A delay in processing for repeated letters: Evidence from megastudies

Iliyana V. Trifonova and James S. Adelman

University of Warwick

Corresponding author:

Iliyana V. Trifonova

Department of Psychology

University of Warwick

COVENTRY

CV4 7AL

United Kingdom.

I.V.Trifonova@warwick.ac.uk

Abstract

Repetitions of letters in words are frequent in many languages. Here we explore whether these repetitions affect word recognition. Previous studies of word processing have not provided conclusive evidence of differential processing between repeated and unique letter identities. In the present study, to achieve greater power, we used regression analyses on existing mega-studies of visual word recognition latencies. In both lexical decision (in English, Dutch, and French) and word naming (in English), there was strong evidence that repeated letters delay visual word recognition after major covariates are partialed out. This delay was most robust when the repeated letters occurred in close proximity but not in immediate adjacency to each other. Simulations indicated that the observed inhibitory pattern of repeated letters was not predicted by three leading visual word recognition models. Future theorizing in visual word recognition will need to take account of this inhibitory pattern. It remains to be seen whether the appropriate adjustment should occur in the representation of letter position and identity, or in a more precise description of earlier visual processes.

Keywords: visual word recognition; repeated letters; regression; megastudies; computational modeling; letter processing

1. Introduction

Reading alphabetic languages requires the successful identification of letters and their position within the word. In this way, the perceptual system discriminates between lexical units that bear strong form resemblance. It can determine the difference between two words with the same length but differing by a single letter ("orthographic neighbors", such as *farm– form*; Coltheart, Davelaar, Jonasson, & Besner, 1977), have the same letters but in a different order (*from-form*), or have different length, but many common letters (*though - through*). One of the goals of orthographic processing research has been to explain how this discrimination is achieved and how bottom up sublexical processes such as encoding of letter identities and their position mediate recognition of the whole word unit. The empirical results have motivated the development and revision of visual word recognition models in which those initial perceptual stages are implemented in their encoding schemes. These schemes determine the models' predictions regarding the word candidates that are considered and, ultimately, the factors affecting lexical selection. An example of such a factor that will be on focus in this work is the presence of letter repetitions in words.

Letter repetitions are frequent in many languages, especially in words with more than one syllable. Understanding how strings with repetitions are processed is therefore vital for understanding reading. The exploration of repeated letter units could be informative as to how the visual system processes identical elements in early stages of word recognition and is intrinsically related to the letter identity and letter position encoding. In this context, repeated letter effects have received relatively little attention in the orthographic processing literature which has been more focused on the problem of letter positional uncertainty (e.g. Kinoshita & Norris, 2009; Lupker, Perea, & Davis, 2008; Perea & Lupker, 2003; 2004; Peressotti & Grainger, 1999; Schoonbaert & Grainger, 2004; Van Assche & Grainger, 2006; Welvaert, Farioli, & Grainger, 2008) and the need to incorporate a mechanism for positional ambiguity in the encoding schemes of the visual word recognition models (e.g. Davis, 2010; Gomez, Ratcliff & Perea, 2008; Grainger, Granier, Farioli, Van Assche, & van Heuven, 2006; Grainger & van Heuven, 2003; Norris & Kinoshita, 2012) .

The problems of letter repetition and positional uncertainty might be deeply intertwined: The initial ambiguous location of a single letter *a* might well be perceived as two letters *a* in two distinct locations and vice versa. Exploring the repeated letters effect in more depth might therefore have important implications for further developing the encoding schemes of the visual word recognition models. We will therefore first review some relevant computational models (Davis, 2010; Grainger & van Heuven, 2003; McClelland & Rumelhart, 1981; Norris, 2006) and their encoding schemes with an emphasis on the

processing of letter repetitions and its consequences. We will further review the relatively inconsistent previous experimental evidence for repeated letter effects. This will motivate an alternative approach to the problem of repeated letters: regression analysis on response times from large databases of visual word recognition tasks (*megastudies*). The results of the regression analysis on the empirical data will then be compared with the predictions of the discussed computational models regarding effects of letter repetitions.

Research on letter position encoding schemes is often motivated by a comparison with the one of the most influential models in visual word recognition, the Interactive-Activation (IA) model (McClelland & Rumelhart, 1981). This model explains lexical selection by means of spreading activation in a cascaded and interactive manner between representations located in three levels of a hierarchical structure (feature, letter, word). Word nodes that are consistent with the perceptual input get activated, while inconsistent ones get inhibited. Once activated, word candidates suppress each other through lateral inhibitory links and compete until the activity of a single word reaches an activation threshold associated with lexical selection. This model has a slot-based scheme in which every letter is assigned to a specific slot - which defines its position within the word - and projects its activity only with regard to this slot. The letter *a* in first position will activate words containing the letter *a* in first position and will inhibit others that start with a different letter. It will not activate words in which *a* is in another position. This model could therefore not predict any effect of letter repetition since its slot-based encoding scheme treats two identical letters in different positions as two different letters. The IA model could explain how the reading system could discriminate between words with equal length and different letter identities in some of the positions but is not in accordance with evidence from the masked priming literature suggesting that the perceptual system has a considerable degree of positional tolerance.

Contrary to the prediction of the IA model (McClelland & Rumelhart, 1981), primes formed by letter transpositions (anwser-ANSWER) produced as strong priming effects as those identical to the targets (Forster, Davis, Schoknecht, & Carter, 1987). In addition, letter strings formed by transposing letters from a base target word have been demonstrated to be orthographically more similar to the target than strings with replaced letters in the corresponding mismatched positions. This finding has been observed with adjacent transpositions, such as *jugde* from the target *judge,* as opposed to the replaced letter control prime *jupte* (Perea & Lupker, 2003), and has been extended to nonadjacent cases (e.g., caniso–CASINO; Lupker, Perea, & Davis, 2008; Perea & Lupker, 2004). The slot-based scheme was also falsified by studies demonstrating a priming effect with relative position primes in which the absolute order of the letters was disrupted by either letter deletions (e.g., blcn-BALCON, Grainger et al., 2006; Peressotti & Grainger, 1999), or insertions in the primes (e.g., juastice–JUSTICE; Van Assche & Grainger, 2006;

Welvaert, Farioli, & Grainger, 2008). These results suggest that an absolute position specific encoding in the slot-based scheme is inaccurate and motivated the proposal of alternative letter position and identity encoding mechanisms.

Several interactive activation framework models with different encoding schemes were later developed. In this work, the focus will be on two of them: The relative position parallel open bigram model (Grainger & van Heuven, 2003) and the Spatial Coding Model (Davis, 2010), which are both based on the interactive-activation architecture but have entirely different schemes with which they account for the transposed letters and relative position priming effects. In the parallel open bigram model, the letter level representations from the IA model are replaced by representations of open bigrams. They are in the relative position map level, at which the letter positions are encoded after letters have been initially identified. Position encoding is achieved through activating bigram units containing information of the position of one letter relative to the others, or more specifically, whether a letter is located to the left or to the right of the other letters in the word (up to two intervening letters). The word SILENCE is therefore represented by the open bigrams SI, SL, SE, IL, IE, IN, LE, LN, LC, EN, EC, EE, NC, NE, CE. In analogy to the letter nodes in the IA model, the open bigram nodes are connected to orthographic word representations with excitatory and inhibitory links. Bigrams consistent with the input get activated and feed forward to the word nodes. Words containing active bigrams receive activation, while those lacking consistent bigrams get inhibited. As in the IA model, lexical competition mechanisms are implemented by inhibitory lateral connections between the nodes at the word level. In this encoding scheme, transposed letter and relative position primes will contain most of the targets' constituent representations in the form of open bigrams. These primes will pre-activate targets more strongly than corresponding control primes that contain fewer consistent bigram units. In this way, the model could successfully simulate both transposed letter and relative position priming effects. When stimuli contain letter repetitions, fewer bigrams will be activated than with stimuli composed of unique letters, and therefore the expectation is that the open bigram model would be likely to predict inhibitory effects of repeated letters (as noted by Schoonbaert & Grainger, 2004).

The Spatial Coding Model, unlike the relative position open bigram model (Grainger & van Heuven, 2003), retains the letter representations in its architecture. Unlike the IA model, however, the letter position and identity encoding scheme is not channel specific. In the Spatial Coding Model, a consistent letter of the input could increase the activity of a word node containing that letter, even if it appears in a different position. An important conceptual difference between the Spatial Coding Model and the other two previously described models is that the same (letter) units encode letter identities in different positions, and so the set of letter units involved in representing *from* and *form* will be identical, unlike

IA's channel specific scheme, in which *r* in position 2 is different from *r* in position 3, or the open bigram model which will encode these words with two nonidentical sets of bigram representations (FR, FO, FM, *RO*, RM, OM and FO, FR, FM, *OR*, OM, RM, respectively). The open bigram model, therefore, has a local-context dependent encoding scheme and uses different set of representations to encode words containing the same letters, but in different positions (such as anagrams). Word recognition modeling with contextual representations such as open bigrams was also implemented by Whitney in her SERIOL model (Whitney, 2001). This model was developed independently from the relative position open bigram model and differed in several ways. One major difference was the serial assumption of information acquisition as opposed to the parallel one in the model of Grainger and van Heuven. In SERIOL, there is a location gradient across letter features which affects the accumulation of letter information. The activation of the features decreases from left to right and determines the activity in the letter level. The leftmost letters receive sufficient activation earlier and fire in a sequential manner. Letter order information is thus established through the temporal firing pattern of the letters. This temporal positional encoding is then translated into nontemporal one through the means of open bigrams. Unlike the model of Grainger and van Heuven, in SERIOL the bigrams are generated in a serial manner in which the bigrams containing the leftmost letters are available earlier than the ones containing the rightmost letters in the string. Despite their differences, both SERIOL and the model of Grainger and Van Heuven have a local-context specific encoding scheme. The Spatial Coding Model (Davis, 2010), on the other hand, has a local-context independent encoding scheme, in which the same letter representations will be activated in cases of anagrams.

In the Spatial Coding Model, the letter positions in word representations are represented by spatial patterns. To incorporate letter position uncertainty, each letter position code of the input stimulus has the shape of a normal distribution, rather than a single value, with its spread representing positional uncertainty. The position code of the letter in the stimulus is assigned dynamically after a rapid left to right serial scan. The spatial pattern of the input is compared to the spatial pattern of a stored word representation, a procedure called superposition matching. The matching algorithm includes computing the signal-weight difference functions for each letter of the word representation, sums up all the difference functions and finally divides the obtained peak of the superposition function by the length of the word representation. In the case of an identity prime, a perfect match score of 1 is obtained, as all the difference functions are perfectly aligned with a mean of 0 (and peak value of 1 in the simplified case of no identity uncertainty). In the cases of transposed letter primes, not all the difference functions are perfectly aligned to 0, and so in the cases of the positional mismatches, the individual letter peaks do not align with those of the other letters in the stimulus. Smaller, non-central, values will be added to the superposition function, resulting in a lower overall peak for the superposition function, but the total match

score is still high. Relatively high match scores will also be calculated for primes with (few) insertions or deletions. Separate mechanisms in the model's architecture are responsible for the penalization in cases of length mismatch between the word representation and the input, and for inhibition from inconsistent stimulus letters.

The Spatial Coding Model has an explicit mechanism that deals with repeated letter cases. It uses bins of clones of letter receptors that interact and cooperate with the final goal of achieving the maximum match score between the input and a word representation while not allowing for the same letter unit (repeated or not) to contribute more than once for this calculation. With this mechanism, the Spatial Coding Model effectively treats repeated letters as different items and is unlikely to predict any effect of letter repetition, unless the effect occurs due to other uncontrolled lexical or sublexical factors.

Given these models' different mechanisms relate to repeated letters, comparing simulations of these models to data on repeated letters could be informative for both repeated letter processing *per se* and the conceptual debate of whether the encoding of letter position and identity should have a local-context dependent element or not. Should the encoding of additional letter *a* depend on whether *a* is already present in the string or not?

