Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes

Xenia Schmalz & Claudio Mulatti

Department of Developmental Psychology & Socialisation

University of Padova

PREPRINT VERSION 2

Please address all correspondence to X. Schmalz, xenia.schmalz@gmail.com.

Abstract

Psycholinguistic researchers identify linguistic variables and assess if they affect cognitive processes. One such variable is letter bigram frequency, or the frequency with which a given letter pair co-occurs in an orthography. While early studies reported that bigram frequency affects visual lexical decision, subsequent, well-controlled studies not shown this effect. Still, researchers continue to use it as a control variable in psycholinguistic experiments. We propose two reasons for the persistence of this variable: (1) Reporting no significant effect of bigram frequency cannot provide evidence for no effect. (2) Despite empirical work, theoretical implications of bigram frequency are largely neglected. We perform Bayes Factor analyses to address the first issue. In analyses of existing large-scale databases, we find no effect of bigram frequency in lexical decision in the British Lexicon Project, and some evidence for an inhibitory effect in the English Lexicon Project. We find strong evidence for an effect in reading aloud. This suggests that, for lexical decision, the effect is unstable, and may depend on item characteristics and task demands rather than reflecting cognitive processes underlying visual word recognition. We call for more consideration of theoretical implications of the presence or absence of a bigram frequency effect.

Key words:

Null hypothesis, reading, research methods.

Busting a myth with the Bayes Factor: Effects of letter bigram frequency in visual lexical decision do not reflect reading processes

In the 1980s, Cutler published a paper with the title "Making up materials is a confounded nuisance: or Will we be able to run any psycholinguistic experiments at all in 1990?" (Cutler, 1981). The paper laments the ever-increasing number of linguistic variables that have been shown to affect psychological processes: a psycholinguistic experiment must control for all these in order to ensure that an observed effect is a result of the manipulation, and not due to a confound. Thus, showing whether or not a linguistic variable has an effect on cognitive processing is necessary both in order to know whether it should be treated as a control variable, and – as is the general goal of psycholinguistics – to inform our understanding of how the cognitive system works.

In the current paper, we focus on letter bigram (hereafter: bigram) frequency, which is defined as the frequency with which each adjacent letter pair within a word co-occurs in the written language. Bigram frequency measures usually take into account specific position in which a bigram occurs. For example, "ll" has a low frequency in beginning positions (as in "llama"), but higher when it occurs in the middle ("yellow") or end of a word ("wall"). It is common for reading researchers to control for this variable, even though its theoretical value is rarely discussed, and reports of its effects on reading latencies are equivocal at best. We discuss the reasons that reading researchers might be reluctant to give up matching for bigram frequency, and we present an analysis of bigram frequency effects using Bayes Factor analyses on lexical decision and reading aloud data. Finally, we discuss the theoretical implications of finding either the presence or absence of a bigram frequency effect.

History of bigram frequency effects

Reports of bigram frequency effects on visual word recognition latencies in English readers started to emerge in the 1960s, from experiments using tachistoscopic word identification (Biedermann, 1966; Broadbent & Gregory, 1968; McClelland & Johnson, 1977; Rumelhart & Siple, 1975), and lexical decision tasks (Rice & Robinson, 1975). These studies reported significant bigram frequency effects, generally for low-frequency words, but not for high-frequency words or nonwords (reviewed in Chetail, 2015; Gernsbacher, 1984). Interestingly, across experiments, the effects for low-frequency words were sometimes facilitatory, and sometimes inhibitory. To get to this bottom of this inconsistency, Gernsbacher (1984) collected subjective word frequency ratings for the items used by previous studies. She argued that word frequency counts for low-frequency words are less reliable than word frequency counts for high-frequency words: unless a corpus contains a large number of words, the number of tokens of low-frequency words will not suffice to obtain reliable word frequency estimates. Indeed, the subjective ratings showed that the bigram frequency effects found by previous studies could be accounted for by a confound with word frequency, and when low-frequency words were matched on subjective frequency, the effect of bigram frequency disappeared.

Subsequent studies have not strengthened the case for a bigram frequency effect. Andrews (1992) manipulated orthographic neighbourhood (for any letter string, the number of real words that can be created by exchanging one letter; Coltheart, Davelaar, Jonasson, & Besner, 1977) and bigram frequency. As these two variables are highly correlated, this served to show that effects of orthographic neighbourhood could not be attributed to bigram frequency. Andrews (1992) found an effect of orthographic neighbourhood, but no effect of bigram frequency in lexical

decision. In reading aloud, a bigram frequency effect for low-frequency words was found in one experiment, both in a standard and delayed naming task. However, a follow-up experiment with more rigorously controlled items could not confirm that this effect was not due to a confound with initial phoneme characteristics.

More recently, mega-studies have provided comprehensive overviews of psycholinguistic variables that affect lexical decision (Balota et al., 2007; Ferrand et al., 2010; Keuleers, Diependaele, & Brysbaert, 2010; Keuleers, Lacey, Rastle, & Brysbaert, 2012). These databases contain data for thousands of words, and therefore have strong statistical power for testing the effects of various psycholinguistic variables on visual word recognition. The databases report no findings of any effects of bigram frequency. In an analysis of the English Lexicon Project (ELP; Balota et al., 2007), bigram frequency gets but a mention: "Other factors such as bigram frequency […] were not included in the analysis because they were not significant" (p. 48; New, Ferrand, Pallier, & Brysbaert, 2006). Similarly, the British Lexicon Project (BLP; Keuleers et al., 2012) replicates the findings of Andrews (1992), and finds no effects of bigram frequency in lexical decision. It is worth noting, however, that the presence or absence of an effect in a large database analysis depends to some extent on the exact constellation of variables that are used in the model (Rayner, Pollatsek, Drieghe, Slattery, & Reichle, 2007). In contrast to the footnotes by New et al. (2006) and Keuleers et al. (2012), an analysis of the ELP and BLP data on the processing of hiatus words found inhibitory effects, or shorter RTs for words with rare bigrams, in naming and lexical decision data (Chetail, Balota, Treiman & Content 2015).

