, and the first that reaches the threshold wins; the race versions have been respectively proposed and simulated

by LaBerge (1962) and Usher et al. (2002).

An illustration of the SWM as a model for the behavior is provided in the left plot of Figure 2. Here a single trial is modeled. The fluctuating black line is a representation of evidence that is accumulating over time for the lexical target; note that the evidence begins at a value of 0, and increases (with noise) over time, until it hits the necessary threshold value, here a value of 40. Upon reaching the threshold, the response is initiated. Parameter γ indicates the rate of evidence accumulation, α is the value of the threshold needed to initiate the behavior, and θ is the time to perform the behavior, such as for response encoding to execution (here abbreviated TEA for time external to the accumulation process), and may also include time for perceptual processes.

In the right plot of Figure 2, many trials (e.g. a subject within an experimental design cell) are modeled with the same three parameters that simulated the single trial in the left plot. Note that all of these finishing times, of when the evidence accumulates to the necessary threshold, plus the TEA (θ), are the response times (RTs), and these are quantified directly by the probability density function of the SW distribution, also known as the three-parameter inverse Gaussian distribution. Finally, it should also be clarified that for illustrative simplicity, here θ (TEA) is placed before the evidence accumulation begins (at $\theta = 200$ ms). However whether θ is placed before, after, or split around the actual accumulation process (e.g. accounting for both concept/visual recognition and response execution time), all of these options are quantified equally (mathematically).

4. Linking the Shifted Wald to Conventional Psycholinguistic Theory

The relationship of the SWM to conventional psycholinguistic theories, such as the traditions by Dell & O'Seaghdha (1992) to Chen & Mirman (2012), can be easily made by discussing the relationships between Figures 1 and 2 within the context of a lexical retrieval experimental paradigm, such as the canonical picture naming task. In the picture naming task with healthy speakers for example, it is often the case that the participant responds with the correct lexical target in near or more than 95% of trials, yet so at varying latencies. Thus when the SWM is applied to cells with these correct trials, the activation X_t is always corresponding to the response in which the activation of the lexical target (e.g. 'Fish' in Figure 1) won the lexical selection process (e.g. the type of lexical selection process that conventional theory claims). We are thus estimating a very simplified type of network model in which we zoom in on the activity progression of just a single node (in this case the lexical target) and its selection, though fortunately said model can be quantitatively fit to the data. Thus in relationship to the psycholinguistic theory, the drift γ refers to activation rate of the lexical target, and α corresponds to the amount of activation needed to select the word; α can also be interpreted as inverse to starting activation of the lexical target. Then in cases of experiments where words and pictures are balanced in their complexity, parameter θ will handle time for visual recognition and response encoding/execution (as in Figure 1), because it typically accounts for process times that are mostly invariant or balanced across trials; though even if trials are not balanced but blocks are, these effects should be averaged out in this parameter across cells.

5. Demonstrating the Evidence Accumulation Approach to Specific Theory

An interesting conventional psycholinguistic theory to choose for demonstrating the empirical and theoretical contributions of the evidence accumulation approach, here by the SWM, is the dark-side theory, or model (DSM, Oppenheim et al., 2010), which is an extension of an important lexical retrieval theory by Howard et al. (2006). The dark connotation results from the theory's major development to demonstrate the power that learning-induced forgetting can have on lexical retrieval (a.k.a. retrieval-induced forgetting, by semantic connections that weaken as a result of learning or retrieving other words). Next, the DSM aims to make statements on both lexical choices and RTs, and it provides one of the only NNMs that can adequately simulate time statistics (the number of network boost cycles, along with lexical choices) for canonical unimpaired speakers, as well as the error rates and error types for canonical aphasic speakers (e.g. see Caramazza & Hillis, 1990; DeLeon et al., 2007; Dell et al., 1997; Foygel & Dell, 2000), that correlate well with mean RTs. Thus the DSM provides a solid NNM. In addition, a fundamental reason why we choose the DSM is because its implementation of lexical selection (in the NNM) can very well be argued as a form of evidence accumulation.

For example, the DSM locates lexical retrieval choices in the time domain, and does so stochastically, through the following process: the resultant lexical activations that occur from excitatory and inhibitory semantic inputs (see Figure 1) are activity accumulation rates, and the first item to reach a fixed threshold is modeled as the lexically-selected word. Such a process includes nearly everything of a standard $k > 2$ alternative choice evidence accumulation model: (a) drift rates for each choice alternative, and (b) a threshold (in this case shared by all lexical activity accumulators); however it does lack (c) a non-decision or motor response time, but this may be sensible after all, since the process is used only to describe the lexical selection time portion. By using this kind of accumulation process at lexical selection, the model does well to concurrently simulate lexical choices and relative time statistics for both unimpaired and impaired speakers. Despite this advantageous inclusion however, the DSM as the author notes, still cannot quantitatively fit the RT data (p. 248, Oppenheim et al., 2010), because it is overparameterized for the current types of data we have available, and it models only a portion of the process (see 'Semantics' to 'Lexical Selection' in Figure 1) by using a non-RT statistic (boost cycles).