This is important, because as had been noted, although the manner at which letters can be combined to form words is quite rich, letter repetitions are common. This is particularly true for words with more than one syllable, in which the likelihood of observing a repeated letter identity is higher, especially in the cases of vowels or high frequency consonants. Word recognition experiments targeting letter repetitions have produced mixed and inconclusive results. Schoonbaert and Grainger (2004) used the masked primed lexical decision task to compare nonword primes formed by deletion of either repeated or unique letters (*balace* vs *balnce* from the target BALANCE) and no difference in priming was found. Nor was a repeated letter effect found in their subsequent experiment, in which they used the items that had served as primes as nonword targets: Items formed by deletion of a repeated letter were rejected as quickly as those formed by deletion of a unique letter. However, the design of their masked-priming experiment also included a between-target manipulation of letter repetition: Both words and nonwords containing letter repetition took significantly longer to recognize than items without repeated letters.

In another masked-priming lexical decision study, with prime manipulations using insertions, rather than deletions in primes (Van Assche & Grainger, 2006), a difference was not found between three related prime conditions, some of which containing repeated letters. They were constructed by doubling a letter in the target (jusstice-JUSTICE), inserting a letter already present in the target in another nonadjacent position (justisce), and inserting a different letter in the target (juastice). These related primes all

produced the same priming effect relative to an unrelated control. In sum, the masked-priming lexical decision procedure of these studies did not provide evidence of differential processing of repeated and unique letter identities within words.

Results from two other studies, however, suggest that the presence of letter repetition could affect processing difficulty. Gomez et al., (2008) reported a two-alternative forced choice perceptual identification experiment in which nonword letter strings were recognized significantly less accurately than strings without repeated letters. The authors interpreted these results as an evidence that repeated letters were more difficult to perceive causing preference towards a foil with no repetition. Norris, Kinoshita and van Casteren (2010) argued that an effect of letter repetition was probably not observed in the Schoonbaert and Grainger (2004) masked-primed lexical decision study due to the lack of sensitivity of the task as well as the use of long stimuli with nonadjacent repetitions. Using shorter stimuli, Norris et al. demonstrated a stronger priming effect in the cases in which a two-replaced-letter primes were constructed by doubling a letter from the target (uueer-UNDER), than using two different letters (ulger-UNDER), with a masked-priming same-different task but not with lexical decision task. Again with shorter stimuli, they also showed that deletion of an adjacent repeated letter (anex-ANNEX), has a smaller disruption of the form priming effect than deletion of a unique letter (eupt-ERUPT), suggesting differential cost of deleting a repeated versus deleting a unique letter from the target. The authors interpreted these results as evidence of imprecise position encoding at early stages and "leakage" of letter identities to nearby positions, enhancing priming in cases of adjacent repetitions.

To support the identity "leakage" explanation of the obtained advantage of the repeated substitution primes over the nonrepeated ones, Norris et al., (2010) simulated these experiments with the Bayesian Reader model of word recognition (Norris, 2006) with an extension for positional uncertainty. The Bayesian Reader is a model from a different class than the interactive activation framework models described so far. Like the IA model (McClelland & Rumelhart, 1981), however, the original version of the Bayesian Reader had a position and length-specific encoding scheme. In the Bayesian Reader, words are represented as points in multidimensional space. Word coordinates are implemented by concatenating vectors which encode the presence or absence of a letter in a specific position, thus effectively acting as slots. The original model incorporates only letter identity uncertainty by adding Gaussian noise to each of the words' coordinates. Unlike the Interactive Activation Model and its successors, word recognition is not achieved through spreading activation mechanisms. Rather than increasing the activation of word representations consistent with the input, word recognition is implemented by gradual accumulation of noisy perceptual evidence and optimal mapping of the noisy input to one of the lexical entries. In the Bayesian Reader framework, readers are described as ideal observers who use Bayesian inference to

combine perceptual evidence with prior linguistic knowledge (word probabilities) to perform a word recognition task in an optimal way. With the accumulation of evidence, word likelihoods are calculated, that is, the probability of observing the input given that a particular word was presented. This calculation is based on the variance of the input distributions and the distance between the mean coordinates of the input distributions at a given time and the ideal (average) coordinates of each word in the perceptual space. The posterior word probability, that is, the probability that the input was generated by a particular word, is then calculated for each of the lexical entries from the prior word probabilities (their frequency) and the likelihoods of the words.

Though recognition through accumulation of noisy evidence is one of the basic assumptions of the Bayesian Reader, as a simplification, its original version implemented only identity uncertainty. The extended version described by Norris et al., (2010) included positional uncertainty, but the model's encoding scheme remained length specific. The positional ambiguity in the sampling process was achieved by randomly drawing letter positions from normal distributions centred over the real letter positions rather than being represented as a single value (for which the authors draw parallels with Gomez et al.'s, 2008 model). Although with this extension, the Bayesian Reader successfully simulated the repeated letter advantage in substitution primes, thus supporting the "leakage" mechanism of letter identities described by the authors, the model could not simulate the deletion prime results from the same study due to its length-specific encoding scheme. The length-specific limitation was later addressed by Norris and Kinoshita (2012) who proposed a different version of the model that also included calculations of likelihoods of letter insertions and deletions that affected the calculation of the posterior word probabilities.

The results of Norris et al. (2010) provided evidence that the repetition of letter identity might play a role in the recognition of the whole word unit. They also demonstrated that such an effect could be simulated successfully with the Bayesian Reader model of word recognition (Norris, 2006) when the model incorporates positional uncertainty. Furthermore, the study of Norris et al. (2010) raised an important methodological concern regarding the choice of approach to problems such as repeated letters. They could have confirmed some methodological issues in investigating repeated letter effects, such as lack of sensitivity of the masked-primed lexical decision task for researching a phenomenon occurring at early perceptual stages and possibly susceptibility to top-down lexical influences. It is not, however, immediately clear how the mechanism of leakage they propose could explain the results of Gomez et al. (2008) that suggest that repeated letter targets are harder to process as well as the overall target type effect reported by Schoonbaert and Grainger (2004) also hinting at inhibitory, rather than facilitatory effect of letter repetition. It is also not clear how such a leakage mechanism would affect nonadjacent repetitions

with longer distances between the repeated identities and whether a repeated letter effect will still be present in longer items, in which it is in fact more common for a repetition to occur.

A methodology that could be effective for the investigation of repeated letter effects for several reasons is a regression approach on megastudies data containing reaction times of visual word recognition tasks in several languages - English, Dutch, and French (e.g., Balota et al., 2007; Brysbaert, Stevens, Mandera, & Keuleers, 2016; Ferrand et al., 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012). The English, British, Dutch 2 and French Lexicon projects (ELP, BLP, DLP2, and FLP) contain disyllabic and multisyllabic words in which letter repetitions occur often. The number of repeated letter observations could provide greater power for observing a possible effect than that in previous studies exploring repeated letters using factorial designs. The regression approach could also overcome other limitations of factorially designed studies such as correlation problems (as factors could be covaried out), particularly problematic in visual word recognition, experimental biases in the stimuli selection, list effects, dichotomization of continuous variables (see Balota, Yap, Hutchison, & Cortese, 2012 for a detailed discussion of the advantages of this methodology over factorial designs). It will also allow for a direct cross linguistic comparison of the effect, exploring whether effects generalize across the processing of different languages or is linked to the idiosyncrasies of individual languages.

Additionally, due to the large number of items, the regression approach allows a deeper investigation of the effect. This includes exploration of repetition effects of different distances (i.e., adjacent repetitions, nonadjacent repetitions with various number of intervening letters) by adding separate predictors encoding the presence or absence of a repetition within each distance. Additional evidence in this direction could help reconcile inconsistencies in the literature, such as differential effects observed with primes with deleted repeat vs unique letters in adjacent repetitions (Norris et al., 2010), and the absence of that effect in the case of nonadjacent repetitions (Schoonbaert & Grainger, 2004).

Another theoretically motivated question relates to possible dependence of the repetition effect on the consonant-vowel status of the repeated letters. There is evidence suggesting a differential role of consonants and vowels in word processing and the importance of the word CV structure at initial stages of lexical access (e.g., Acha & Perea, 2010; Chetail, Balota, Treiman, & Content, 2015; Chetail, Drabs, & Content, 2014; Chetail, Treiman, & Content, 2016; Lupker et al., 2008). Moreover, vowel letters and vowel repetitions are more common (see Table 1) and their higher frequency could modulate any letter repetition effect. To investigate possible differences in the processing of repeated consonants and vowels, with so many items, separate variables can be constructed for each letter type.

The present study explored whether the presence of letter repetitions affects visual word processing difficulty and whether repetition distance and consonant-vowel status moderate this effect. The investigation was conducted with the regression approach on response times from visual word recognition tasks in English, Dutch and French. Simulations of the open bigram model (Grainger & van Heuven, 2003), the Spatial Coding Model (Davis, 2010) and the Bayesian Reader model (Norris & Kinoshita, 2012) were also included with the purpose of evaluating the models' predictions for the effect and the plausibility of their letter position and identity encoding schemes. Norris et al. (2010) have already addressed the problem of repeated letters with adjacent repetitions and have demonstrated empirically and computationally facilitatory effect of repeated letters in primed word recognition tasks. More extensive simulations on megastudy data will clarify of whether the repeated letter effect in the Bayesian Reader could be extended to unprimed lexical decision task within different distances between the repeated letters.

2. Method

2.1. Dependent Variables

The mean correct item latencies across participants of the lexical decision task were obtained from the English Lexicon Project (Balota et al., 2007), the British Lexicon Project (Keuleers et al., 2012), the Dutch Lexicon Project 2 (Brysbaert et al., 2016), and the French Lexicon Project (Ferrand et al., 2010). Word naming latencies were acquired from the English Lexicon Project. We used the raw unscaled response time (RT) measures to facilitate the interpretation of the results.

2.2. Independent Variables

2.2.1. Variables of Interest (Repeated Letters)

The variables of interest were constructed with the purpose of observing the effect of repetition of letter identity on visual word recognition.[1] Two factors were considered while calculating the variables. These were the consonant-vowel status of the repeated letter and the distance between the letter repetition, or how far from each other the two letters with the same identity were. The repetition distance was measured

[1] The R codes including methods and results are available in the Trifonova & Adelman (2019) dataset.

by the number of intervening letters between the repeated ones. There were separate variables for each possible repetition distance. Such a division allowed us to explore whether the letter repetition effect was dependent on the distance between letters with same identities. The variables of interest therefore encoded all the possible instances in which a consonant or a vowel letter could be repeated within a certain distance. Each repeated letter variable represented the number of times a consonant or vowel repetition within a specific distance occurs in a word. Here are several examples with English words and their corresponding repeated letters values. The variable *repetition of consonants with distance 0,* summed the number of repeated consonant letters with no intervenor (adjacent repetition), such as *c* in the word *accept* (aCCept) and *d* and *s* in *address* (aDDreSS). For these words the variable had the values 1 and 2, respectively. The variable *consonant repetitions with distance 1* (one intervening letter between the repeated ones) had a value 1 for the word *coconut* (CoConut), as one letter, *c*, is repeated within that distance, and 2 for the word *suspended (*SuSpenDeD), in which both *s* and *d* appear twice. The variable with a *consonant repetition with distance 2* had the values 1 for *hundred* (hunDreD) and 2 for *accountant* (accouNTaNT) and so on. The vowel repetition variables were constructed in the same way. If no repetition was present for a certain condition, the corresponding variable had a value of 0. Words in which one letter appeared more than twice and therefore had two or more possible distances between repetitions of the same letter were discarded from the analyses to avoid additional complexity.