In summary, the evidence for effects of bigram frequency in visual word recognition stems mainly from early studies, and may often be explained by the use of unreliable word frequency statistics, confounds, or depends on the exact covariates

included in a large-scale analysis. Yet, psycholinguists take bigram frequency seriously: most item-matching programs provide information about bigram frequency (e.g., N-Watch; Davis, 2005; and WordGen; Duyck, Desmet, Verbeke, & Brysbaert, 2004). Bigram frequency counts can also be easily retrieved from the large-scale databases along with the behavioural data (Balota et al., 2007; Ferrand et al., 2010; Keuleers et al., 2012), and a dedicated website provides numerous types of bigram frequency statistics (Medler & Binder, 2005). To make matters more complicated, the counts do not coincide across these platforms. Different corpora, trimming procedures, ways of treating position-specificity, summing or averaging the frequency across a word, or using type versus token counts, yield different values. For example, for the word *cat*, N-watch gives averaged type and token frequency counts of 10.5 and 1520, respectively; the BLP a summed bigram frequency of 28,844; the ELP a summed bigram frequency of 5,984; and the MCWord database gives position-specific type and token counts of 11 and 1462.13 and position-independent type and token counts of 30,418.17 and 5,332.5, respectively. The availability of linguistic data on bigram frequencies is, beyond doubt, valuable, but it raises the question of why bigram frequency receives this much attention when there is little empirical support for the notion that it has any effect on reading processes, nor any consensus about how to measure it.

Why is the myth of bigram frequency still alive?

We propose two possible causes for the current state of affairs regarding bigram frequency. The first is methodological. Using frequentist methods, it is impossible to provide statistical support for the absence of an effect. If a *p*-value exceeds the conventional cut-off of 0.05, a null-hypothesis significance test is considered to provide evidence against the null hypothesis, but failing to reach this

threshold does not provide evidence for it (Dienes, 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009). In the case of bigram frequency, this leaves us in a stale-mate position: no number of studies showing non-significant bigram frequency effects can provide evidence for the absence of an effect. The empirical value of the current paper lies in the use of Bayesian analysis methods. Here, two models are created, one which includes an effect of bigram frequency (H_1 model), and one which excludes it (H_0). The Bayes Factor quantifies the degree to which the data is compatible with H_1 over H_0. If the data is more compatible with H_0 than with the pre-specified H_1, we get support for the null hypothesis (Rouder et al., 2009). This quantification measure is a ratio: if the data is more compatible with the H_1 model over the H_0 model, its value will increase; a value approximately equal to one suggests that both models provide an equally good fit and that more evidence is required for a conclusion. As the data becomes increasingly more likely under H_0 than H_1, the Bayes Factor approaches zero; as the data becomes increasingly more likely under H_1 than H_0, the Bayes Factor approaches infinity. Bayes Factors can be interpreted as a continuous measure of the strength of evidence for or against the presence of an effect, if we contrast a model containing our effect of interest to an identical one which does not contain it. By convention, values above 3 may be interpreted as support for the numerator hypothesis (e.g., the H_1 model). Similarly, Bayes Factor values below 1/3 provide evidence for the denominator hypothesis (e.g., the H_0 model) and against the numerator hypothesis (e.g., the H_1 model; Rouder et al., 2009). The Bayes Factor thus provides the means for a hypothesis testing procedure, which allows us to provide report support for one statistical hypothesis over another.[1]

[1] Note that the statistical debate of hypothesis testing versus parameter estimation is orthogonal to the frequentist-versus-Bayesian cleft. The latter contrast refers to differences in the view of probability: frequentist statistics is concerned with the

The second reason for the resilience of bigram frequency might be theoretical. Despite the existing empirical work, there is little discussion of theoretical implications of bigram frequency effects (or the absence thereof). Bigram frequency is mainly treated as a potential confound variable: rather than questioning its impact, researchers seem to match for it during item selection, as a conservative approach to creating a well-controlled item set. As a consequence, results relating to bigram frequency in published papers are generally mentioned as a side-note, and receive little attention from the reader (see Gigerenzer, 1998, for a discussion of the importance of theories for the survival of data).