Thus since the DSM can be connected to evidence accumulation theory, but still cannot quantitatively fit data, it provides for a nice example case for how evidence accumulation modeling and psycholinguistic theory can be complementary with one another. Therefore, we select it as our example theory in our empirical evidence accumulation exercise.

6. Application to Lexical Retrieval Data

In this section, we demonstrate the SWM as a first approach of quantitative evidence accumulation fitting of the lexical retrieval paradigm, on a large new experiment of the *blocked-cyclic* picture naming task for lexical retrieval. Indeed a major canonical paradigm for studying lexical selection is this naming from pictures (or picture naming) task (Cattell, 1886; Carroll & White, 1973; Oldfield & Wingfield, 1964). The task is to rapidly name pictures as they sequentially and independently appear, while the participant's perceptual focus is fixated on the center of the display. The picture paradigm avoids the confound of simply phonologically reproducing a printed word, and insures that the participant lexically retrieves, albeit after the visual recognition of a picture. An alternative, non-pictorial paradigm for lexical selection is the naming from

definition task (e.g. Marques, 2005), in which participants name from reading a description of a concept. In both cases, after a visual or conceptual recognition, yet before response production, it is generally accepted that there is an intermediary process in which a number of semantic features are combined in order to select the lexical target (for illustration, see Figure 1).

In blocked-cyclic picture naming, a same set of pictures, sharing a certain degree of semantic context (or relatedness), is repeated by different orderings within mini-blocks (for more information, see Damian et al., 2001). The three main experimental factors of primary interest in this picture-naming task are: semantic context (degrees in which pictures in the set belong to the same semantic category), repetition cycle (the number of times the picture set has been repeated within the block), and lag (the number of other pictures seen since last seeing the current picture). The new experiment involves two additional intermediary levels of semantic context, six levels of repetition cycle, and nearly ten lag levels; as well as a larger number of trials (864 per participant after warm-up) and healthy-speaking participants (23) than a number of other studies (Damian et al., 2001; Howard et al., 2006). The semantic conditions from lowest to highest context are of the ordering 16, 23, 32, 61: the first digit refers to the number of pictures per set, and the second digit refers to the number of picture sets in the block. The total experiment resulted in 98.2% of trials being correct responses and 1.8% as errors. A full methods description of the experiment is included in the Appendix.

Fitting Approach

The fitting method utilized for the SWM combines techniques of maximum likelihood and deviance criterion minimization (as developed in Anders et al., in review). In this approach, an independent SWM is fit for each unique experimental design cell, in which main effects and interactions can be observed on the parameters of evidence accumulation. Note that for every design cell, parameters: drift γ, threshold α, and TEA θ are estimated. The design cells we fit pertain to factors that are already demonstrated in prior literature to be important for the paradigm: semantic context, repetition, and lag.

However, here it is important to illustrate that even with arguably the simplest evidence accumulation model, factor levels need to be combined to have enough trials in each design cell, to quantitatively fit the model. By collapsing the repetition factor into two levels (1-3 and 4-6) and the lag factor into two levels (2-5 and 6-12), the result is $N = 23 \times 4 \times 2 \times 2 = 368$ cells that will be fit, each having an appropriate average cell size of 53 trials, with standard deviation (SD) of trials 13, and range $(28, 78)$. Note that here we are fitting by subject and conditions, and an even better fit to aspire to in future experiments, is to have enough trials to fit also along items, which can be a large source of variance. Though at least in this experiment, the same 36 items are randomly balanced and used equally in all blocks.

Finally before fitting, on each cell a very light processing of potential contaminant RTs (see Barnett & Lewis, 1994; Ratcliff & Tuerlinckx, 2002) was performed by an elimination criterion of below three or above six median absolute deviations (MADs, see Leys et al., 2013) from the cell RT median (preserving the long tail RT values), resulting in $19,168$ trials for analysis out of the original $19,506$ (1.7% of trials omitted). Then to fit the model, for each cell, maximum likelihood method of moment estimators were used to calculate the three SW parameters when a shape parameter is proposed (β, as in Anders et al., in review). Then searching across the near-entire range β, the optimal parameter set is selected according

to the minimal difference between the data to model-predicted RT quantiles, by 100 equally-spaced quantile points in the range of $(.02, .98)$.

Results

In this section, the results of the SWM applied to the lexical retrieval data are presented. First model fit checks are assessed to check the degree to which the model appropriately fits the data, such that if the parameter results are appropriate to interpret. Secondly following a satisfactory fit, the main parameter differences are reported. Then finally, an interpretation of these parameters and their relationships to psycholinguistic theory is provided.

Model Fit Checks

The top plot in the left column of Figure 3 provides the quantile-quantile (QQ) matching of the deciles of the fit-simulated distributions against the observed distributions; it contains all 368 cells. The QQ plot may show overall trends in systematically misfitting quantiles of the distribution, as well as misfit outliers; it also gives an idea about the scale and range of the data. The importance of this check is to observe critically any curvatures in the plot, which is a strong sign of misfit. As one can see, there is no systematic curvature in the plot and the SWM performs systematically well on the data set, with minimal outliers.