The repeated letter variables were constructed in this way for all languages under investigation, with two variations in French. The diacritic-sensitive version took into account diacritic marks and treated letters with and without diacritics as different. With this algorithm, there was no letter repetition in *zèbre*. The diacritic-insensitive version counted letter repetitions only after all diacritic marks were removed from the items; the rare diacritics in English and Dutch were always removed. This calculation encoded a repetition of *e* in *zebre* as it was not sensitive to the presence of the diacritic mark. The two separate calculations were performed to provide an appropriate baseline for the evaluation of visual word recognition models with no implementation of diacritic marks in their encoding schemes. Comparing the effect of letter repetition in behavioral data and in models' simulations therefore required diacritic insensitive repeated letter measures. Apart from employing two separate algorithms that treated diacritics differently, the repeated letters variables were constructed in the same way as for the analyses of the other lexicon projects.

2.2.2. Control variables

Due to the fact that the investigation was crosslinguistic and the three languages (English, Dutch, French) had their own idiosyncrasies, the list of covariates was not identical across the different lexicons.

However, care was taken so that important control factors were included in the regression models of each of the languages.

English Lexicon Project. The control variables obtained from the English Lexicon Project (Balota et al., 2007) were: logarithmically transformed subtitle (SUBTLEX-US; Brysbaert & New, 2009) contextual diversity and word frequency measures; number of letters; orthographic neighborhood size (number of orthographic neighbors, Coltheart N; Coltheart et al., 1977); phonological neighborhood size (the number of words differing by a single phoneme); Levenshtein orthographic distance (OLD20, the average orthographic Levenshtein distance of the 20 nearest neighbors, Yarkoni et al., 2008); Levenshtein phonological distance (PLD20); number of morphemes; as well as two different measures of bigram frequency: mean bigram frequency and bigram frequency by position. Apart from avoiding confounds in the behavioral regression results, bigram frequencies were important controls as the predictions of one of the word recognition model under evaluation could be sensitive to these measures (the relative position open bigram model; Grainger, & van Heuven, 2003). To control for any possible effects of letter frequency, we calculated and included in the model mean type 1-gram position specific and position nonspecific frequencies based on the frequency of occurrence of each letter in the English Lexicon word list. The quadratic term of the number of letters variable was also included as a predictor (New et al., 2006; Yap & Balota, 2009).

In addition to the control factors provided in the English Lexicon Project (Balota et al., 2007), several additional phonological variables were constructed and added to the regression models, based on phonological transcriptions acquired from CELEX (Baayen, Piepenbrock, & van Rijn, 1993). Words that were not found in the CELEX database were excluded from the analysis. The additional variables were: first phoneme, primary lexical stress position, number of phonemes, number of syllables, and several phonological consistency measures. The first phoneme variable was entered as a categorical variable whose levels were each of the possible phonemes that could appear as an initial sound in English. The primary lexical stress position variable was dummy coded and reflected the stressed syllable in a word, primary stress on first syllable serving as a baseline. The consistency measures reflected the consistency of mapping of print to sound and included feedforward onset, feedforward rime, feedback onset and feedback rime consistency of the first syllable as well as four composite measures of the same type which represented the mean consistency across all the syllables in a word (see Yap & Balota, 2009, for a detailed discussion of consistency measures). As the construction of the consistency measurements depended on the syllabification of the words, special care was taken so that orthographic and phonological syllabifications matched before performing the calculations. In the cases of inconsistent syllabification between the phonological and orthographic forms, the orthographic syllable was adjusted

to the phonological one. The ratio between the orthographic Levenshtein distance (OLD20) and the phonological Levenshtein distance (PLD20) was also included as a separate measure of phonological consistency (Yap & Balota, 2009). Heterophonic homograph entries such as *bow* that had multiple pronunciations for the same orthographic form and therefore multiple values of the phonological variables (phonological consistencies, phonological neighborhood, stress pattern) were not included in the analyses (N = 370).

British Lexicon Project. The control variables obtained from the British Lexicon Project were: two different measures for orthographic neighborhood size (Coltheart N and OLD20), number of letters, as well as the number of syllables in a word. The quadratic term of the number of letters variable was also included as a predictor (New et al., 2006; Yap & Balota, 2009). The logarithmically transformed word frequency measures in Zipf scale and the contextual diversity measures were obtained from SUBTLEX-UK and were also added as control variables in the regression analyses (van Heuven, Mandera, Keuleers, & Brysbaert, 2014).

As in the English Lexicon Project (Balota et al., 2007) analyses, the additional phonological variables, calculated based on the phonological transcription in CELEX (Baayen et al., 1993) were also included as control predictors. These were first phoneme, primary lexical stress, number of phonemes, all first syllable and composite phonological consistency measures. In addition, the phonological neighborhood size and the average phonological Levenshtein distance of the 20 nearest neighbors (PLD20) were calculated using the *vwr* package (Keuleers, 2015) as implemented in R version 3.4.1 (R Core Team, 2017). The ratio between OLD20 and PLD20 was included as an additional phonological consistency measure (Yap & Balota, 2009). Heterophonic homographs were not included in the analyses (N = 298).

In addition, mean type-based bigram frequency was calculated by counting the number of times a bigram appears in all English words from the CELEX database (Baayen et al., 1993), regardless of bigram position and word length (see Westbury & Buchanan, 2002, for the description of a similar measure). For example, the bigram frequency of ac was increased after encountering both *back* and *act*. The frequencies of all bigrams in a word were then summed and divided by the number of word letters minus one. The *vwr* package in R (Keuleers, 2015) was used for acquiring the list of words. The mean positional bigram frequency was calculated in a similar way, the only difference being that the bigram counts were performed for the specific bigram position, rather than for all positions. This measure was bigram position, but not word length specific, i.e. all words that contained the bigram in the specific position (relative to the beginning of the word) contributed to its count. The bigram *ac* in position 1 was counted in both *act* and *action*, but not in *back*, where *ac* in position 2 was counted instead. Mean positional and

nonpositional 1-gram frequencies were calculated in an analogous manner and were included to control for possible letter frequency effects.

To control for possible morphological effects, two morphological variables were included in the analyses. The first variable was constructed by counting the number of morphemes after immediate segmentation of the lemma. The second variable represented the number of elements after the inflectional transformation of the wordform. The morphological database from CELEX (Baayen et al., 1993) was used for the construction of these variables.

Dutch Lexicon Project 2. The control variables obtained from the Dutch Lexicon Project 2 were number of letters, number of syllables, SUBTLEX2 word frequency (added as a control after a logarithmic transformation), number of phonemes, orthographic Levenshtein distance (OLD20), the phonological Levenshtein distance provided in the lexicon (PLD30)[2], Coltheart N. As this lexicon also provided ratings of concreteness and age of acquisition, these were also added as control variables.

In addition, bigram and 1-gram frequency measures were calculated in an identical way as for the analyses of the British Lexicon Project. Mean bigram and 1-gram frequencies and positional bigram and 1-gram frequencies were therefore included in the model. The number of elements after immediate segmentation of the lemma was added as a morphological factor. This calculation was performed on the morphological analyses provided by CELEX (Baayen et al., 1993).

French Lexicon Project. The control variables provided in the French Lexicon Project (Ferrand et al., 2010) and included in the regression model were two measures of word frequency (cfreqmovies and cfreqbooks). The logarithmically transformed sum of both frequencies and their quadratic term were added as suggested by Ferrand et al. (2010) as a frequency measure accounting for largest amount of variance. Other important controls provided by the lexicon and included in the model were number of letters and number of syllables.[3] In addition to these variables, several other important lexical characteristics were obtained from the *Lexique 3* database (www.lexique.org; New, Brysbaert, Veronis, & Pallier, 2007; New, Pallier, Brysbaert, & Ferrand, 2004). The phonological transcription of the words was

[2] The Dutch Lexicon 2 project provided phonological neighborhood distance measure that was equal to the average phonological Levenshtein distance of the 30 nearest neighbors, unlike the measure in the English Lexicon project, in which the number of neighbors was 20.

[3] The are methodological differences between the lexicon projects. In the English Lexicon Project, unlike the other projects, the stimuli were presented in uppercase and the nonword material in most of the cases was generated by changing one letter from a baseword. The other lexicons used a lowercase presentation and nonword generation algorithms that were design to minimize serial scanning mechanisms for example by changing the proportions of different letters between the nonwords and the basewords as a function of the stimulus length. Methodological differences such as these might be the source of some of the observed effects, such as a stronger word length effect in the English Lexicon Project.

used to extract the identity of the first phoneme, which was entered as a categorical variable. The orthographical Levenshtein distance (OLD20) and the phonological Levenshtein distance (PLD20) were entered as measures of neighborhood densities. The number of morphemes of the word item was also added as a predictor. In the cases in which several possible values were matched to the same orthographic form, a preference was given to the biggest value (larger number of morphemes). In addition, nonpositional and positional mean type bigram and 1-gram frequencies were calculated for each of the items. The calculations were based on the *Lexique 3* word list with items without spaces and dashes. The list was generated from the *vwr* package in R (Keuleers, 2015).

3. Results

3.1. Lexical Decision Task

A hierarchical regression analyses was performed on the lexical decision latencies of the selected items from each of the lexicons.[4] The first step included all described control variables. The total proportion of variance explained in each of the regression models was $R^2 = 54.37\%$ for the English Lexicon Project (ELP), $R^2 = 44.39\%$ for the British Lexicon Project (BLP), $R^2 = 43.48\%$ for the Dutch Lexicon Project 2 (DLP2), and $R^2 = 42.94\%$ for the French Lexicon Project (FLP). In the next step, the repeated letter variables were added to the models. Their inclusion significantly improved all models: English, $F(26, 28202) = 10.304$, $p < .001$; British, $F(20, 24872) = 7.949$, $p < .001$; Dutch, $F(28, 18021) = 11.44$, $p < .001$, and French, $F(28, 31532) = 8.144$, $p < .001$. These variables accounted respectively for an additional $\Delta R^2 = 0.43\%$, $\Delta R^2 = 0.35\%$, $\Delta R^2 = 0.99\%$, and $\Delta R^2 = 0.41\%$ unique variance.

3.1.1. Adjacent repetitions

The coefficients of the repeated letter variables can be seen in *Figure 1*. When the repeated letters were adjacent, there was a 5 ms significant facilitation effect for vowel repetitions and a small (3 ms) nonsignificant facilitation effect for consonant repetitions in the ELP regression. In the BLP regressions, the 4 ms inhibitory effect was significant for consonants, while there was no effect (0 ms) for adjacent vowel repetitions. In the DLP 2 regression model, the results for the adjacent repetitions showed dissociation between consonants and vowels, with consonants having a significant 5 ms inhibitory effect, while vowels produced significant 7 ms facilitation effect. In the FLP regressions, the 9 ms effect of

[4] Full regression tables are available as a supplementary material.

adjacent consonant repetitions was inhibitory and significant. The effect of adjacent vowels was different, depending on whether diacritic marks were disregarded or not, with the variables, constructed with the diacritics sensitive algorithm having an inhibitory nonsignificant 6 ms effect, while the same variable produced significant 29 ms facilitation when diacritics were not taken into account. Overall, the results suggested a small facilitation effect of vowel repetitions, except for the diacritics-sensitive vowel variable in the French Lexicon project, while the effect of the adjacent consonants trended towards small inhibition, except for in the ELP model, in which it was not significant[5].

3.1.2. Nonadjacent repetitions (1-3 intervening letters)

When the repetitions were within a 1-to-3-intervening-letters distance, all consonants and vowels variables in all lexicons indicated a significant inhibitory effect. In ELP this effect was in the range 11 ms to 13 ms when the distance was within 2 intervening letters and dropped to 4 ms and 8 ms for the 3 intervening letters variables, vowel and consonant repetitions, respectively. In BLP, the effect size was in the range 7 ms to 11 ms for the vowel repetitions and 4 to 12 ms for consonant repetitions. The effect was slightly smaller in size, but quite consistent in DLP2. It ranged between 3 ms and 9 ms with a similar pattern for vowels and consonants. In FLP, the effect of repeated letters within the 1 to 3 letters distance was in the range from 7 ms to 12 ms for vowels and 7 ms to 9 ms for consonants.