The lack of theoretical discussion about bigram frequency is unfortunate, because bigram frequency effects (or their absence) are bound to provide some information about the cognitive processes underlying reading (Chetail, 2015). Currently, we know of no theories that could either be supported or refuted by bigram frequency effects. Given the scarcity of studies focussing on bigram frequencies, the conflicting results, and the question of whether or not the published positive effects are replicable, it is difficult to provide an integrative theory that would account for what we know so far. Though the common approach in psycholinguistic research is to expand on theoretical predictions and test whether the empirical data support the theory, we take here the opposite approach of attempting to establish whether the effect exists, and following up with a theoretical discussion. Our current approach is

---

probability of long-run observed events in light of an assumed fixed parameter, while the Bayesian philosophy allows for the assignment of probabilities to a given statistical hypothesis, thus bypassing the need of an assumed fixed parameter. For the former contrast, psychological scientists are predominantly concerned with hypothesis testing, as evidenced by the wide-spread use of frequentist *p*-values for drawing inferences. A Bayes Factor is a Bayesian equivalent of *p*-values, in the sense that it is used for hypothesis testing. In Bayesian parameter estimation, a posterior distribution is derived by combining the prior probability with incoming data. Some proponents of Bayesian statistics oppose the use of Bayes Factors to confirm or disconfirm statistical hypotheses (see, e.g., Gelman & Rubin, 1995; Kruschke, 2014).

in line with the principle of Hyman's Maxim, which states that one should not try to explain something until one is sure that there is something to be explained (Loxton, 2015; Schmalz, 2015). Once one has established whether the effect exists or not, it can serve as a benchmark for theories and computational models of visual word recognition. Without a consensus about the absence or presence of an effect, there is a risk of models and theories becoming needlessly complex, because they take into account isolated findings which may reflect experimental noise. We reserve an overview of the potential theoretical value of bigram frequency effects for the discussion section, when we can determine whether a psychologically valid theory should predict the presence or absence of bigram frequency effects.

Methods

The empirical aim of this paper is to provide either evidence for or against the bigram frequency effect. To this end, we used Bayes Factor (BF) analyses on existing databases of word reading in English. The analyses were performed in R (Version 3.3.1; R Core Team, 2013), with the package BayesFactor (Version 0.9.12-2; Morey & Rouder, 2014). We seek to provide converging evidence from two approaches: a factorial analysis of a set of matched items, and an analysis of available large-scale databases which statistically controls for covariates.

Each of these two approaches has advantages and disadvantages. A factorial design gives us more control over potential confounds, while large-scale regression analyses provide more power for detecting potential effects or interactions. The problem of inter-correlations between psycholinguistic variables is particularly relevant here: Bigram frequency is theoretically distinct from but necessarily correlated with orthographic neighbourhood (as neighbours, almost by definition, share many bigrams, thus increasing their bigram frequency), morphological

complexity (morphologically complex words have high average bigram frequency, because the letter strings of bound morphemes, such as *-ing*, *-er*, or *-ed*, occur very frequently), and monogram (letter) frequency (as letters that occur often necessarily occur more frequently in bigrams). Although regression models can include these variables as covariates, they are not immune to problems of assigning shared variance to one predictor over the other. As a result, a regression analysis of a strongly correlated item set can either inflate an observed effect by assigning to it variance from a correlated variable, or (of relevance to a model-comparison approach) suppress an effect, by assigning the shared variance instead to the correlated variable.

A factorial design allows us to choose items such that there is variability on the dimension of bigram frequency, while keeping constant the values of correlated variables. As pointed out by Cutler (1981), this restricts the number of possible items that can be used, especially when the potential confounds are correlated with the variable of interest. An additional issue is that selecting a set of well-matched items for an orthogonal design on two correlated variables forces the researcher to choose items with unusual characteristics. This may lead to an item set which is not representative of the actual orthography of interest. In sum, well-matched factorial designs tend to rely on small and unrepresentative sample sizes, which is problematic when drawing inferences unless supplemented by a large-scale analysis. In the current study, we analyse trial-level data from the BLP (i.e., data which is not averaged by participants or by items). Here, an additional caveat is that items and participants are not fully crossed, as not all participants responded to the same items. To address the short-comings of factorial and large-scale approaches individually, we present both a factorial design aimed to assess the influence of bigram frequency, and subsequently a

large-scale database analysis of three different datasets, in an attempt to provide converging evidence.

Items for factorial design

For a factorial analysis, we created two item sets, each of which contained words of different lengths, namely 5 or 7 letters. There were 120 items in total: each length condition contained 30 words with high and 30 words with low bigram frequency. A word was classified as having a high bigram frequency if its count fell above the median bigram frequency for its length, and as low bigram frequency if its count fell below the median. The dichotomisation served for us to be able to match the two groups of items on potential covariates. In general, dichotomisation of continuous variables is undesirable, because it reduces the sensitivity of the analysis. Therefore, in the next section we follow up on the dichotomised analysis with a large-scale database analysis, where we treat bigram frequency as a continuous variable.

We chose to look at the bigram frequency effect while keeping length constant because bigram frequency can be measured either as the sum of the frequency all bigrams within a word, or as the average frequency, where this sum is divided by the number of letters. If we use the former measure, bigram frequency counts are strongly correlated with the number of letters. It is unclear which variable is more closely associated to the potential impact of bigram frequency on behavioural outcomes, but this becomes irrelevant if we look at bigram frequency effects only across words of the same length.

All items were morphologically simple English words from the BLP, and had a lexical decision accuracy of greater than 80%. The item characteristics of the low and high bigram frequency words across the different lengths are presented in Table 1. The items, full data, and R script are available online: https://osf.io/apg4z/.

TABLE 1 ABOUT HERE

Large-scale analyses

We used two mega-studies for the large-scale analyses: (1) the BLP, which contains lexical decision data for bisyllabic words from British undergraduate students, and (2) the ELP, which contains two datasets: one with lexical decision and one with reading aloud data, for words of different lengths, read by American undergraduate students. We removed all words with orthographic prefixes or suffixes (i.e., this included both morphologically complex words, such as "stronger", and pseudo-affixed words, such as "brother"). Note that this creates a quasi-factorial design. Furthermore, we included only words with average lexical decision accuracies of >80%, to exclude words that are unknown to a large proportion of participants. This left us with 4560 words from the BLP and 3857 words from the ELP.