Then the bottom plot provides the distribution of residuals for each of the nine deciles across the 368 cells fit. In this model fit check, one might optimally see a distinct ordering of decile residual distributions, due to the property that residual magnitude tends to correlate with RT data variance and magnitude (e.g., see Anders et al., in review). In this application, the ordering is slightly less distinct, which may be due to data noise or that multiple repetition or lag levels are aggregated into two levels. However, it is nicely shown that the fit does well to recover the deciles of each cell, with a mode residual of only 4-6 ms for every cell; and most importantly, there are no outlying decile distributions, showing no particularly poor recovery of one decile from another. Furthermore, note that the larger variance of the last decile distribution is appropriately located, as it contains the largest outlying RT values. Thus given that the model fit also appears to be supported in the assessment of this model check, we proceed to an analysis of the fitted parameter results.

Parameter Analysis

The right column of Figure 3 provides the parameter main-effect results of the SWM fit to the data. The three plots contain the main-effect means, and pairwise-difference standard errors, of the model-fit measurements of the three SWM parameters: γ, α, and θ, by experimental factor: the four semantic context levels, the two repetition groups, and the two lag groups. The main-effect means are calculated by the mean of within-subject means for a given experimental level. The pairwise-difference standard errors are calculated for each pair of adjacent experimental levels, by computing the standard error of the within-subject differences between a pair of adjacent experimental levels; these standard error bars are indicative of the significance levels on the parameters that our ANOVA analyses return.

Beginning with the effect of the semantic context levels in Figure 3, the RTs are slowed distinctly with a decrease in the signal accumulation rate parameter, γ, and secondarily with an increase in the signal criterion parameter, α, and no effect

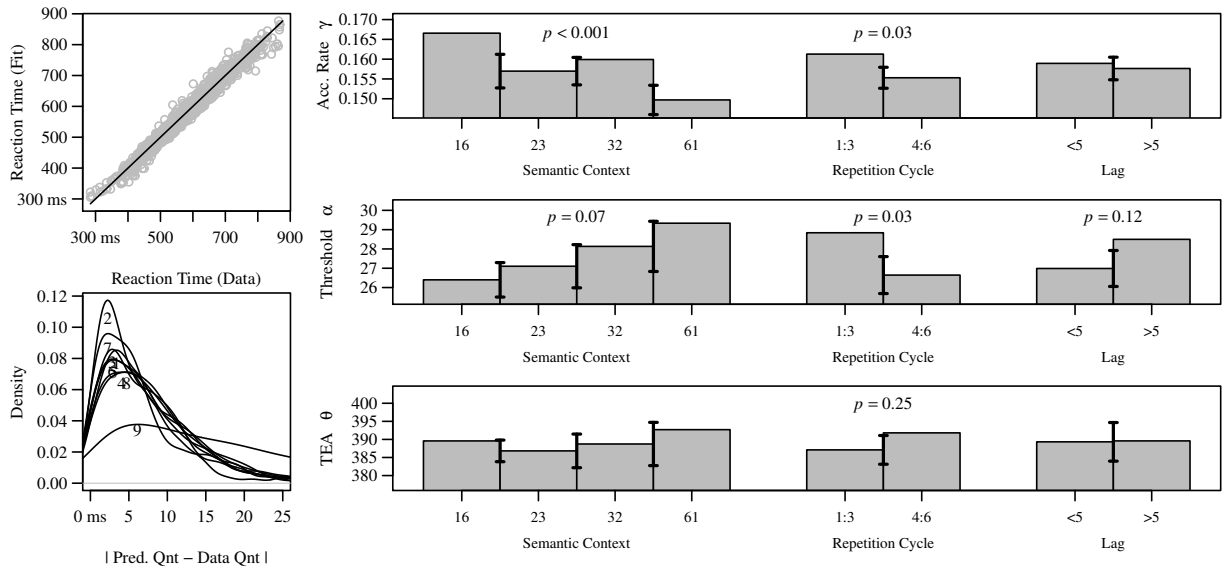

Figure 3: SWM fit to the observed data. The top-left plot shows the quantile-quantile match for the nine deciles (0.1 to 0.9) for each of the 368 cells. The bottom-left plot shows the distribution of residuals for each of the nine quantiles across the 368 cells; where the residual is the absolute difference (in ms) between the observed quantile and the model-predicted quantile. The next column provides mean parameter values with bars representing standard error of the mean ($N = 368$) grouped by experimental factor. From top to bottom are respectively the parameter values for activity accumulation rate, γ, inverse baseline activation (or threshold), α, and time external to activity accumulation (TEA), θ; and ANOVA p-values are provided for marked differences.

in external time θ; this is supported by ANOVAs across the three parameters ($F_\gamma(3,66) = 6.26, p < .001, \eta_p^2 = 0.22, \eta_G^2 = 0.036$; $F_\alpha(3,66) = 2.45, p = .071, \eta_p^2 = 0.10, \eta_G^2 = 0.016$; and $F_\theta(3,66) = 0.45, p = 0.72, \eta_p^2 = 0.02, \eta_G^2 = 0.001$)[1]. The absence of an effect on θ indicates that this effect of semantic context is not simply shifting similar distributions and their means, which perhaps some previous research assumes, but rather, the distribution shapes are changing with respect to the experimental factors.