3.1.3. Nonadjacent repetitions (more than 3 intervening letters)

When the repetitions were within a longer than 3 intervening letters distance, the pattern of results was not as consistent and clear as the one in the previous distance interval. However, as in the previous distance interval, the pattern was generally inhibitory. There were some differences within the lexicons as well as between the consonants and vowel repetitions. In ELP, 7 of the repeated letter variables within that distance indicated significant inhibitory effect. In this lexicon the vowel repetitions in the longer distances peaked and indicated large effects. A similar peak was also observed in the BLP results for the vowel repetitions, while none of the consonant variables within that distance interval showed a significant effect. In DLP2, there were significant inhibitory effects of the 4 letters distance for both letter types, as well as for the 5 and 9 letters distance vowels repetition. In FLP, the only significant variables that

[5] To investigate the possible effect of the frequency of the geminates (adjacent repeated letters), a separate analysis was performed on subsets of words containing one geminate. This frequency measure replaced the variables encoding adjacent repetition. It represented the number of occurrences of the geminates in the wordlists of the main regression analyses. All other variables were identical to the one in the main regression analyses. The results indicated significant facilitation effects of geminate frequency in ELP ($\beta = -0.018$, $t = -4.759$, $p < .001$); BLP ($\beta = -0.011$, $t = -2.923$, $p = .003$), DLP2 ($\beta = -0.004$, $t = -4.271$, $p < .001$), and a significant inhibitory effect in FLP ($\beta = 0.009$, $t = 4.110$, $p < .001$).

encoded repetition within that distance interval were the consonant repetition variables with 5 and 6 intervening letters between the repetitions.

3.1.4. Lexical Decision (up to ten-letters-long words)

Regression models were also fitted on subsets of the words from all lexicons, that were up to ten letters long. This was done with the purpose of providing a common basis of comparison between the behavioral data on these subsets and simulation results of computational models, where not all models accommodate longer words in their simulators. The diacritics-insensitive variables were used for the French Lexicon Project analyses, as diacritic marks were not implemented in two of the models under examination, and coefficients from the Bayesian Reader did not substantially differ when diacritics were used. The coefficients of the repeated letter variables on these subsets can be seen in *Figure 2*. Overall, the patterns did not deviate from those observed with the slightly bigger datasets. The adjacent vowel repetitions were either nonsignificant (ELP, BLP) or trended towards facilitation (DLP2, FLP with no diacritics), while the adjacent consonant repetitions had either null or small inhibitory effect. The results at the 1-3 intervening letters distance were broadly significant and inhibitory. The pattern at longer distances was mainly inhibitory, but not all variables were significant.

3.2. Lexical Decision Task Simulations

3.2.1. Spatial Coding Model

Lexical decision task simulations were run with the Spatial Coding Model (SCM; Davis, 2010) on the data sets of words up to ten letters long from ELP, BLP, DLP2 and FLP using the SCM simulator[6]. Prior to each simulation, the model's vocabulary was set to the word list of the corresponding lexicon project. The vocabularies contained all words from the lexicon projects, except for those longer than 10 letters. The resultant lexical decision latencies of the correctly recognized words were then entered in regression models as the dependent variable. The control predictor variables were identical to the ones in the corresponding regression models on behavioral data and were held constant across models' simulations. These variables explained R^2 = 36.11%, R^2 = 75.81%, R^2 = 67%, R^2 = 57.73% of the variance in the model's reaction times for ELP, BLP, DLP2 and FLP, respectively. In the next step, the repeated letter variables were added to the models. All models were significantly improved, except for the English Lexicon Project model: English, $F(18, 25314) = 1.452$, $p = .097$; British, $F(18, 24837) = 8.887$, $p < .001$;

[6] Downloaded from http://www.pc.rhul.ac.uk/staff/c.davis/SpatialCodingModel/

Dutch 2, $F(18, 18388) = 8.322$, $p < .001$, and French, $F(18, 26779) = 62.34$, $p < .001$. These variables accounted respectively for an additional $\Delta R^2 = 0.07\%$, $\Delta R^2 = 0.15\%$, $\Delta R^2 = 0.27\%$, and $\Delta R^2 = 1.7\%$ unique variance.

The effects of the repeated letter variables predicted by SCM for the four datasets can be seen in *Figure 3*. The model's predictions did not agree with the patterns observed in the empirical data. The consistent inhibitory pattern in the distance of 1-to-3 intervening letters was not present in the SCM simulation results. The model predicted several very small significant effects in the different lexicons, many of which were in the opposite direction. In FLP, SCM predicted a relatively strong inhibitory effect for adjacent vowel repetitions (4 cycles). This effect was, however, facilitatory in the corresponding empirical diacritics-insensitive regression model.

3.2.2. Relative Position Open Bigram Model

Lexical decision task simulations were also run with the Relative Position open bigram Model (RPM; Grainger, & van Heuven, 2003) on the same reference data sets as the ones used for the SCM simulations. The RPM simulations were run with the *easyNet* software (http://adelmanlab.org/easyNet/). The model's vocabulary was set to the corresponding lexicon word list prior to each simulation. Only the correctly recognized words were included in the analyses (see *Figure 4* for number of observations in each dataset and patterns of repeated letter effects). The control predictors explained $R^2 = 23.23\%$, $R^2 = 23.74\%$, $R^2 = 22.82\%$, $R^2 = 19.41\%$ of the variance of the model's reaction times for ELP, BLP, DLP2 and FLP, respectively. Adding the repeated letter predictors improved all four models: English, $F(18, 22938) = 6.853$, $p < .001$; British, $F(18, 21869) = 11.263$, $p < .001$; Dutch 2, $F(18, 17455) = 4.771$, $p < .001$, and French, $F(18, 21054) = 9.086$, $p < .001$. These variables accounted respectively for an additional $\Delta R^2 = 0.41\%$, $\Delta R^2 = 0.7\%$, $\Delta R^2 = 0.38\%$, and $\Delta R^2 = 0.62\%$ unique variance.

As can be seen in *Figure 4*, RPM tended to predict inhibitory, rather than facilitatory, effects of repeated letters. However, in most of the datasets, the effects were larger for the adjacent repetitions and were not consistent in the 1-3-letters distance interval, therefore also failing to capture the empirical data pattern.

3.2.3. Bayesian Reader Model

Lexical decision task simulations were run with the Bayesian Reader Model (BRM; Norris & Kinoshita, 2012) on the data sets of words up to ten letters long from ELP, BLP, DLP2 and FLP. The simulation

software was that provided by Norris and Kinoshita as a supplementary material to their paper. Prior to each simulation, the model's vocabulary was set to the word list of the corresponding lexicon. Words longer than 10 letters were excluded from the vocabularies. The mean reaction times of the correctly recognized words ("Yes" responses) were then entered in the regression models as the dependent variable. The "yes" threshold was set to .90. The control predictor variables were identical to the ones in the corresponding behavioral data regression models and the simulations with the previous two models. These variables explained R^2 = 62.29%, R^2 = 84.85%, R^2 = 86.76%, R^2 = 79.35% of the variance in the model's reaction times for ELP, BLP, DLP2 and FLP, respectively. When the repeated letter variables were added, all but the DLP2 models were significantly improved: English, $F(18, 24637) = 2.419$, $p < .001$; British, $F(18, 24837) = 8.844$, $p < .001$; Dutch 2, $F(18, 18388) = 1.308$, $p = .171$, and French, $F(18, 26779) = 5.093$, $p < .001$. These variables accounted respectively for an additional ΔR^2 = 0.07%, ΔR^2 = 0.1%, ΔR^2 = 0.02%, and ΔR^2 = 0.07% unique variance. As the Bayesian Reader Model has an implemented version that is diacritics-sensitive, separate French Lexicon Project simulation was also run with items and vocabulary with preserved diacritics. The control variables explained R^2 = 81.07 % of the variance in the model's reaction times. The repeated letter variables improved the model significantly, $F(18, 28120) = 2.157$, $p = .003$. The results of the Bayesian Reader Model simulations are displayed in *Figure 5.* The model did not predict a consistent pattern of the repeated letter effects and did not capture the inhibitory repeated letter effects in the shorter repetition distances.

3.3. Word Naming Task

A linear regression model was also fitted with the latencies from the word naming task, obtained from the English Lexicon Project (Balota et al., 2007). The same control variables were entered in the model as the ones in the lexical decision model for this lexicon project. These variables explained R^2 = 54.52% of the variance. The model was significantly improved after adding the repeated letters predictors, $F(26, 28206) = 11.822$, $p < .001$. They explained additional ΔR^2 = 0.49% unique variance. The effects of repeated letters can be seen in *Figure 6.* Overall, the pattern was the same as the one in the lexical decision data. In the word naming results, however, both adjacent repetitions had a significant facilitation effect and were slightly larger in size than those in the lexical decision results (10 ms for vowel repetitions and 7 ms for consonant repetitions). This facilitation effect was followed by consistent inhibitory effects in the interval of 1-to-3 intervening letters for both letter types. In the more than three letters distance interval, most of the variables, especially the vowel repetitions were significant, and, as in the lexical decision data, there was a peak of the inhibitory effect in the long-distance vowel repetitions.

4. Discussion

The presence of letter repetitions in words significantly predicted reaction times of lexical decision and word naming tasks, after controlling for important lexical and sublexical variables. Regressions showed that the presence of nonadjacent letter repetition delays visual word recognition. The inhibitory repeated letter effect was robust in the cases in which the repeated letters were intervened by one, two or three letters. The effect was less consistent but still generally present with increasing distance between the repetitions. The effect was observed in all three languages under consideration: English, Dutch, and French. This replication implies that the inhibition of letter repetition might be linked to a general mechanism in visual word recognition rather than to idiosyncrasies of a single language. The obtained results are in accordance with the reported target type effect in the methodologically similar primed lexical decision task in the Schoonbaert and Grainger's (2004) study, in which targets containing nonadjacent repetitions took longer to process. However, they disagree with the methodologically less similar masked priming data, reported by these authors and by Van Assche and Grainger (2006), which provided no evidence for differential processing between repeated and unique letter identities. The results are also broadly consistent with Gomez et al.'s (2008) perceptual identification data that showed lower accuracies for stimuli containing repetition in comparison to stimuli with no repetition. A difference between their results and the results presented here is that Gomez et al. reported inhibitory repeated letter effects in both adjacent and nonadjacent repetitions, while the inhibitory pattern in the present study is more consistent in the cases of nonadjacent repetitions.

In the cases of adjacent repetitions, the effect had alternating patterns, suggesting that adjacent repetitions might be a special case of letter repetitions. The results also implied a possible dissociation between consonants and vowels and double letter processing in different languages. In English, the inhibitory effect was less consistent for consonants and trended towards facilitation in the cases of vowel repetitions. In Dutch, the vowel-consonant dissociation was clearly observed, with repeated consonants producing a small but significant inhibitory effect, while repeated vowels had a small but significant facilitatory effect. In French, the adjacent consonants preserved their inhibitory pattern. It should be noted that adjacent vowel repetitions are extremely rare in French, mostly occurring in loanwords, while in Dutch they are quite common. Given the constraints of French orthography, the sensitivity of the pattern for vowels to the processing of diacritics suggests that words with the bigram *ée* are read quicker than those containing *ee* or *oo*, which are very often loanwords. The differential effects might therefore result from the adjacent double letters idiosyncrasies of the language.

The different patterns of adjacent repetition might be due to counteractive processes that are not associated with cases of nonadjacent repetition and possible benefits of the letters being in contiguous positions. Such an interpretation is consistent with the idea that adjacent repetition (letter doubling) should be coded as an additional dimension, that is separate from letter identity and letter position. This idea arises in studies and models of spelling (e.g., Fischer-Baum & Rapp, 2014; Glasspool & Houghton, 2005; Tainturier & Caramazza, 1996). Tainturier and Caramazza argue that graphemic representations are multidimensional structures and grapheme doubling is encoded separately from grapheme identity. In their study, they reported a case of dysgraphic patient in which the presence of adjacent letter repetition was preserved even when the letter identity and order information was severely distorted (e.g., giraffe - GAFFICATE). The special status of repeated adjacent letters has also been supported by evidence from visual word recognition research. It has been demonstrated that participants are more likely to misperceive the number of letters in a word and report a word with repetition (WEED instead of WED) if it was presented with a distractor word with another double letter (WOOD) than with a distractor word with no repetition (WORD; Fischer-Baum, 2017).