Results

Factorial design

We retrieved the trial-level data for the items from the BLP lexical decision database (Keuleers et al., 2012). The average accuracy and RTs across bigram frequency conditions and number of letters are shown in Table 2. In a Bayes Factor model comparison, we included only correct responses with complete data in all cells. This left us with 6389 out of 6858 observations for the 120 selected words. The raw RT values were transformed into inverse RT (-1000/RT), and we removed one outlier point with an inverse RT < -4. A Q-Q norm plot showed an approximately normal distribution of this trimmed variable[2]. The Bayes Factor models, for each letter length,

[2] Inverse RT transformations are common in psycholinguistic research to reduce the skew of raw RTs (e.g., Baayen, 2008); the advantage of this measure in reading

included previous RT, and items and subjects as a random effect. We compared these base models against models which, in addition, contained the effect of bigram frequency. Bigram frequency, in line with the item set matching procedure, was coded as a binary variable, with high versus low bigram frequency items coded as 0.5 and -0.5, respectively. This contrast coding serves to obtain a slope estimate for average values of bigram frequency. In the absence of such contrast coding, R outputs the slope estimates for the condition which comes first in alphabetical order (i.e., "high" if we were to label the two conditions as "high" and "low"). Repeating the analyses with bigram frequency as a continuous predictor did not change any of the results. Overall, we had 2204 data points for the five-letter analysis and 2177 for the seven-letter analysis. The Bayes Factor value provided evidence against a model containing the bigram frequency effect, with Bayes Factors below the conventional cut-off of 1/3: For a model including bigram frequency model, BF = 0.18 (±1.88%) for five-letter words, and BF = 0.18 (±1.25%) for seven-letter words.

TABLE 2 ABOUT HERE

Previous simulations have shown that the default settings provided by the BayesFactor package may provide evidence for a null hypothesis in a case where there is a small real effect in the population (Simonsohn, 2015). To address this potential concern, we re-did the analysis above while changing the prior scale for the standardised effects from the default 0.5 to 0.1. This reduces the width of the prior Cauchy distribution, making larger values of the effect size less likely. The small-prior analyses also provided evidence against an effect of bigram frequency, BF < 0.2.

---

research is that they give a practically meaningful value, namely the number of words read per second. Using raw RTs did not change any results.

Large-scale analyses

In this second set of analysis, we adopted a large-scale database analysis approach. To build a model which would provide a good description of the obtained data, we started with a linear model which contained only the previously established effect of word length, word frequency, and their interaction (Coltheart, Rastle, Perry, Langdon & Ziegler; Weekes, 1997). We continued with a model-building procedure, where we added variables which are correlated with bigram frequency, but theoretically distinct. We included them in the model if BF > 1, suggesting that the data was more in line with the presence than the absence of an effect.

Bigram frequency is correlated with orthographic neighbourhood, or the number of words with a similar spelling. This can be quantified by Coltheart's N (Coltheart et al., 1977), where the N is the number of neighbours, or real words that can be created by substituting one letter in any position (e.g., *pat, cut, cap* are neighbours of the word *cat*). For polysyllabic words, analyses have shown that an alternative quantification explains more variance of response latencies: the OLD20 measure calculates the average number of substitutions, additions or deletions that need to be undertaken to get the distance from a target word to its nearest 20 neighbours. We compared the OLD20+Frequency*Length model against the Frequency*Length model with a Bayes Factor, and used the cut-off of BF = 1 to decide whether to retain the effect of OLD20. In the next step, we included the effect of monogram (or letter) frequency: again, this variable is correlated with bigram frequency, but reflects a slightly different theoretical construct. We retained this effect in the model if BF > 1. We compared the final model which additionally included the effect of bigram frequency against the final model without the effect of bigram frequency.

For the BLP database, we obtained strong evidence for an effect of OLD20 in addition to the length and frequency main effects and interaction, BF > 200,000. For a model also containing monogram frequencies, we obtained weak evidence against this effect, BF = 0.78. Comparing the model with the length and frequency effects and interaction plus OLD20 and bigram frequency against an identical one excluding the effect of bigram frequency, we obtained evidence against an effect of bigram frequency, BF = 0.22. Changing the prior distribution by scaling the default width parameter to 0.1 still gave us evidence against the effect of bigram frequency, BF = 0.22. As bigram frequency and OLD20 are correlated, it is possible that OLD20 explains some of the variance that is, in reality, attributable to bigram frequency. However, when we compared a model with only the length and frequency effects and interaction plus bigram frequency against one excluding bigram frequency, we still obtained evidence against an effect of bigram frequency, BF = 0.03.

For the ELP lexical decision data, we again obtained strong evidence for an effect of OLD20 in addition to the length and frequency effects and interaction, BF > 1,000,000. There was also weak evidence for an effect of monogram frequencies, BF = 1.12. In the final model comparison, including bigram frequency, monogram frequency, OLD20, and the length and lexicality effects and interaction, we get weak evidence against the effect of bigram frequency, BF = 0.73. Again, to establish whether the absence of a bigram frequency may be attributable to some of the covariates, we conducted model comparisons (1) excluding OLD20, and (2) excluding monogram frequency. For the former, we obtained evidence against the model Length*Frequency + monogram frequency + bigram frequency, versus an identical model excluding bigram frequency, BF = 0.04. For the latter, we obtained evidence for a model including the length and frequency effects and interaction, plus

bigram frequency and OLD20, compared to one excluding the effect of bigram frequency, BF = 19.10. To establish the directionality of the bigram frequency effect, we generated the slopes for the length*frequency + OLD20 + bigram frequency linear model. The slope was positive (estimate = 0.0003, SE = 0.0001), suggesting an inhibitory effect, where words with higher bigram frequency require more processing time.