With regard to the effect of repetition, the largest effect on RT magnitude is observed by a decrease in signal criterion level, α, which supports faster RTs with more repetition; secondarily at a smaller effect on RT magnitude, there is a distinct opposing effect[2] on signal accumulation rate, γ, that in contrast supports slower RTs with repetition. Across repetition, again there is no distinct effect observed in parameter θ. These results are also supported by ANOVA ($F_\gamma(1,22) = 5.08, p = .034, \eta_p^2 = 0.19, \eta_G^2 = 0.009$, $F_\alpha(1,22) = 3.88, p = .032, \eta_p^2 = 0.19, \eta_G^2 = 0.010$, and $F_\theta(1,22) = 1.40, p = 0.249, \eta_p^2 = 0.06, \eta_G^2 = 0.001$). Then for picture lag levels, increasing lag provided for a trend in the signal criterion parameter, in the direction that supports slower RTs. However, the ANOVA results do not support significance of this effect ($F_\gamma(1,22) = 0.20, p = .656, \eta_p^2 = 0.01, \eta_G^2 = 0.000$, $F_\alpha(1,22) = 2.62, p = .12, \eta_p^2 = 0.11, \eta_G^2 = 0.008$, and $F_\theta(1,22) =$

[1]For an explanation of effect sizes η_p^2 and η_G^2, see Bakeman (2005).

[2]Note that these kinds of opposing effects on the distribution increase the chance for a Type II error, if one were to do classical tests (e.g. ANOVA) on the raw RT means rather than on these parameters.

$0.00, p = 0.961, \eta_p^2 = 0.00, \eta_G^2 = 0.000$). Finally, no significant interaction effects were found between the factors in the ANOVA results.

Interpretation

In this section we demonstrate how the parameter results may be interpreted in respect to psycholinguistic theory. As mentioned previously, we select the DSM theory as an exemplary candidate for demonstrating how the evidence accumulation approach can be integrated.

I. In respect to semantic interference in the DSM (that arises from greater semantic context), the SWM results are nicely consistent with the DSM theory. The evidence of the most semantically-relevant target accumulates more slowly when there is greater semantic interference; then as a secondary smaller effect, there is also a higher threshold with greater semantic interference. Both of these results are consistent with predictions by the DSM theory, in that for their racing accumulation mechanism, the drift of the lexical target is slower in cases of semantic interference; and since the threshold is a function of the difference between the target and the best alternative starting drift rates, this also predicts a larger threshold during interference, which is what we observe with the SWM fit on the data. The results are also further consistent with the DSM in more simpler ways, as based in their lexical retrieval NNM, the drift and threshold are always negatively correlated with one another, and the threshold itself has less importance (p. 244 Oppenheim et al., 2010). Thirdly, no effect was observed on external process time θ, however while the DSM does not aim to make inferences about such processes, it is fortunate to see that with a quantitative fit of the data by an evidence accumulation model, it confirms that the variance in the RTs are not significantly different due to response execution or recognition duration differences (for example), and are rather different from lexical selection duration differences, which is supportive of the DSM theory.

II. Next in regard to repetition, the SWM results are also nicely consistent with the DSM theory, and furthermore, the SWM fit may offer a disambiguation of the theory. For instance, during repetition cycles, it is the case that both *priming* and interference are occurring before repeating the same picture. Thus according to DSM theory, one can expect both a facilitatory and inhibitory effect on the lexical retrieval process, and these kinds of opposing effects can be seen in the increasing-and-then-decreasing selection time trends in the DSM simulations across cycles (e.g. Figures 10a, 11a, 12c in Oppenheim et al. 2010). This is indeed nicely consistent with what we observe with the SWM: a facilitory effect is expressed in a lower threshold that needs to be reached, while the inhibitory effect is expressed in a slower drift rate. Finally, no effect was observed on external process time θ.

The DSM theory offers two ways in which these facilitory and inhibitory trends may be jointly modeled: (*i*) by learning (semantic input weights change over trials which result in different lexical target drift rates) combined also with a competitive racing process (in which the threshold is always negatively correlated to the drift rate of the lexical target), or (*ii*) by only learning, in which the threshold is free to vary without a necessary correlation to the drift rate, which may be considered a non-competitive racing process. Simulations of the DSM were unable to conclude favor of one method over the other, yet importantly here, the empirical evidence accumulation approach offers a disambiguation for the theory, a case for (*ii*). This is because the lower target drift rates observed by the inhibitory (or interference) effect, imply higher lexical alternative drift rates by the DSM, and since the theory's evidence criterion is a subtractive function of the lexical target and the alternative

drift rates (see Eq. 4 in Oppenheim et al. 2010), method (*i*) necessarily predicts a higher threshold. However empirically, instead we observed a lower threshold, despite there being lower target drift rates due to inhibition, which method (*ii*) can account for. This may hence suggest that it is possible for lexical selection to occur as a non-competitive racing process.