Another possible cause of the diminished inhibitory effect in the cases of adjacent repetitions might be linked to the proposed letter identity "leakage" to nearby positions (Gomez et al., 2008; Norris et al., 2010) that could possibly be less detrimental in the cases of same adjacent identities than in the cases of nonadjacent ones. It should also be noted that two adjacent repeated letters could differ from two unique ones, especially in the cases of consonants, in that in most cases, in the languages considered in this study, these would map into a single phoneme, rather than into two phonemes. This would lead to shortening of the phonological length (number of phonemes) relative to a case with two different adjacent letters. However, any phonological contribution to the effect due to this reason was dealt with by including the number of phonemes as a covariate.

The finding of inhibitory effects of letter repetitions suggests that there is a mechanism related to the rapid processing of identical sublexical elements (such as letters) that has not been previously described in the visual word recognition literature. Such a mechanism might be linked to a smaller number of activated abstract representations in the cases of repetitions compared to cases with no repetitions. This explanation is conceptually consistent with the architecture of the open bigram model (Grainger & van Heuven, 2003) as well as with the argument of Schoonbaert and Grainger (2004) that the open bigram model would predict an inhibitory effect of repeated letters due to duplications of open bigrams and therefore a smaller number of active representations in cases of repeated letters. In accordance with their proposal, the open bigram model indeed predicted consistent inhibitory effects of repeated letters and in this sense was conceptually most accurate. However, the specific prediction of the model on word lists of

four lexicon projects did not match the observed pattern as it overestimated the inhibitory adjacent-repetition effect and underestimated the nonadjacent one, therefore depicting the wrong inhibitory pattern than the one in the lexicon projects.

A possible explanation for the wrong predictions of the open bigram model (Grainger & van Heuven, 2003) could be the encoding scheme of the model and more specifically, the manner of which open bigrams are constructed. One factor affecting the predictions is the three-letter distance constraint on how open bigrams are generated (*silence* consists of SI SL and SE but not of SN, SC and another SE). Another factor affecting the model's predictions is the left-to-right manner of open bigram construction that privileges internal letters as they participate in more bigrams. This tends to disadvantage of the last three letters of the word, for example, as they appear in the first position of only two, one and zero open bigrams, respectively (*nce* from *silence* will generate respectively NC, NE, and CE, and no bigram). In this way, in words in which one letter does not appear more than twice, as in the present lexicon subsets, a repeated bigram is only possible to occur either when the repeated letters are adjacent or when they are intervened by only one other letter. The latter case is further constrained on the repeated letters not appearing at external positions within the word. For these reasons, many words with nonadjacent letter repetitions are in fact not encoded with fewer units, a possible prerequisite for a prediction of an inhibitory effect. As the pattern of predictions of the open bigram model by Grainger and van Heuven (2003) does not agree with the nonadjacent inhibitory effects of repeated letters, a revision of the encoding scheme might be appropriate that could include some relaxation of the current constraints.

It was, perhaps, less surprising that the spatial coding scheme (Davis, 2010) also did not contribute to a successful simulation of the repeated letter effects, as the scheme contains explicit mechanism that prevents repeated letter identities from playing a different role than nonrepeated ones. This mechanism comprises cooperative competitive interactions between banks of receivers (clones of receptors) for each letter in a word representation (word template). When there are repeated letters in both the input and the word template, multiple receivers in a single bank would become active. This would force a competition between the activated receivers for the selection of a winner. This selection is based on the signal-weight differences between the input and the word template, which equal 0 in the cases in which the letters appear in their expected positions. In the current implementation of the model, the resolution in cases of multiple activated receivers occurs immediately and without any cost. This could explain why the Spatial Coding Model did not predict the observed inhibitory pattern of repeated letters and in cases even predicted effects in the opposite direction. The results obtained in this study therefore present a challenge for the Spatial Coding Model and possibly signal the necessity of implementing a time penalization for resolving cases with repeated letters. This solution was, in fact, already suggested by Davis (2010) for the

discrepancy between the model's mechanisms and the reported inhibitory effect of repeated letters in targets of the study of Schoonbaert and Grainger (2004). The disambiguation and allocation to the corresponding positions of two identical units (letters) already recognized by the system as separate objects could therefore be another possible explanation of the observed inhibitory repeated letter effect.

Another way to approach the repeated letter problem could be in terms of the Bayesian Reader framework described by Norris and Kinoshita (2012), in which there is graded positional uncertainty: Evidence for a letter in a given position is also (weaker) evidence for nearby positions. A Bayesian decision maker receiving noisy data from two identical letters in different positions should therefore place some likelihood on the source of these data being a single instance of that letter somewhere in the middle, which will be evidence for a different string, typically a nonword, slowing responses to words in lexical decision. By contrast, this possibility does not need to be considered when there are no repetitions of letters. Although the Bayesian Reader model (Norris & Kinoshita, 2012) did not predict a consistent inhibitory pattern from repeated letters, such an explanation is conceptually consistent with the broader theoretical framework with a specific model enhanced with a more perceptually veridical likelihood calculation.

Another possible explanation to an extent follows up from the problem of a smaller number of abstract representations and more specifically, the shared identity of letter objects in cases of repetition. It might be the case that repeated identities force serial processing due to functional specialization and modularity of the processing components (letter receptors) associated with each letter identity. In such a scenario, processing of two identical elements might not be achieved in the rapid parallel manner, in which it has been previously demonstrated that letters are perceived (Adelman, Marquis, & Sabatos-DeVito, 2010), as the same component might be involved in encoding of two or more letters at once, therefore resulting in some processing delay, albeit not in predictable sequence. This delay differs from that considered for the Spatial Coding Model (Davis, 2010) both in that it affects one or other of the letters, but not both, and in that this delay is tied to an early stage of perception rather than a later stage of resolution of letter repetition in the lexical unit.

The problem of repeated letter identities might also be approached in terms of exploring how the visual system establishes the numerosity of objects with similar identities. It might be the case that the inhibitory repetition effect might result from limitations of the perceptual system in either very early low-level stages or later more abstract levels in the form of incapability of dissociating between two identical (letter) units. The perceptual system could perceive the two identical letters but cannot perceive that they are two, that the same letter representation is present at two different locations. In this case, the identity of both letters is recognized but it is the numerosity that is yet to be established. Such an assumption is

consistent with the repetition blindness phenomenon and the problem of registering two different tokens of the same type. Kanwisher (1987) found that participants were unable to report the second occurrence of a word from a word list in a rapid serial visual presentation (RSVP), when words were presented for a 117 ms followed by a mask and when the two occurrences were intervened by a small number of items. Since the effect was larger with a smaller number of intervenors, Kanwisher argued that the effect was not due to people forgetting the items. She also discarded the possibility that the tokens were not perceived. In Experiment 3 of her study, she demonstrated that when the word lists were truncated after the second occurrence of the repeated word, or its corresponding nonrepeated control word, and participants had to report the final item in lists of various length, their accuracy was higher when the final word was repeated than when it was not. Kanwisher posited that the repetition blindness phenomenon is not related to inability of recognizing the type of the occurrences, but rather to "individuating" the items as two tokens of the same type.

Two apparent differences between Kanwisher's study (1987) and the present are: the serial versus the parallel presentation of the repeated items; their scale (words vs. letters). However, the problem of perceiving several occurrences (tokens) of the same letter (type) has also been explored with a parallel presentation of letter strings containing repetitions. Kanwisher (1991) used word and nonword stimuli which were displayed in the original RSVP presentation, a "moving" RSVP presentation with each incremental letter shifted one character to the right, and a simultaneous presentation in which the whole string was briefly flashed. The word stimuli containing repetition were chosen so that the deletion of a second occurrence generated another genuine word (manager – manger). The repetition blindness effect was observed in all three presentation conditions, with the simultaneous repeated condition being least detrimental.

Another account of the repetition blindness effect was proposed by Luo and Caramazza (1996). They argued that the recognition system might be less sensitive to a type node shortly after it has already been perceived (a refractory period). They suggested that the processing of successive or concurrent stimuli continues after their presentation and the recognition of two identical items might overlap in time in both RSVP and simultaneous visual presentations. In their study, they manipulated the onset of the encoding of two critical items in both type of presentations. The authors assumed that in the simultaneous presentation, the distance between the repeated elements affects the temporal overlap of their processing. In their experiment, five letters and three dollar signs were displayed in eight consecutive clocklike positions within a circle. In the RSVP presentation, the letters appeared sequentially in a clockwise direction and were either preceded or followed by the dollar signs. In the repeated condition, the repeated letters appeared within a distance of one intervening letter and were either the second and the fourth or the

third and the fifth within the letter sequence. The start location of the stimuli presentation was any of the eight possible clock locations. In the simultaneous presentation, all letters and dollars were briefly displayed at the same time. The results indicated lower accuracy in the repeated conditions in both presentation modes with a stronger repetition blindness effect in the RSVP condition. In a follow-up experiment, the authors further investigated the repetition blindness effect in simultaneous presentation by manipulating the distance between the repeated elements in terms of number of other intervening letters (0-3). The results showed that the effect was stronger within distances 1 and 2 than distances 0 and 3 and indicated an inverted U-shaped repetition blindness effect as a function of distance. The authors explained the results in terms of a type refractory period in which the sensitivity of the recognition system to a certain type is reduced after stimulation and is recovering to its resting levels. If two identical stimuli are encoded with a minimal delay and the second stimulation is received shortly after the first one, the excitation level of the type node is still peaking above its resting level, a priming facilitation effect might occur (repetition priming). As, however, after peaking, the activity of the node drops below its resting level, if the second stimulation happens while the node recovers to its resting state (the refractory period), this will lead to a delayed encoding of the second stimuli and an inhibitory effect proportional to the distance away from the resting level. Luo and Caramazza suggested that the effect of the repetition lag (distance) can be explained with the activity of the type node which after peaking first decreases below resting level and then recovers back to resting level resulting in the reversed U-shape of the repetition blindness effect. The reported repetition distance effects by these authors are consistent with the results of the present study, in which the adjacent repeated letter condition induces less inhibition than the immediate nonadjacent conditions, which in turn have more consistent effect than the longer repetition distances.

In another study related to the repetition blindness effect, Mozer (1989) showed that participants were less accurate when they had to report the number of letters in a briefly presented display when a string of letters was formed by repeating the same letter (DDDD) than when it was formed by different letters (NRVT). He referred to this effect as “homogeneity effect” and argued that it is dependent on the common form of the items. He extended his findings and demonstrated that a repeated letter effect can also operate on the level of abstract letter identities. When the task required recognition of the letter identities, rather than just counting the visual objects, the number of two nonadjacent repeated letter targets presented in different case was still harder to perceive than the number of two distinct letters. Participants were less accurate when they reported the total number of As and Es in a display of two CVC strings when the display contained repeated letters (BEC mes) than when the target letters were not repeated (ner TAL). Mozer proposed a model of parallel processing of information from different visual fields. The repeated letter (homogeneity) effect is caused by spatial uncertainty and imprecision in

retrieving the exact location information of the two repeated objects. As the system differentiates between the two instances of a single object by the difference in their location, in stages that have insufficient spatial information it is unable to process their number. Mozer argued that the repetition blindness phenomenon described by Kanwisher resembles the homogeneity effect in his study. The difference in the mechanisms of the two effects, according to him, were that the insufficient processing time in the rapid serial visual presentation led to inability for the events to be tagged to a temporal serial position in a sequence, while the homogeneity effect in parallel processing were caused by inefficient spatial tagging. In both cases, the information of the number of identical objects therefore could not be retrieved. Mozer also suggested that serial attentional scanning could decrease spatial uncertainty and weaken the homogeneity effect.