For the ELP reading aloud data, we obtained very strong evidence for the presence of an effect of OLD20 in addition to the length and frequency effects and interaction, BF > 1,000,000. Adding monograms also increased model fit, BF > 7,000, as did adding the effect of bigram frequency, BF > 5,000. Here, the bigram frequency slope was negative (estimate = -0.0004, SE = 0.0001), suggesting a facilitatory effect, where items with a high bigram frequency count are processed faster.

Consistency of results across different bigram frequency counts

As mentioned in the introduction, different psycholinguistic packages provide sometimes vastly different frequency counts. Therefore, it is of interest to determine whether the results from the previous section are stable regardless of the source of the bigram frequency counts. We repeated the analyses on the BLP and ELP lexical decision databases, using different type frequency counts: counts provided by the BLP (summed frequency), summed and mean counts provided by the ELP, and counts provided by N-watch (averaged frequency; Davis, 2005). For both databases, we calculated the Bayes Factor for a model containing the length and frequency main effects and interaction, OLD20, and bigram frequency against a model excluding bigram frequency. The counts were not available for all items. For the BLP database, the ELP contained bigram frequency counts for 4448 items, and N-watch for 4534 (as

it cannot calculate bigram counts for words with two letters or less, or more than 11 letters). For the ELP database, the BLP provided items for 3219 items, and N-watch for 3823.

For the BLP database, we obtained evidence against an effect of bigram frequency for all counts, for the BLP (summed) counts BF = 0.22, for N-watch (averaged) counts BF = 0.1, for ELP averaged counts BF = 0.04, and for ELP summed counts BF = 0.07. For the ELP database, we obtained evidence for an effect using the N-watch (averaged) counts, BF = 5.9 (positive slope, estimate = 0.0004, SE = 0.0001), and for the ELP summed counts, BF = 46 (positive slope: estimate <0.0001, SE < 0.0001). For the ELP averaged counts, the evidence for the presence of a bigram frequency effect was equivocal, BF = 0.70, and for the BLP (summed) counts, there was evidence against an effect, BF = 0.21.

Discussion

Here, we set out to resolve a conflict in the literature on bigram frequency effects in visual word recognition. We got clear-cut results from the BLP database, with all analyses suggesting the absence of a bigram frequency effect. However, the results from the ELP database proved less consistent. A factorial approach, using data from the BLP, suggests that there are no bigram frequency effects in lexical decision, as does an analysis of the whole database. The absence of the bigram frequency effect in the BLP database is robust to changes in the prior distribution, in the constellation of predictor variables, and holds regardless of the database used to obtain the bigram frequency counts. The data from the ELP suggests that there is a facilitatory bigram frequency effect in reading aloud, and shows inconsistent inhibitory effects in lexical decision. Lexical decision and reading aloud involve different cognitive processes;

thus, it is perhaps not surprising that we get different results. We will discuss the theoretical reasons for a task difference in a later section.

At face value, the difference between the ELP and BLP lexical decision data is difficult to reconcile. However, there is an important systematic difference in the nonword foils: the ELP database used manually created nonwords, while the BLP used computationally generated nonwords, which were specifically designed to be matched to words on sublexical characteristics such as bigram frequency. These characteristics of words versus nonwords may be picked up by participants, throughout the duration of the experiment, and can act as an additional, non-lexical cue to facilitate lexical decision. The nonwords of the ELP were created by exchanging letters from real words, which made them very similar to words. Using a set of simulations, Keuleers and Brysbaert (2011) showed that the similarity creates a bias, where nonwords become increasingly more likely to be perceived a more word-like than words. The algorithm, called LD1NN, is based on a Levenshtein metric (the number of subsitutions, additions, or deletions between letter strings), which is calculated, for each upcoming item, based on all previously presented items. This metric is likely to be correlated with bigram frequency, however, because items which are similar to many previously presented items are likely to contain frequency letter bigrams. Thus, within the context of the ELP lexical decision experiments, participants may have developed a bias, where high-frequency bigrams were associated with a “no”-response and low-frequency bigrams with a “yes”-response.

We tentatively attribute the presence of an inhibitory bigram frequency effect in some of the ELP analyses to this systematic difference between the databases, as all analyses of the BLP database showed consistent evidence for the absence of a bigram frequency effect. An alternative interpretation would be that, for whatever

reason, our analyses of the BLP database were not sensitive to pick up a true effect. This could be, for example due to our specific constellation of independent variables or choice of bigram frequency counts. Thus, confirming whether our interpretation is correct would require future empirical work. A straight-forward prediction based on our interpretation is that we should find a facilitatory bigram frequency effect in the later trials of a lexical decision experiment when words have higher bigram frequency than nonwords; and conversely, a bigram frequency effect should be inhibitory towards the end of an experiment when words have lower bigram frequency than nonwords.