The empirical evidence in support of (*ii*) however, provides a new challenge for the DSM theory. Particularly, how will the DSM predict what the actual threshold values are in both semantic interference (higher thresholds, negative correlation as discussed in **I.**) and repetition cycle level increases (lower thresholds, positive correlation discussed here in **II.**) if their only formulaic method of specifying the threshold previously, provided for a decision threshold that is always negatively scaled with the drift? Since the DSM cannot quantitatively fit the data, it must propose an additional feature that can serve to generally formulate the behavior of the threshold for these two patterns in the non-competitive paradigm. The model fit thus provides a motivation for additional development of the theory.

III. Finally in respect to lag, the SWM results provide correspondence to the DSM theory, and also a motivation for further research. Lag effects may be considered indicative of what is known as *decay* in psycholinguistic theory (Anderson, 1983; Berg & Schade, 1992), of the lexical target activation. The SWM fit provided for no lexical selection parameter differences to be notably significant in the case of lag, which is consistent with the DSM theory that assumes regular decay effects to be small, or insignificant in the lexical selection process.

One may note however that the fit does provide a suggestive difference in the threshold parameter for large lag distances, which may be a motivation for future work. Specifically, note that lag distances in the experimental setting, in contrast to the pure concept of time-based decay, correspond to greater probabilities of trials intervening that are causing the 'dark-side' of learning (retrieval-induced forgetting), by naming semantically-related words for example. Thus one might expect to see similar but much fainter trends as was seen in the semantic interference condition, and indeed the fit resulted in such larger thresholds and smaller drift rates on average, when there are increased lag magnitudes. However, clearly these results are far from conclusive and should be taken with caution, additional accumulation model fits would be necessary to answer this question. We propose that the evidence accumulation modeling approach fit to other experimental data sets that make improvements on the balancing, and number of trials, per level of lag distance over the current one analyzed, may be informative for helping settle the debate of decay or dark-side learning mechanisms in psycholinguistic theory.

7. General Discussion

We demonstrated evidence accumulation to be a worthwhile model for explaining the lexical selection process of lexical retrieval within the RT domain. A number of current psycholinguistic theories consider lexical selection as a process related to selecting a lexical target from a number of lexical alternatives which each have varying activations (or signal supports), that are largely resultant of an initial stimulus recognition. How these activations develop into influencing the decision process of selecting the appropriate lexical target, can be described as a racing evidence accumulation process. We showed how such a process can be directly related to psycholinguistic theory and their simulatory instantiations, e.g. NNMs. We selected a candidate psycholinguistic theory for the demonstration, the DSM. Then the quantitative fit of the evidence accumulation model for lexical selection, driven entirely by the real data, provided for parameter results that confirmed

or supported the lexical selection theory components of the DSM, which prior to this, were supported only by simulation exercises to the data. Then in other aspects, the quantitative application motivated future development of the DSM theory, and supported lexical selection as possible to occur as a non-competitive process.

Evidence accumulation is already supported in a large number of other cognitive domains, as well as the neurosciences. We have argued that the time is ripe to truly weigh in the possibility that evidence accumulation is occurring at the level of lexical selection. We have developed a thorough case for this argument and how evidence accumulation may be implemented in the domain: for example theoretically, embedded within existent or future psycholinguistic theory, or empirically, providing quantitative fits of lexical selection, that better inform psycholinguistic theory.

There are a number of racing accumulation models that already exist which provide a good foundation for meshing with psycholinguistic theory: the race model (LaBerge, 1962; Usher et al., 2002), which features many accumulators of the SWM; the LBA, identical except that evidence accumulates over time without noise in single trials, but there is noise between trials–which perhaps requires fewer trials to fit (Brown & Heathcote, 2008). Then there are also more complicated variants, such as the LCA (Usher & McClelland, 2001), that includes features even more parallel to NNM designs, such as lateral inhibition and decay in the accumulation process, though consequently includes more parameters. Each of these models can be embedded theoretically, or used in a system of layers acting as a network. For the empirical approach however, these racing models will require an adequate number of observations for a quantitative fit; and thus without adequate numbers, these approaches become only simulatory instead of quantitative, just like these major psycholinguistic NNMs (for example, see Dufau et al., 2012).

Given that, it is important to note that unfortunately, most current picture naming data sets lack sufficient observations to quantitatively fit the optimal racing evidence accumulation models. New experiments, which either increase observation sizes, or reduce the number of pertinent lexical alternatives, could be promising to the field. As an interim solution however, we demonstrated that a single accumulator model, the SWM, can be used to model the lexical target activation in the lexical selection process, and in this case, one forgoes directly quantifying the activations of the alternatives. There is additional support for this simplified approach however, where notably, Zandbelt et al. (2014) find that often the activity of a single accumulator, like the SWM, can sufficiently describe the activity of many accumulators operating at the same time.

Our new picture naming data set, is quite large compared to current standards in the blocked-cyclic naming paradigm. Using the SWM empirical approach on the data, we were able to quantitatively fit the model along four factors of the data: semantic context, repetition, lag, and participants. There are two notable ways we would have liked to improve the specificity of the fit. Firstly we would have liked to have enough observations to also capture the variance, and hence fit by, items (pictures), which can be a significant source of variance. At least in this case, fortunately the same 36 pictures were used equally throughout all trials per participant. A second large source of variance could consist of inter-trial effects (Baayen & Milin, 2010; Barr et al., 2013; Wagenmakers et al., 2004). We were only able to capture inter-trial effects coarsely by two lag levels (lags less than or greater than five), and repetition cycles (every six trials, up to 36); if there were enough observations, it would be much more efficient to include a larger range of lag levels, which may have reduced the residuals and helped to extract clearer factor effects.