Another possible cause of processing delay could be the increased positional ambiguity when letters are repeated. With identical signals coming from separate locations, making it more difficult for the visual system to allocate two identical letters to their corresponding positions than two different letters. In this case, the identity of both letters is established, and it is the positional encoding that is delayed, possibly leading to increased processing difficulty. Such an interpretation is, however, not consistent with the evidence provided by Adelman (2011) from a two-forced choice perceptual identification task showing a similar take-off of the accuracy function in conditions in which the foils were formed by identity substitution and a condition with transposition, suggesting that information about identity and position gets resolved in a similar timeframe.

This evidence was considered in the design of the Letters in Time and Retinotopic Space (LTRS; Adelman, 2011) model. Although LTRS describes perceptual timing mechanisms that could explain effects of different prime types, it does not describe downstream lexical processing mechanisms that would be affected by these timings when entirely consistent information (i.e., the target alone) is presented. The model perceives items of information at different times, so identifying one of a set of repeated letters is ambiguous between several possibilities unless there is other disambiguating information. So, although the model is highly sensitive to (excess) repeated letters in primes, it is sensitive to repeated letters in lexical decision targets only insofar as they make some non-identical primes more ambiguous. The model could be extended to include less efficient processing of repeated letters, by adjusting the attention equation so that repeated letters do not receive as high a processing rate as unique letters. This would delay responding to these items, because the total processing rate would be lower, increasing the average latency to perception of the first letter, and hence delaying the start of lexical processing. This would not provide evidence as to the nature of the delays in letter processing, although there is evidence from other paradigms that such delays exist.

In summary, the possible causes of the repeated letter effects include lower number of active representations, resolution of the position of the repeated letters at later processing stages, capacity limitations due to serial processing of repeated letters, inability of the processing system to tokenize the letter types at early perceptual stages, increased ambiguity in cases of repeated letters. It might also be the case that multiple of the aforementioned mechanisms underly the inhibitory repeated letter effect. For example, it might be the case that the perceptual system is indeed unable to establish the numerosity of the letters at early stages and requires additional time to establish that the positional uncertainty for a single identity is higher than that of other identities (even after a certain period of time) and therefore the source of that information should come from (at least) two separate locations. There is also a possibility that the repeated letters effects might be driven by other non-perceptual phenomenon not yet addressed in the current literature. The control variables that were included, however, consisted of the major lexical and sublexical factors, including but not limited to frequency and size measures in multiple levels. We, therefore, consider it unlikely for the effect to be driven by effects such as letter frequencies or any of the known possible factors. The clear presence of an inhibitory effect within megastudies, established here, will require complementary experimental efforts to determine its cause.

In conclusion, the results of the present study demonstrated a robust inhibitory effect which was replicated in all three languages under consideration. The effect was stable in the cases of one, two and three intervening letters, it was broadly consistent in larger distances and decreased in the cases of adjacent repetitions. The observed inhibitory pattern of repeated letter effects in the regression analyses was not predicted by three leading computational models and it is not yet reflected in any mechanism in the visual word recognition literature. These results have important implications for development of theories of visual word recognition and, more specifically, for understanding letter position and identity encoding and sublexical orthographic processes mediating lexical access. The obtained effects should motivate additional research in that direction, as well as additional theoretical and computational modelling effort with the purpose of the better understanding of the processes involved in the repeated letters effects.

Acknowledgments

This work was supported by the Leverhulme Trust (Project Grant RPG-2013-408) and an Early Career Fellowship awarded to the first author by the Institute of Advanced Study, University of Warwick, United Kingdom.

References

Acha, J., & Perea, M. (2010). On the role of consonants and vowels in visual-word processing: Evidence with a letter search paradigm. *Language and Cognitive Processes, 25*, 423–438. https://doi.org/10.1080/01690960903411666

Adelman, J. S. (2011). Letters in time and retinotopic space. *Psychological Review, 118*, 570.

Adelman, J. S., Marquis, S. J., & Sabatos-DeVito, M. G. (2010). Letters in words are read simultaneously, not in left-to-right sequence. *Psychological Science, 21*, 1799–1801. https://doi.org/10.1177/0956797610387442

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database (on CD-ROM)*. Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

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

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods* (pp. 90–115). Hove, East Sussex: Psychology Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*, 977-990.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 441–458. https://doi.org/10.1037/xhp0000159

Chetail, F., Balota, D., Treiman, R., & Content, A. (2015). What can megastudies tell us about the orthographic structure of English words*? The Quarterly Journal of Experimental Psychology, 68*, 1519–1540. https://doi.org/10.1080/17470218.2014.963628

Chetail, F., Drabs, V., & Content, A. (2014). The role of consonant/vowel organization in perceptual discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 938–961. https://doi.org/10.1037/a0036166

Chetail, F., Treiman, R., & Content, A. (2016). Effect of consonant/vowel letter organisation on the syllable counting task: evidence from English. *Journal of Cognitive Psychology, 28*, 32–43. https://doi.org/10.1080/20445911.2015.1074582

Coltheart, M., Davelaar, E., Jonasson, J., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological review, 117*, 713.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42*, 488–496. https://doi.org/10.3758/BRM.42.2.488

Fischer-Baum, S. (2017). The independence of letter identity and letter doubling in reading. *Psychonomic Bulletin & Review*, *24*, 873–878. https://doi.org/10.3758/s13423-016-1149-8

Fischer-Baum, S., & Rapp, B. (2014). The analysis of perseverations in acquired dysgraphia reveals the internal structure of orthographic representations. *Cognitive neuropsychology*, *31*, 237-265.

Forster, K. I., Davis, C., Schoknecht, C., & Carter, R. (1987). Masked priming with graphemically related forms: Repetition or partial activation? *The Quarterly Journal of Experimental Psychology Section A*, *39*, 211–251. https://doi.org/10.1080/14640748708401785

Glasspool, D. W., & Houghton, G. (2005). Serial order and consonant–vowel structure in a graphemic output buffer model. *Brain and language*, *94*, 304-330.

Gomez, P., Ratcliff, R., & Perea, M. (2008). The overlap model: A model of letter position coding. *Psychological Review, 115*, 577–600. https://doi.org/10.1037/a0012667

Grainger, J., Granier, J.-P., Farioli, F., Van Assche, E., & van Heuven, W. J. B. (2006). Letter position information and printed word perception: The relative-position priming constraint. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 865–884. https://doi.org/10.1037/0096-1523.32.4.865

Grainger, J., & van Heuven, W. J. B. (2003). Modeling letter position coding in printed word perception. In P. Bonin (Ed.), *The mental lexicon* (pp. 1–23). New York: Nova Science.

Kanwisher, N. G. (1987). Repetition blindness: Type recognition without token individuation. *Cognition, (27)*, 117-143.

Kanwisher, N. G. (1991). Repetition blindness and illusory conjunctions: Errors in binding visual types with visual tokens. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 404.

Keuleers, E. (2015) *vwr: Useful functions for visual word recognition research* (R package version 0.3.0). Retrieved from https://CRAN.R-project.org/package=vwr

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*, 287–304. https://doi.org/10.3758/s13428-011-0118-4

Kinoshita, S., & Norris, D. (2009). Transposed-letter priming of prelexical orthographic representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1–18. https://doi.org/10.1037/a0014277

Luo, C. R., & Caramazza, A. (1996). Temporal and spatial repetition blindness: Effects of presentation mode and repetition lag on the perception of repeated items. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 95.

Lupker, S. J., Perea, M., & Davis, C. J. (2008). Transposed-letter effects: Consonants, vowels and letter frequency. *Language and Cognitive Processes*, *23*, 93–116. https://doi.org/10.1080/01690960701579714

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: An account of basic findings. *Psychological Review*, *88*, 375–407. http://dx.doi.org/10.1037/0033-295X.88.5.375

Mozer, M. C. (1989). Types and tokens in visual letter perception. *Journal of Experimental Psychology: Human Perception and Performance*, *15*, 287–303. https://doi.org/10.1037//0096-1523.15.2.287

New, B., Ferrand, L., Pallier, C., Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, *13*, 45–52.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics; New York*, *28*, 661–677.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2 : A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*, 516–524. https://doi.org/10.3758/BF03195598

Norris, D. (2006). The Bayesian reader: explaining word recognition as an optimal Bayesian decision process. *Psychological review*, *113*, 327.

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: Why there's nothing special about the perception of orthography. *Psychological Review*, 119, 517–545.

Norris, D., Kinoshita, S., & van Casteren, M. (2010). A stimulus sampling theory of letter identity and order. *Journal of Memory and Language*, *62*, 254–271. https://doi.org/10.1016/j.jml.2009.11.002

Perea, M., & Lupker, S. J. (2003). Transposed-letter confusability effects in masked form priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked priming: The state of the art* (pp. 97–120). Hove, UK: Psychology Press.

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory and Language*, *51*, 231–246. http://dx.doi.org/10.1016/j.jml.2004.05.005

Peressotti, F., & Grainger, J. (1999). The role of letter identity and letter position in orthographic priming. *Perception & Psychophysics, 61*, 691–706.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved From https://www.R-project.org/

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes, 19*, 333–367. https://doi.org/10.1080/01690960344000198

Tainturier, M. J., & Caramazza, A. (1996). The status of double letters in graphemic representations. *Journal of Memory and Language*, *35*, 53-73.

[dataset] Trifonova, I. V., Adelman, J. S. (2019). *Data for: A delay in processing for repeated letters: Evidence from mega-studies.* Mendeley Data, v1.

Van Assche, E., & Grainger, J. (2006). A study of relative-position priming with superset primes. Journal of Experimental Psychology: *Learning, Memory, and Cognition, 32*, 399–415. https://doi.org/10.1037/0278-7393.32.2.399

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology, 67*, 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Welvaert, M., Farioli, F., & Grainger, J. (2008). Graded Effects of Number of Inserted Letters in Superset Priming. *Experimental Psychology, 55*, 54–63. https://doi.org/10.1027/1618-3169.55.1.54

Westbury, C., & Buchanan, L. (2002). The probability of the least likely non-length-controlled bigram affects lexical decision reaction times. *Brain and Language, 81*, 66-78.

Whitney, C. (2001). How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychonomic Bulletin & Review, 8*, 221-243.

Yap, M. J., & Balota, D. A. (2009). Visual word recognition of multisyllabic words. *Journal of Memory and Language, 60*, 502–529. https://doi.org/10.1016/j.jml.2009.02.001

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971–979. https://doi.org/10.3758/PBR.15.5.971

		Lexicon Project			
Distance	CV	ELP	BLP	DLP2	FLP
0	C	14.5	14.6	12.3	18.9
0	V	3.7	5.1	17.9	0.1
1	C	5.4	3.9	4.4	5.4
1	V	10.2	5.0	14.0	11.3
2	C	8.5	6.9	9.3	8.4
2	V	10.6	8.0	12.3	15.2
3	C	8.0	6.4	6.9	6.6
3	V	7.5	4.8	7.0	8.9
4	C	6.7	5.3	6.9	6.3
4	V	6.6	2.4	6.2	7.1
5	C	4.8	2.9	5.1	4.5
5	V	4.0	0.8	4.5	5.0
6	C	3.0	1.7	3.4	3.1
6	V	2.5	0.2	2.4	2.8
7	C	1.6	0.7	2.2	1.9
7	V	1.3	0.0	1.3	1.4
8	C	0.9	0.3	1.2	1.1
8	V	0.7	0.0	0.5	0.7
9	C	0.5	0.1	0.4	0.5
9	V	0.4	NA	0.1	0.3
10	C	0.2	0.0	0.2	0.2
10	V	0.1	NA	0.1	0.1
11	C	0.1	NA	0.1	0.1
11	V	0	NA	0	0
12	C	0	NA	0	0
12	V	0	NA	0	0
13	C	0	NA	0	0
13	V	0	NA	NA	0
Any	Any	65.7	54.2	74.1	68.4

Table 1.