Given our interpretation, the results suggest the absence of a bigram frequency effect in lexical decision, provided that the items are matched across lexicality and bigram frequency cannot be used as a non-lexical cue to derive a correct lexical decision response. Thus, it seems that when creating item sets for a lexical decision experiment, matching for bigram frequency may not be not necessary. As bigram frequency is correlated with many psycholinguistic variables that are of theoretical interest (e.g., morphological complexity, regularity, orthographic neighbourhood), this will provide more flexibility for creating larger item sets. However, our interpretation stresses the importance of matching words and nonwords on sublexical characteristics, in order to avoid a non-lexical cue which may bias lexical decision responses (Keuleers & Brysbaert, 2011). In the BLP, the results were stable regardless of the bigram frequency measure, while there were some differences in the ELP analyses. It is noteworthy that we obtained the strongest Bayes Factor favouring the presence of a bigram frequency effect in the ELP database when we used the ELP counts: the finding that the effect is strongest for counts that were derived from the same database is in line with the suggestion that the bias reflects sensitisation to

sublexical characteristics of the items that are presented throughout the experiment (Keuleers & Brysbaert, 2011). The effect of bigram frequency found for the other measures would then reflect a correlation between the frequency of bigrams presented during the experiment and the corpus-derived frequency counts.

In reading aloud, the ELP data support a model including the effect, where words with higher bigram frequencies are associated with faster responses. At face value, a bigram frequency effect for reading aloud but not lexical decision might suggest that it reflects a sublexical decoding process. Lexical decisions can, in principle, be performed purely based on lexical activation, while reading aloud also requires computing a pronunciation. In this case, we would also expect an interaction of bigram frequency and word frequency: for high-frequency words, lexical access occurs more quickly, and sublexical processes are relatively less important than for low-frequency words (Coltheart, et al., 2001). However, comparing the model from the results section plus a bigram frequency by word frequency interaction against a base model (Length*Frequency + monogram frequency + bigram frequency) provided evidence against such an interaction ($BF < 0.0001$; excluding monogram frequency from this model: $BF = 0.96$).

It is possible that the significant bigram frequency effect in reading aloud but not lexical decision reflects a phonological effect, which might be more accurately measured by biphone frequency. This would be in line with research on biphone effects in spoken word perception. Here, studies show that a high phonological neighbourhood (which is strongly and positively correlated with biphone frequency) increases reaction times, due to competition between the target word and similar words. However, when the task involves only nonwords and lexical competitors

become irrelevant to task performance, biphone frequency has a facilitatory effect on response latencies (for a summary, see Vitevitch & Luce, 2016).

Similarly to our results of a bigram frequency effect in the reading aloud database, Andrews (1992) found an effect of bigram frequency in reading aloud, though the effect was not found in a subsequent experiment where the items were matched for first phoneme characteristics. Despite the methodological difference between the two experiments, it is possible that the bigram frequency effect of the first reading aloud experiment reflected a true bigram or biphone frequency effect and the latter was a false negative. If this is the case, it is noteworthy that the bigram frequency effect was found both for immediate and for delayed naming, suggesting an effect on the articulatory level. This would contradict the proposed explanation from the previous paragraph, that biphone frequency may facilitate speech production on a sublexical phonological level. Future research should focus on manipulating biphone frequency in oral language tasks, in order to establish whether biphone effects have an effect on speech production, and if so, whether this effect interacts with lexicality or frequency, suggesting a sublexical phonological locus, or if it is stable across delayed and immediate naming tasks, suggesting an articulatory locus.

From a theoretical perspective, the absence of a bigram frequency effect in lexical decision can serve as a benchmark for theories and models of reading. We know of no descriptions of current models of visual word recognition that make explicit predictions about the presence or absence of bigram frequency effects across tasks, but the absence of such an effect provides valuable information about whether specific processes are likely to be psychologically valid or not. The theoretical potential of orthographic redundancy has been raised in a recent review (Chetail, 2015). Orthographic redundancy is a broader concept than bigram frequency, as it

includes measures such as the existence of a bigram in the orthography (e.g., the bigram *xx* never occurs in English, while the bigram *ll* occurs in words like *yellow*) and position-specific letter frequency. The role of orthographic redundancy or bigram frequency may emerge at the early stages of word processing, by facilitating letter identity or position coding (Frankish & Barnes, 2008; Perea & Carreiras, 2008). At the later stages, bigram frequency may also be involved in parsing words into linguistically meaningful orthographic units. Furthermore, low bigram frequency may make words more distinct from other lexical entries and thus facilitate lexical access.

Although our results are not consistent with the notion of a lexical locus of a bigram frequency effect, future research is needed to determine whether an effect may emerge for tasks tapping letter identity and position coding. Some findings might indicate such an early-processing influence: a neuroimaging study that has found an effect of bigram frequency in a non-lexical task on activation in the visual word form area (Binder, Medler, Westbury, Liebenthal, & Buchanan, 2006). Binder et al. (2006) used a task which specifically discouraged any type of lexical processing: Their items were unpronounceable consonant strings with very low overall bigram frequency counts (mean positional bigram frequency averages of 0, 25, 147, and 463 across four conditions). The task was based on the stimuli's visual features: the participants responded each time that the consonant string contained an ascending letter (e.g., *t*, *d*, as opposed to *q*, *m*). The behavioural data showed that this task became easier across conditions increasing in bigram frequency, thus providing some support for a bigram frequency effect on early visual processes, though, due to the nature of the task, it is unclear whether this is associated with word processing.