7.1. Concluding Remarks

We provided the first quantitative evidence accumulation application to semantic (picture/concept) lexical retrieval data, with the SWM. Provided it is a first, additional applications to other data sets are important to determine the replicability and stability of such accumulation parameter trends that were observed in this data set. As mentioned previously, an important future step is to analyze a similar experiment with enough trials that allow fitting each of the repetition and lag levels separately, without a need to combine them into groups of levels. Given this, the results provided herein are hence considered a precursory beginning to the new paradigm of evidence accumulation as a model for lexical selection. However, it is a powerful quantitative approach above the previous analysis of RT means and standard deviations. For example, it is important to note that (as indicated in Figure 1 by Anders et al. in review) the same means and standard deviations of RTs, may be obtained by notably different shapes and onsets of RT distributions; it is hence easier to replicate the RT means across experiments than RT distributions.

Parameters in our suggested modeling approach, that describe the shapes and onsets of the RT distribution, rather than only the means and standard deviations, are hence more specific, and provide richer information for psycholinguistic theory. Furthermore, it should be more difficult to replicate these parameters that describe the distributions of data across experiments, although they provide more information. Therefore, if adequate replication is indeed found, it may be a strong case that the trends found in lexical selection by evidence accumulation, indeed behave in such patterns due to (for example) semantic interference, repetition, lag, and participant types. We hope to have provided a convincing argument, as well as an interesting demonstration, of evidence accumulation as a plausible and useful model for the lexical selection process of lexical retrieval.

Acknowledgements

We acknowledge funding by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant agreement n^{o} 263575), and the Brain and Language Research Institute (Aix-Marseille Université : A*MIDEX grant ANR-11-IDEX-0001-02 and LABEX grant ANR-11-LABX-0036). We thank the "Féderation de Recherche 3C" (Aix-Marseille Université) for institutional support.

References

Alario, F.-X., & Ferrand, L. (1999). A set of 400 pictures standardized for french: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*, 531–552.

Anders, R., Alario, F.-X., & Van Maanen, L. (in review). The shifted Wald distribution for response time data analysis, .

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American Psychologist*, *51*, 355.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*, 12–28.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384.

Balota, D. A., & Yap, M. J. (2011). Moving beyond the mean in studies of mental chronometry the power of response time distributional analyses. *Current Directions in Psychological Science*, *20*, 160–166.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language*, *59*, 495–523.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* volume 3. Wiley New York.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Belke, E., & Stielow, A. (2013). Cumulative and non-cumulative semantic interference in object naming: Evidence from blocked and continuous manipulations of semantic context. *The Quarterly Journal of Experimental Psychology*, *66*, 2135–2160.

Berg, T., & Schade, U. (1992). The role of inhibition in a spreading-activation model of language production. i. the psycholinguistic perspective. *Journal of Psycholinguistic Research*, *21*, 405–434.

Bonin, P., Peereman, R., Malardier, N., Méot, A., & Chalard, M. (2003). A new set of 299 pictures for psycholinguistic studies: French norms for name agreement, image agreement, conceptual familiarity, visual complexity, image variability, age of acquisition, and naming latencies. *Behavior Research Methods, Instruments, & Computers*, *35*, 158–167.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, *340*, 95–98.

Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from? *Cortex*, *26*, 95–122.

Carpenter, R. (1981). Oculomotor procrastination. *Eye Movements: Cognition and Visual Perception*, (pp. 237–246).

Carroll, J. B., & White, M. N. (1973). Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, *25*, 85–95.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*, 63–65.

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*, 417.

Damian, M. F., Vigliocco, G., & Levelt, W. J. (2001). Effects of semantic context in the naming of pictures and words. *Cognition*, *81*, B77–B86.

DeLeon, J., Gottesman, R. F., Kleinman, J. T., Newhart, M., Davis, C., Heidler-Gary, J., Lee, A., & Hillis, A. E. (2007). Neural regions essential for distinct cognitive processes underlying picture naming. *Brain*, *130*, 1408–1422.

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon: Friend or foe? In N. O. Schiller, & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 9–37). Walter de Gruyter.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, *42*, 287–314.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*, 801.

Dell, G. S., Schwartz, M. F., Nozari, N., Faseyitan, O., & Coslett, H. B. (2013). Voxel-based lesion-parameter mapping: Identifying the neural correlates of a computational model of word production. *Cognition*, *128*, 380–396.

Donkin, C., & Van Maanen, L. (2014). Piérons law is not just an artifact of the response mechanism. *Journal of Mathematical Psychology*, *62*, 22–32.

Donner, T. H., Siegel, M., Fries, P., & Engel, A. K. (2009). Buildup of choice-predictive activity in human motor cortex during perceptual decision making. *Current Biology*, *19*, 1581–1585.

Dufau, S., Grainger, J., & Ziegler, J. C. (2012). How to say no to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1117.