Proportions of items containing any letter repetition (at least one repeated letter) and at least one of each of the repetition types in the English Lexicon Project (ELP), British Lexicon Project (BLP), Dutch Lexicon Project 2 (DLP2) and French Lexicon Project (FLP) datasets. The repetition types are presented in terms of the repetition distance (number of intervening letters between the repeated) and consonant-vowel (CV) class of the repeated letter. Words in which the same letter appears more than twice are excluded.

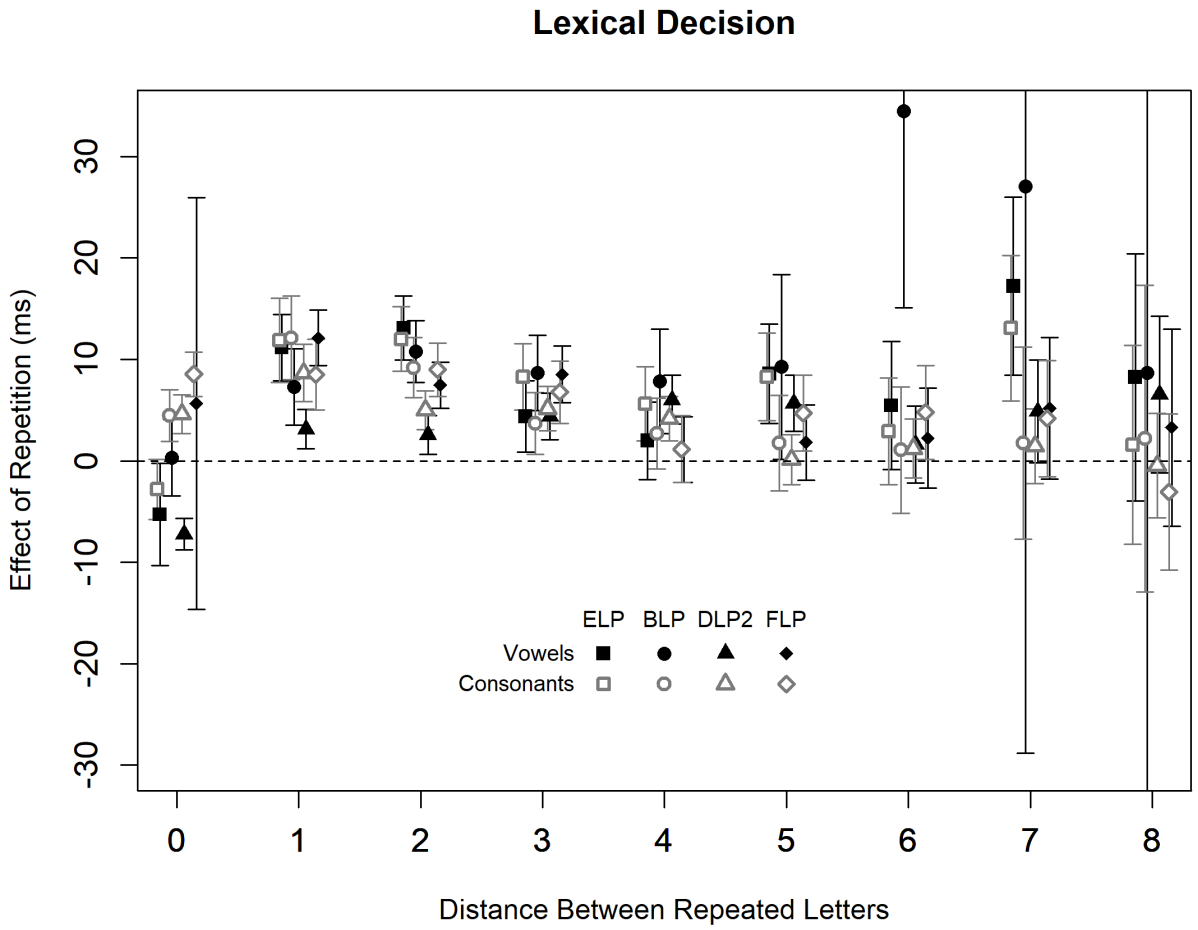

Figure 1. Repeated letter effects in English Lexicon Project (ELP), British Lexicon Project (BLP), Dutch Lexicon Project 2 (DLP2) and French Lexicon Project (FLP; with diacritics). Positive values indicate inhibition, negative values indicate facilitation. The distance is measured with the number of intervening letters between the repeated ones. Error bars represent 95% confidence intervals.

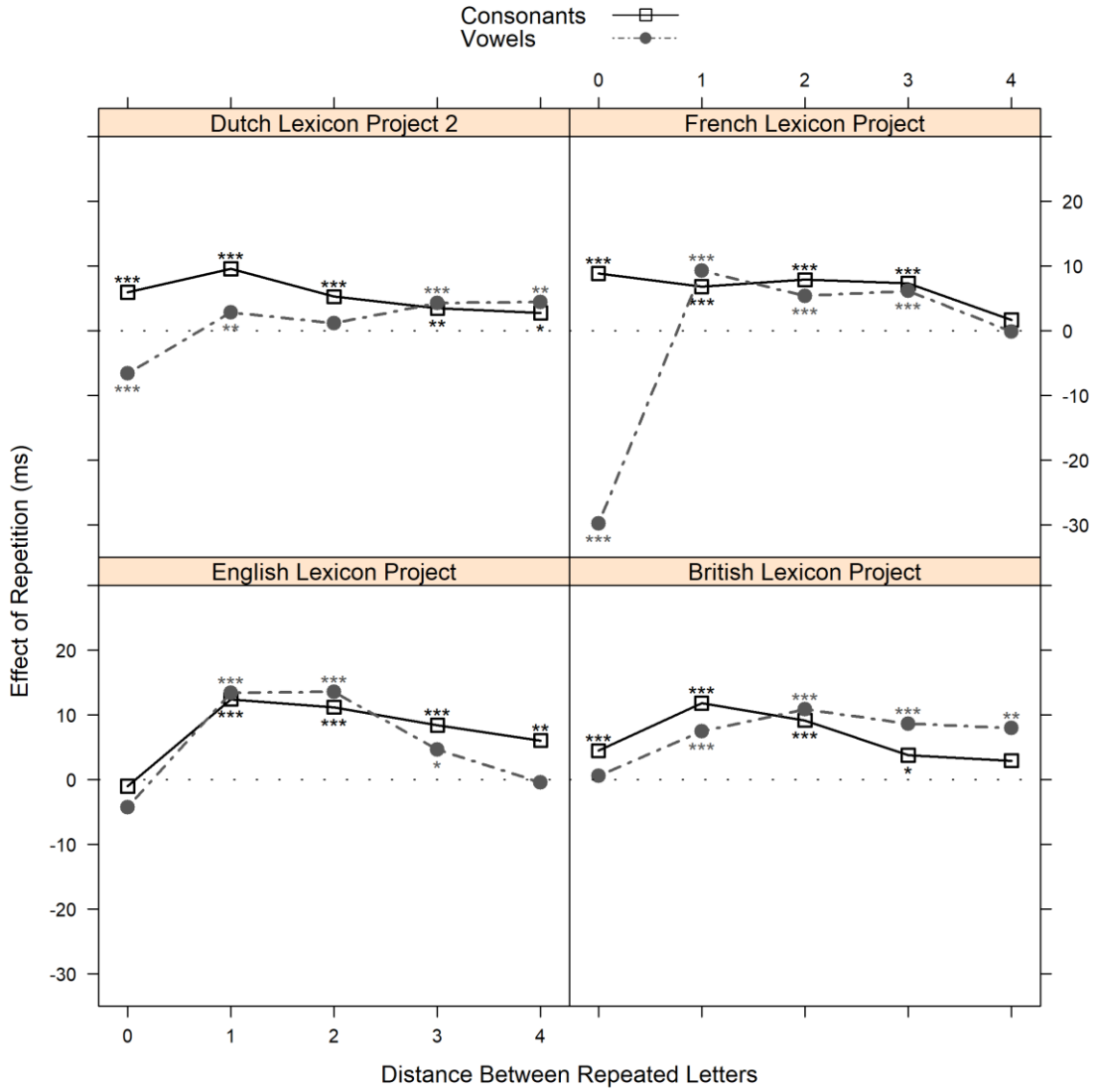

Figure 2. Repeated letter effects in subsets of the lexicon projects with words no longer than ten letters. Positive values indicate inhibition, negative values indicate facilitation. *** $p < .001$; ** $p < .01$; * $p < .05$

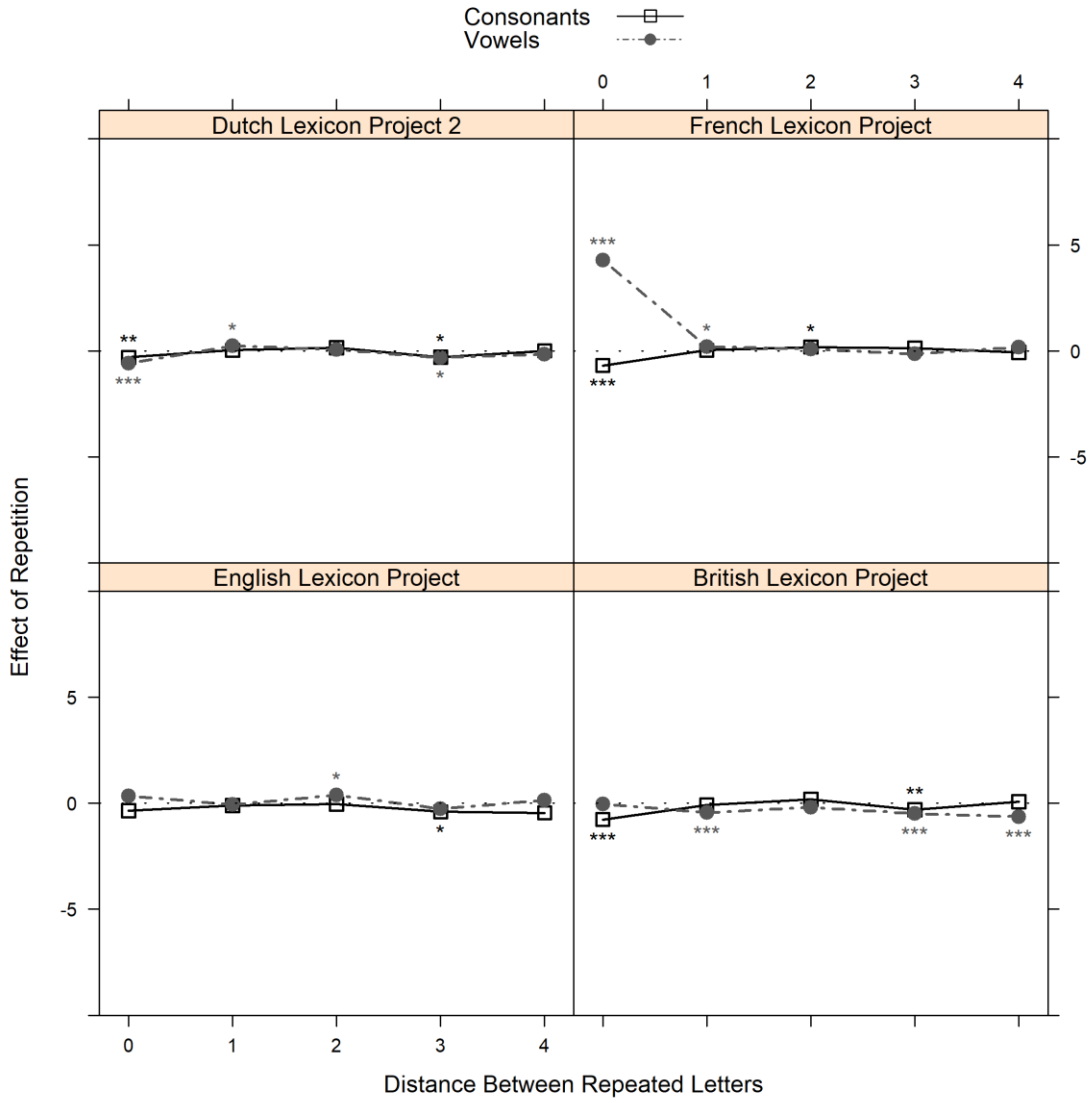

Figure 3. Spatial Coding Model's predictions of repeated letter effects. Positive values indicate inhibition, negative values indicate facilitation. *** $p < .001$; ** $p < .01$; * $p < .05$