Turning to existing theories of reading processes, letter bigrams play a major role in some recent theories of orthographic coding. Open bigram theories (Grainger

& Whitney, 2004; Schoonbaert & Grainger, 2004; Whitney, 2001) propose that coding letters and their relative order is achieved by processing a word as a series of open bigrams (but see also Kinoshita & Norris, 2013). These open bigrams are not only letter pairs that are adjacent (e.g., *ca* and *at* for the word *cat*), but also non-adjacent letter pairs, which can have one or two intervening letters (*ct* in *cat* and *cart*, but not in *court*). The strength of each open bigram is proposed to depend on the distance between the two letters, such that open bigrams are activated to a lesser extent when there are intervening letters (Whitney, 2008). In these models, one might expect that bigram frequency should matter: if adjacent letter pairs co-occur often, the system might be facilitated in coding these, resulting in faster word recognition. Whether this prediction indeed falls out of the open bigram framework can be tested in future work, using computational implementations. It is worth noting that open bigrams are proposed to reflect an orthographic process that is directly linked to whole-word access, as opposed to a sublexical decoding procedure (Grainger & Ziegler, 2011). Therefore, if anything, the open bigram account would predict a bigram frequency effect for lexical decision but not reading aloud, not vice versa.

A related proposal about bigram processing has been made by Dehaene and colleagues (Dehaene, 2009; Dehaene, Cohen, Sigman, & Vinckier, 2005). Here, the authors suggest that visual word recognition of a single word reflects the hierarchical structure of the brain: at the lowest level of the visual system, the receptive fields are small, and code basic visual features (e.g., a horizontal and a vertical line for the letter *T*). Further up the visual system, the receptive fields get progressively larger, which allows them to respond to letters, followed by letter combinations (i.e., bigrams), and, finally, whole words. Bigram neurons are proposed to selectively respond to a given bigram. Therefore one might predict that bigram frequency should have a facilitatory

effect on visual word recognition: if a bigram occurs often, its corresponding receptive field might be well-connected and/or have a low activation threshold. Conversely, if a bigram occurs rarely, it may not even have a dedicated bigram neuron. Although neither Dehaene's bigram neuron hypothesis, nor the open bigram views make explicit statements about effects of bigram frequency, the absence thereof in a lexical decision task provides a challenge for these theories. Future theoretical work might be useful to clarify whether and how bigrams can be coded during orthographic processing without being dependent on bigram frequency.

Conclusion

We set out to address an issue which has been in the background of reading research for half a century, namely, whether bigram frequency has an effect on visual word recognition. Using Bayes Factor analyses, we found evidence against an effect of bigram frequency in lexical decision data of the BLP. The ELP showed inconsistent inhibitory effects of bigram frequency, which we attribute to sublexical characteristics of the nonwords used in this experiment (Keuleers & Brysbaert, 2011). This interpretation suggests that researchers do not need to include bigram frequency when they are matching items for a lexical decision task, provided that there are no systematic differences between words and nonwords.

From a theoretical perspective, showing the absence of a bigram frequency effect in lexical decision also establishes a benchmark which needs to be accounted for by theories and models of orthographic processing. More generally, we argue that it is important not only to search for variables that have psychological reality, but also to determine which variables have no effect on cognitive processing: When two theoretical models make contrasting predictions about the presence or absence of an effect, a null result is theoretically just as interesting as a positive result. Establishing

evidence against an effect is, in practice, more challenging. A combination of experimental approaches, as well as the use of Bayes Factors to provide evidence for the null hypothesis against an alternative, may help researchers to argue for theoretically meaningful null-results.

References

Andrews, S. (1992). Frequency and neighbourhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory & Cognition, 18*(2), 234-254.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge: Cambridge University Press.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445-459. doi: 10.3758/Bf03193014

Biedermann, G. (1966). The recognition of tachistoscopically presented five-letter words as a function of digram frequency. *Journal of Verbal Learning and Verbal Behavior, 5*, 208-209.

Binder, J. R., Medler, D. A., Westbury, C. F., Liebenthal, E., & Buchanan, L. (2006). Tuning of the human left fusiform gyrus to sublexical orthographic structure. *Neuroimage, 33*(2), 739-748.

Broadbent, D., & Gregory, M. (1968). Visual perception of words differing in letter digram frequency. *Journal of Verbal Learning and Verbal Behavior, 7*(2), 569-571.

Chetail, F. (2015). Reconsidering the role of orthographic redundancy in visual word recognition. *Frontiers in Psychology, 6*(645), 1-10. doi: 10.3389/fpsyg.2015.00645

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*(8), 1519-1540.

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

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256. doi: 10.1037//0033-295x.108.1.204

Cutler, A. (1981). Making up materials is a confounded nuisance: or Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition, 10*(1-3), 65-70. doi: 10.1016/0010-0277(81)90026-3

Davis, C. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37*(1), 65-70. doi: 10.3758/bf03206399

Dehaene, S. (2009). *Reading in the brain: The new science of how we read*. London: Penguin.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in Cognitive Sciences, 9*(7), 335-341. doi: 10.1016/J.Tics.2005.05.004

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5,* 1-17. doi: 10.3389/fpsyg.2014.00781

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods Instruments & Computers, 36*(3), 488-499. doi: 10.3758/bf03195595

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*(2), 488-496.

Frankish, C., & Barnes, L. (2008). Lexical and sublexical processes in the perception of transposed-letter anagrams. *The Quarterly Journal of Experimental Psychology, 61*(3), 381-391.