Erlich, J. C., Brunton, B. W., Duan, C. A., Hanks, T. D., & Brody, C. D. (2015). Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. *eLife*, *4*, e05457.

Fitzgerald, T. H., Moran, R. J., Friston, K. J., & Dolan, R. J. (2015). Precision and neuronal dynamics in the human posterior parietal cortex during evidence accumulation. *NeuroImage*, *107*, 219–228.

Folks, J., & Chhikara, R. (1978). The inverse Gaussian distribution and its statistical application–a review. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 263–289).

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124.

Foygel, D., & Dell, G. S. (2000). Models of impaired lexical access in speech production. *Journal of Memory and Language*, *43*, 182–216.

Gerstein, G. L., & Mandelbrot, B. (1964). Random walk models for the spike activity of a single neuron. *Biophysical Journal*, *4*, 41–68.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hanks, T. D., Kopec, C. D., Brunton, B. W., Duan, C. A., Erlich, J. C., & Brody, C. D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature*, .

Heathcote, A. (2004). Fitting Wald and ex-Wald distributions to response time data: An example using functions for the S-PLUS package. *Behavior Research Methods, Instruments, & Computers*, *36*, 678–694.

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: Experimental and computational studies. *Cognition*, *100*, 464–482.

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *The Journal of Neuroscience*, *33*, 19434–19441.

LaBerge, D. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27*, 375–396.

de Lafuente, V., Jazayeri, M., & Shadlen, M. N. (2015). Representation of accumulating evidence for a decision in two parietal areas. *The Journal of Neuroscience*, *35*, 4306–4318.

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

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*, 764–766.

Luce, R. D. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. 8. Oxford University Press.

Marques, J. F. (2005). Naming from definition: The role of feature type and feature distinctiveness. *The Quarterly Journal of Experimental Psychology*, *58*, 603–611.

Melinger, A., & Abdel Rahman, R. (2004). Investigating the interplay between semantic and phonological distractor effects in picture naming. *Brain and Language*, *90*, 213–220.

Moss, H., Abdallah, S., Fletcher, P., Bright, P., Pilgrim, L., Acres, K., & Tyler, L. (2005). Selecting among competing alternatives: selection and retrieval in the left inferior frontal gyrus. *Cerebral Cortex*, *15*, 1723–1735.

Mulder, M., Van Maanen, L., & Forstmann, B. (2014). Perceptual decision neurosciences–a model-based review. *Neuroscience*, *277*, 872–884.

Mulder, M. J., Keuken, M. C., Van Maanen, L., Boekel, W., Forstmann, B. U., & Wagenmakers, E.-J. (2013). The speed and accuracy of perceptual decisions in a random-tone pitch task. *Attention, Perception, & Psychophysics*, *75*, 1048–1058.

Nakahara, H., Nakamura, K., & Hikosaka, O. (2006). Extended LATER model can account for trial-by-trial variability of both pre-and post-processes. *Neural Networks*, *19*, 1027–1046.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*, 1729–1735.

Oldfield, R., & Wingfield, A. (1964). The time it takes to name an object. *Nature*, *202*, 1031–1032.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*, 227–252.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*, 53.

Protopapas, A. (2007). Check vocal: A program to facilitate checking the accuracy and response time of vocal responses from dmdx. *Behavior Research Methods*, *39*, 859–862.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*, 190.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261.

Ricciardi, L. M. (1977). Diffusion processes and related topics in biology, .

Roelofs, A. (2008). Dynamics of the attentional control of word retrieval: Analyses of response time distributions. *Journal of Experimental Psychology: General*, *137*, 303.

Schnur, T., Lee, E., Coslett, H., Schwartz, M., & Thompson-Schill, S. (2005). When lexical selection gets tough, the lifg gets going: A lesion analysis study of interference during word production. *Brain and Language*, *95*, 12–13.

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

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2009). Localizing interference during naming: convergent neuroimaging and neuropsychological evidence for the function of Broca's area. *Proceedings of the National Academy of Sciences*, *106*, 322–327.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and language*, *54*, 228–264.

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, *86*, 1916–1936.

Taft, M., & Forster, K. I. (1975). Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, *14*, 638–647.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*, 179.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*, 550.

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, *46*, 704–715.

Van Casteren, M., & Davis, M. H. (2006). MIX, a program for pseudorandomization. *Behavior Research Methods*, *38*, 584–589.

Van Maanen, L., Grasman, R. P., Forstmann, B. U., Keuken, M. C., Brown, S., & Wagenmakers, E.-J. (2012a). Similarity and number of alternatives in the random-dot motion paradigm. *Attention, Perception, & Psychophysics*, *74*, 739–753.

Van Maanen, L., & Van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research*, *8*, 174–181.

Van Maanen, L., Van Rijn, H., & Borst, J. P. (2009). Stroop and pictureword interference are two sides of the same coin. *Psychonomic Bulletin & Review*, *16*, 987–999.

Van Maanen, L., Van Rijn, H., & Taatgen, N. (2012b). Race/a: An architectural account of the interactions between learning, task control, and retrieval dynamics. *Cognitive Science*, *36*, 62–101.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*, 37–58.