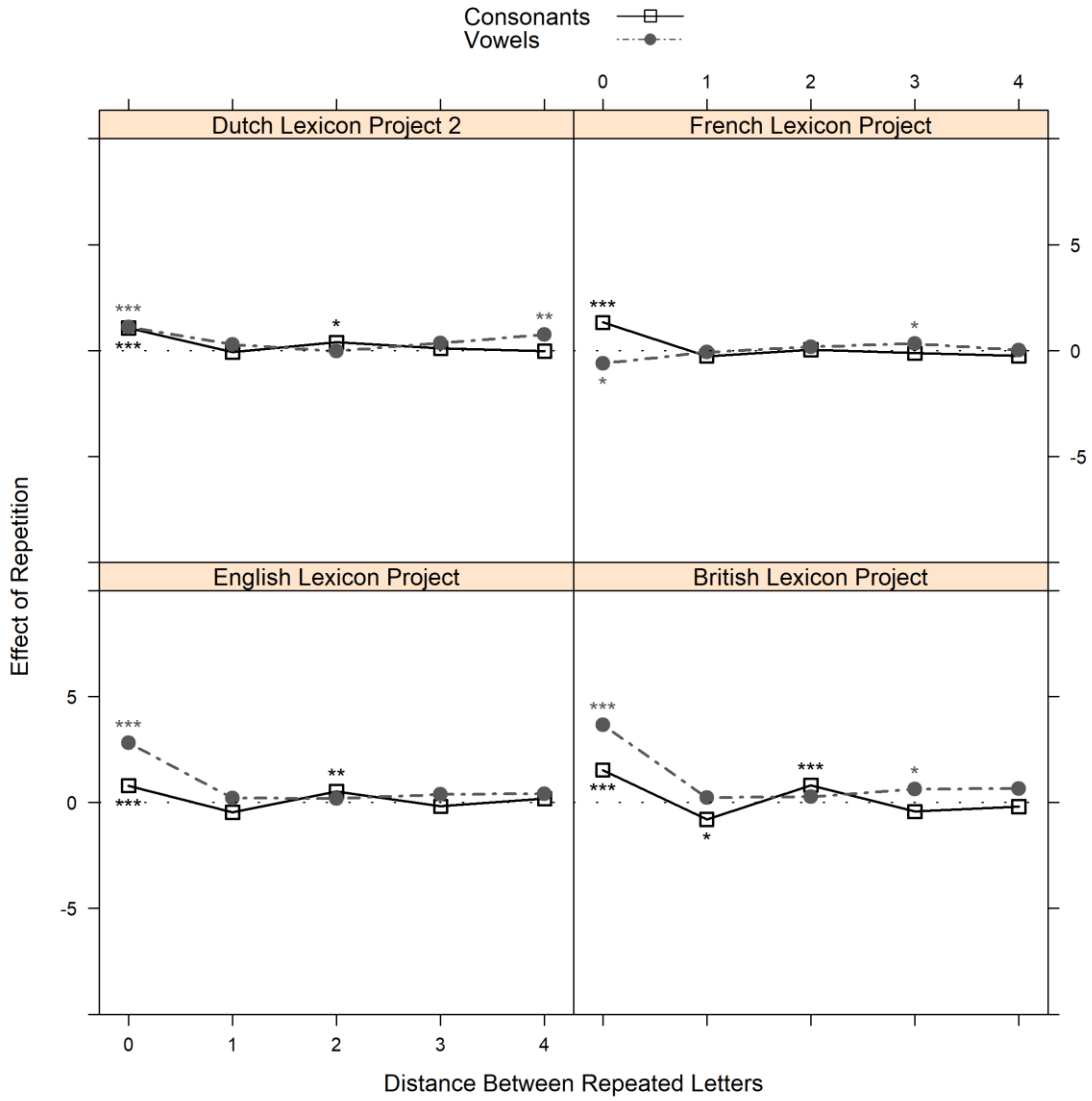

Figure 4. Relative Position Open Bigram Model's predictions of repeated letter effects. Positive values indicate inhibition, negative values indicate facilitation. *** $p < .001$; ** $p < .01$; * $p < .05$

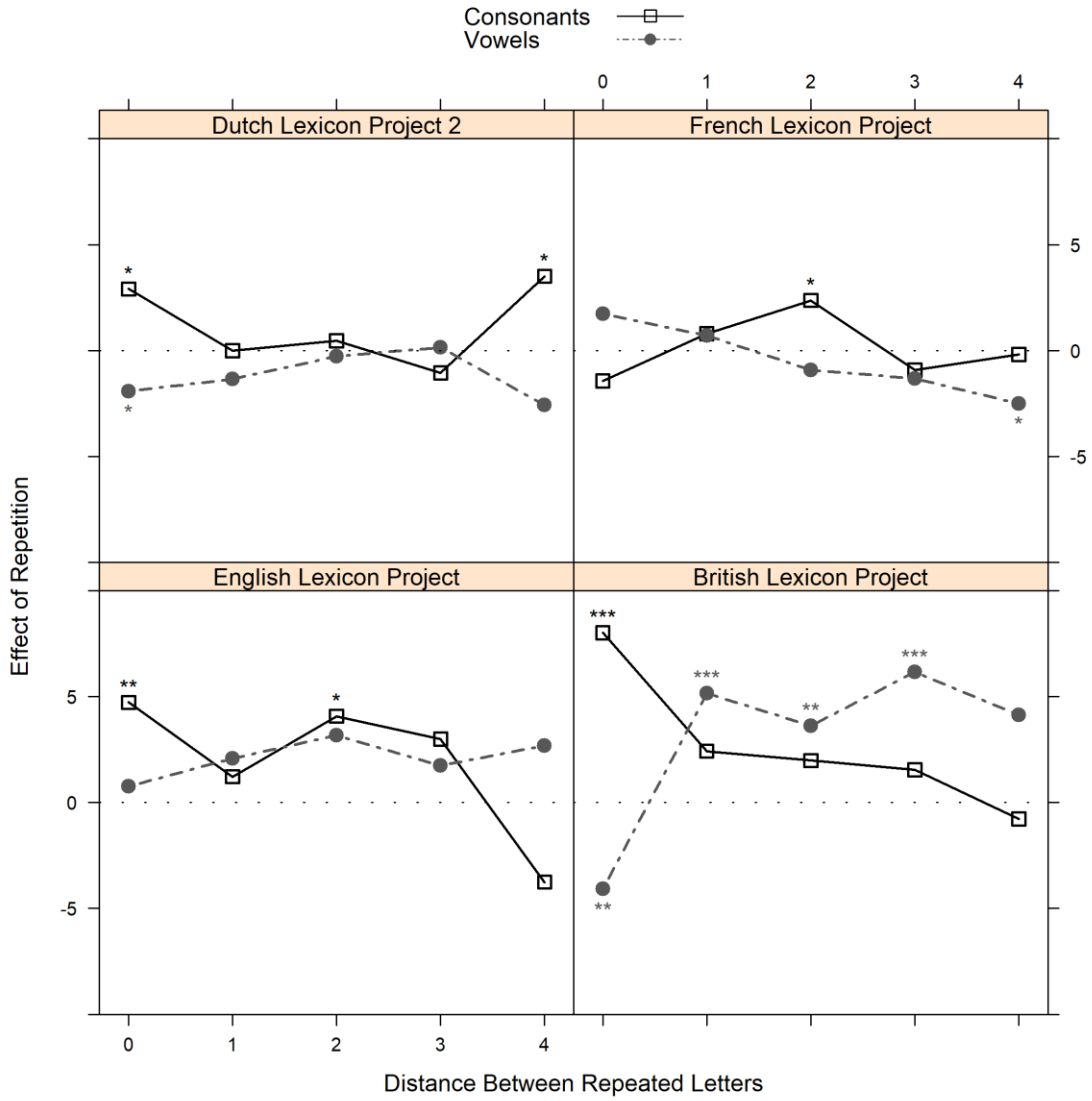

Figure 5. Bayesian Reader Model's predictions of repeated letter effects. Positive values indicate inhibition, negative values indicate facilitation. *** p <.001; ** p <.01; * p <.05

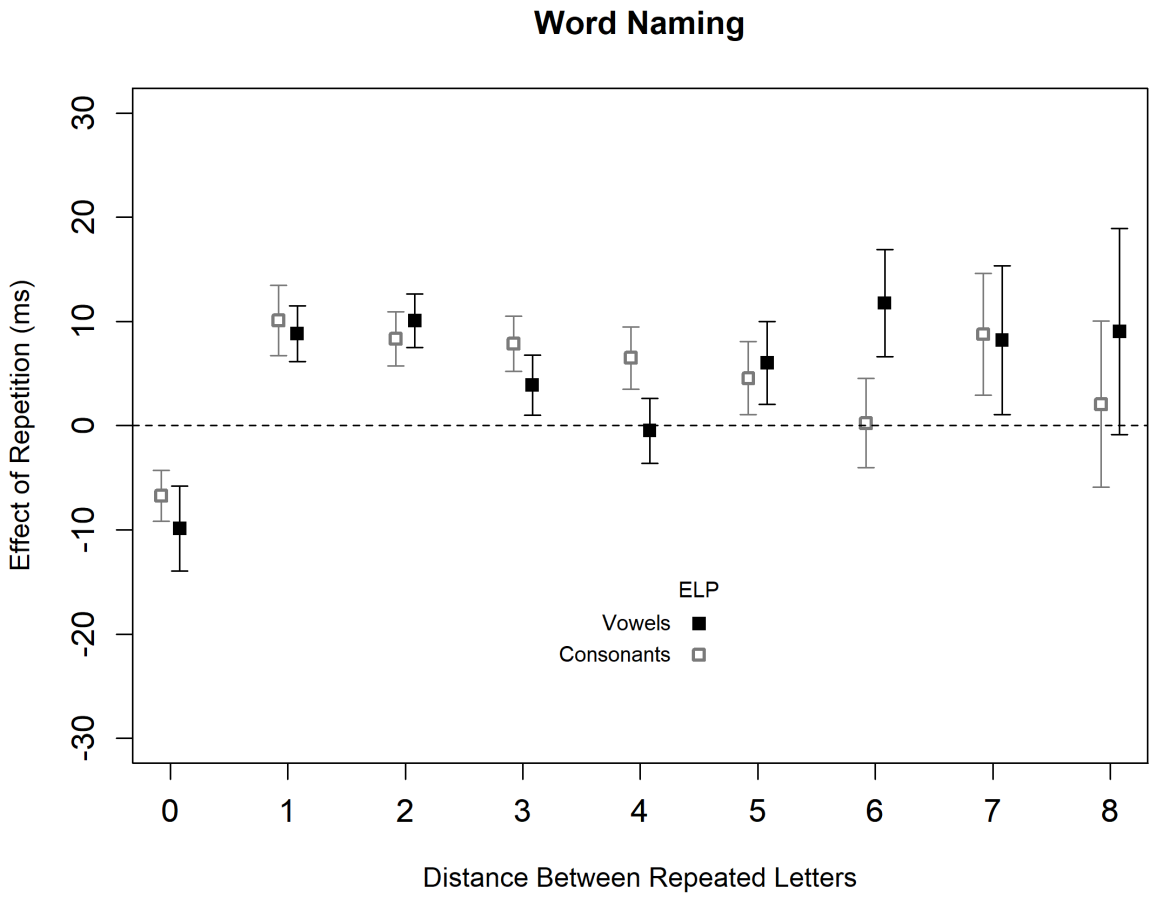

Figure 6. Repeated letter effects in word naming in English Lexicon Project. Positive values indicate inhibition, negative values indicate facilitation. The distance is measured with the number of intervening letters between the repeated ones. Error bars represent 95% confidence interval