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25*, 165-173.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General, 113*(2), 256-281.

Gigerenzer, G. (1998). Surrogates for theories. *Theory and Psychology, 8*, 195-204.

Grainger, J., Dufau, S., Montant, M., Ziegler, J. C., & Fagot, J. (2012). Orthographic processing in baboons (Papio papio). *Science, 336*(6078), 245-248.

Grainger, J., & Whitney, C. (2004). Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Sciences, 8*(2), 58-59.

Grainger, J., & Ziegler, J. (2011). A dual-route approach to orthographic processing. *Frontiers in Psychology, 2*, 1-13. doi: 10.3389/fpsyg.00054.

Keuleers, E., & Brysbaert, M. (2011). Detecting inherent bias in lexical decision experiments with the LD1NN algorithm. *The Mental Lexicon, 6*(1), 34-52.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, *1*, 1-15. doi: 10.3389/fpsyg.2010.00174

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*(1), 287-304. doi: 10.3758/S13428-011-0118-4

Kinoshita, S., & Norris, D. (2013). Letter order is not coded by open bigrams. *Journal of Memory and Language, 69*(2), 135-150.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London: Academic Press.

Loxton, D. (2015, May 24th). History and Hyman's Maxim [Blog post]. Retrieved from http://www.skeptic.com/insight/history-and-hymans-maxim-part-one/

Medler, D. A., & Binder, R. J. (2005). MCWord: An on-line orthographic database of the English language. Retrieved 20.5.2015 from http://www.neuro.mcw.edu/mcword/

McClelland, J. L., & Johnston, J. C. (1977). The role of familiar units in perception of words and nonwords. *Attention, Perception, & Psychophysics, 22*(3), 249-261.

Morey, R. D., & Rouder, J. N. (2014). Package "BayesFactor". Retrieved 9.8.2014, from http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

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*(1), 45-52.

Perea, M., & Carreiras, M. (2008). Do orthotactics and phonology constrain the transposed-letter effect? *Language and Cognitive Processes, 23*(1), 69-92.

R Core Team. (2013). R: A language environment for statistical computing [Computer software manual]. Vienna. Retrieved from http://www.R-project.org/

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T. J., & Reichle, E. D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). *Journal of Experimental Psychology: General, 136*(3), 520-529.

Rice, G., & Robinson, D. (1975). The role of bigram frequency in the perception of words and nonwords. *Memory & Cognition, 3*(5), 513-518.

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

Rumelhart, D. E., & Siple, P. (1974). Process of recognizing tachistoscopically presented words. *Psychological Review, 81*(2), 99-118.

Schmalz, X. (2015, July 17th). Hyman's Maxim: The most important principle in observational sciences? [Blog post]. Retrieved from http://xeniaschmalz.blogspot.de/2015/07/hymans-maxim-most-important-principle.html

Schoonbaert, S., & Grainger, J. (2004). Letter position coding in printed word perception: Effects of repeated and transposed letters. *Language and Cognitive Processes, 19*(3), 333-367.

Simonsohn, U. (2015, September 4th). The default Bayesian test is prejudiced against small effects [Blog post]. Retrieved from http://datacolada.org/2015/04/09/35-the-default-bayesian-test-is-prejudiced-against-small-effects/

van Heuven, W., 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*(6), 1176-1190.

Vitevitch, M. S., & Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics, 2*, 75-94.

Weekes, B. (1997). Differential Effects of Number of Letters on Word and Nonword Naming Latency. *The Quarterly Journal of Experimental Psychology, 50A*(2), 439-456.

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*(2), 221-243.

Whitney, C. (2008). Comparison of the SERIOL and SOLAR theories of letter-position encoding. *Brain and Language, 107*(2), 170-178.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971-979. doi: 10.3738/Pbr.15.5.971

Table 1

Item characteristics for the high- and low- bigram frequency words of different lengths (SDs in brackets)

	5 Letter words		7 Letter words	
	High BiF	Low BiF	High BiF	Low BiF
Summed BiF	45.87 (5.08)	31.81 (6.00)	110.28 (26.12)	70.83 (12.87)
Summed Letter Frequency	329.70 (49.83)	356.31 (74.40)	776.06 (121.09)	802.84 (100.93)
Log Frequency	0.90 (0.84)	0.66 (0.82)	0.19 (0.90)	0.31 (0.97)
OLD20	1.75 (0.10)	1.74 (0.12)	2.44 (0.25)	2.40 (0.27)
Orthographic N	2.93 (1.53)	3.03 (1.99)	0.27 (0.58)	0.53 (0.73)
Number of syllables	1.2 (0.41)	1.1 (0.31)	2 (0)	2 (0)

Note: The orthographic neighbourhood (Coltheart et al., 1977), Orthographic Levenshtein Distance (OLD 20; Yarkoni, Balota, & Yap, 2008) and subtitle frequency (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) values are provided by the BLP database. Bigram and monogram counts are taken from Medler and Binder (2005). These are position-specific sums of the frequencies. BiF = bigram frequency.

Table 2

Behavioural results from the BLP (calculated across item)

	5 letter words		7 letter words	
	High BiF	Low BiF	High BiF	Low BiF
Accuracy (%)	96.0 (5.1)	95.3 (4.6)	95.1 (5.7)	94.5 (5.7)
Reaction times (ms)	582.9 (52.4)	577.0 (49.4)	619.4 (67.2)	614.7 (53.5)