Vickers, D. (1979). *Decision processes in visual perception*. Academic Press.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of 1/fα noise in human cognition. *Psychonomic Bulletin & Review*, *11*, 579–615.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140–159.

Winkel, J., Van Maanen, L., Ratcliff, R., Van der Schaaf, M. E., Van Schouwenburg, M. R., Cools, R., & Forstmann, B. U. (2012). Bromocriptine does not alter speed–accuracy tradeoff. *Frontiers in Neuroscience*, *6*.

Zandbelt, B., Purcell, B. A., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2014). Response times from ensembles of accumulators. *Proceedings of the National Academy of Sciences*, *111*, 2848–2853.

Appendix

Detailed Exeriment Methods

Participants

Twenty-four native French speakers between 18 and 30 years old (20 females), which were students at the Université de Provence, participated in the experiment for course credit. The data of one participant was removed from the analysis due to technical difficulties during voice recording. Each participant declared having no language disturbance, and normal or corrected-to-normal vision.

Materials

A total of 36 common nouns, each with a corresponding black and white picture, were selected. They were issued from six semantic categories (e.g. animals, vehicles). These pictures were drawn from published collections (Bonin et al., 2003; Alario & Ferrand, 1999), and if no such picture was available from prior research, they were drawn from the Internet, or created by the experimenters. The name agreement of each picture was very high. Their dimensions consisted of 245 pixel width, by 240 pixel height, and were presented at the center focus of the computer screen display.

Design

A 6×6 design was used in which six items (pictures) were available for each of six different semantic categories. There were four semantic context conditions divided into four types of blocks. The level of context, or semantic relatedness by category membership, of each block was manipulated, and involved four increasing levels: ranging from no shared category membership, to fully-shared category membership.

In homogenous blocks, all items belong to the same semantic category and this condition is termed "61," where 6 refers to the number of items per category, and 1 refers to the number of semantic categories involved in the condition. In heterogeneous blocks, all items are from different semantic categories, and so this condition is then termed "16." Then there

are also two intermediary conditions: “32,” involving three items from one semantic category and three from another; and condition “23,” with 2 items issued from each of three different semantic categories.

Thus the order of increasing semantic relatedness for the conditions, from full to none, is: 61, 32, 23, and 16. Thus in these condition names, the first digit always corresponds to the number of items per category, and the second corresponds to the number of categories represented. The heterogeneous and intermediary conditions were created by mixing the members of the semantic categories. Three different master lists were created so all possible category combinations could be present in the intermediary conditions.

In each master list, all 36 items participated equally in each of the four conditions, and care was taken to avoid the presence of more than two picture names starting with the same phonological onset (defined in respect to manner of articulation of the first phoneme: voiceless occlusive, voiced oral or nasal occlusive, and liquid, voiced or voiceless constrictive), or more than two picture names with the same rhyme within the same block. Furthermore, the order in which the pictures were presented was pseudo-randomized for each participant, such that identical pictures were always at least five items apart, and there were never four consecutive items with the same number of syllables; this was done using specialized randomization software, called MIX (Van Casteren & Davis, 2006).

Procedure

Participants were tested in a sound-attenuated dimly-lit room. The experiment was controlled by the software DMDX (Forster & Forster, 2003), which allows on-line recording and voice-key triggering of the participants’ verbal responses. Participants were first familiarized with the 36 pictures used in the experiment. The pictures were presented one by one in a random order, and the participant was asked to name each one of them. The experimenter made verbal corrections when an incorrect or unexpected response was produced. The microphone sensitivity was tested and adjusted to the voice of the participant during this familiarization phase. Then, the experimental instructions were delivered and the experiment started. A trial consisted of the following events: (1) a fixation point (“plus” sign presented at the center of the screen) for 500 ms; (2) a picture, which remained on the screen until the participants responded or until a 1300 ms deadline was reached; (3) an intertrial blank screen with a randomized duration between 166 to 666 ms. The following trial started automatically. Participants were instructed to name each presented picture as fast and accurately as possible. Each picture was repeated 7 times per condition. The first of seven sets established a warm-up trial while the remaining six sets consisted of the experimental data. Indeed the focus of this study was not on the transient facilitation effect that tends to take place in the first set, which has been repeatedly, if not consistently, reported in other published studies (Belke & Stielow, 2013; Melinger & Abdel Rahman, 2004). Altogether, there were 1008 trials per participant, composed of 144 as warm-up, and 864 for analysis. The whole experiment lasted about an hour. There were short breaks between each block of 42 pictures.

Data Post-processing

A total of $19,872$ response times were recorded across all participants after warm-up, which consisted of 98.2% correct responses and 1.8% error responses. Each of the 23 participants performed 864 trials post-warm-up, and on average made only 16 error responses.

The accuracy of the reaction time measures provided by the voice-key was checked visually offline, and corrected when necessary using the software CheckVocal (Protopapas, 2007). Trials were excluded from the analysis if the participant did not respond or produced any kind of verbal error (partial or complete production of incorrect words, verbal dysfluencies: e.g., stuttering, utterance repairs, hesitations), leaving $19,506$ trials for analysis.