A word or two about nonwords: frequency, semantic neighborhood density, and orthography-to-semantics consistency effects for nonwords in the lexical decision task

Peter Hendrix
University of Tübingen

Ching Chu Sun
University of Tübingen

Abstract

For the most part, the effects of lexical-distributional properties of words on visual word recognition are well-established. More uncertainty remains, however, about the influence of these properties on lexical processing for nonwords. The work presented here investigates the mechanisms that guide nonword processing through an analysis of lexical decision latencies for 18,547 words and 27,079 nonwords in the British Lexicon Project (Keuleers, Lacey, Rastle, & Brysbaert, 2012) using piece-wise generalized mixed models (PAMMs; Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, Hartl, Day, & Küchenhoff, 2018). The PAMM analysis of the data revealed two novel effects for nonwords in the lexical decision task. First, whereas previous studies reported effects of base word frequency, the current study is the first to document a true nonword frequency effect. Second, we report effects of semantic neighborhood density and orthography-to-semantics consistency; not only for words, but also for nonwords. The effects of frequency, semantic neighborhood density and orthography-to-semantics consistency are facilitatory for words, but inhibitory for nonwords. The PAMM analysis offers insights into the temporal development of the effects of lexical-distributional variables that are not available through more traditional analysis techniques and that shed new light on lexical processing in visual word recognition tasks. The implications of the reported results for models of visual word recognition are discussed.

Keywords: nonwords, nonword frequency, semantic neighborhood density, orthography-to-semantics consistency, lexical decision

Introduction

The perhaps most well-known experimental task in psycholinguistic research is lexical decision (Meyer & Schvaneveldt, 1971). Participants are presented with a sequence of letters and are asked to decide if this sequence of letters is a real word or a nonword. Typically, the analysis of lexical decision data focuses on the response times and the accuracy of the responses for real words. The effects of lexical-distributional properties of words on the performance in the lexical decision task are well-documented and by and large undisputed. Commonly reported effects include the effects of word frequency (Forster & Chambers, 1973; Murray & Forster, 2004; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Keuleers et al., 2012), word length (O'Regan & Jacobs, 1992; Hudson & Bergman, 1985; New, Ferrand, Pallier, & Brysbaert, 2006), and orthographic neighborhood density (i.e., the number of words that are orthographically similar to a word; Yarkoni, Balota, & Yap, 2008; Keuleers, Diependaele, & Brysbaert, 2010; Andrews, 1989, 1992, 1997; Forster & Shen, 1996). The response patterns to nonwords in the lexical decision task have received considerably less attention, although several studies that specifically investigate lexical processing of nonwords in the lexical decision task do exist (Whaley, 1978; Perea, Rosa, & Gómez, 2005). Recently, Yap, Sibley, Balota, Ratcliff, and Rueckl (2015) made a substantial contribution to the nonword reading literature through a large-scale regression analysis of the responses to nearly 37,000 nonwords in the English Lexicon Project (ELP; Balota et al., 2007). Nonetheless, the amount of studies that investigated nonword processing in visual word recognition is less-than-overwhelming.

There are at least two reasons for the relatively limited number of studies on nonword reading. First, it could be argued that normal language processing concerns word reading, not nonword reading. Nonword reading, some may argue, is restricted to the context of artificial experimental tasks. In these tasks, nonwords serve as foils, and are considered uninteresting and uninformative about word recognition processes. The absence (or limited size) of masked identity priming effects for nonwords (cf. Forster, 1998), for instance, is often taken as evidence for the absence of lexical representations for nonwords. The absence of lexical representations for nonwords, however, does not imply the absence of lexical activation for nonwords. A number of studies reported inhibitory effects of orthographic neighborhood density for nonwords in the lexical decision task (Yap et al., 2015; Balota et al., 2004; Carreiras, Perea, & Grainger, 1997; Forster & Shen, 1996; Andrews, 1989; Coltheart, Davelaar, Jonasson, & Besner, 1977). These effects are hard to reconcile with the notion that lexical activation is absent for nonwords altogether. Furthermore, while it may be true that word reading is the default in everyday language use, it is not the case that nonword reading exists in the context of contrived laboratory experiments only. During normal reading, it is quite common to encounter words we are not familiar with. Indeed, processing unknown words is essential in language learning (Norris, 2006; Chaffin, Morris, & Seely, 2001).

Response patterns for nonwords in psycholinguistic experiments provide further information about the mechanisms that underlie nonword processing. In addition, however, the study of experimental data for nonwords has the potential to shed further light on the mental architectures that drive lexical processing for real words. Behavioral patterns in lexical decision latencies for nonwords can thus provide valuable information for the development

of computational models of visual word recognition in the lexical decision task. As an example, the variable deadline for a "yes" response in the multiple read-out model (MROM; Grainger & Jacobs, 1996) for visual word recognition was inspired by the inhibitory effect of orthographic neighorhood density for nonwords (Yap et al., 2015; Balota et al., 2004; Carreiras et al., 1997; Forster & Shen, 1996; Andrews, 1989; Coltheart et al., 1977).

The second reason for the limited number of studies on nonword reading is that the range of lexical-distributional predictors that can be computed for nonwords is more narrow than the range of lexical variables that is available for words. Analyses of response patterns for nonwords have therefore primarily revolved around three concepts: length, orthographic neighborhood density, and base word frequency (cf. Yap et al., 2015). Longer nonwords consistently give rise to longer response times (Yap et al., 2015; Balota et al., 2004; Whaley, 1978), as do nonwords with a large number of real word orthographic neighbors (see above). Base word frequency is an approximation of the frequency of a nonword through either the frequency of the real word it was derived from or the frequency of a nonword's orthographic neighbors. The effects of base word frequency, however, have been less than consistent in previous studies. Yap et al. (2015) and Ziegler, Jacobs, and Klüppel (2001) reported facilitatory effects of base word frequency. By contrast, Andrews (1996) and Perea et al. (2005) documented inhibitory effects. Allen, McNeal, and Kvak (1992) did not observe a significant effect of base word frequency in either direction.

Recent developments have extended the set of lexical predictors that can be computed for nonwords. First, the exponential growth of the world wide web has made available an enormous corpus of real life language use. Lexical-distributional properties of words in this corpus can be gauged through information obtained from graphical user interfaces to algorithms that allow for a comprehensive search of the world wide web, such as Google search. As a result, it is now possible to obtain true frequency counts for nonwords through the number of results returned by a Google search. Frequency effects for nonwords therefore no longer need to be approximated through frequencies of orthographically similar real words. In this paper, we report a robust effect of the Google frequency of nonwords with a considerable effect size. This effect of the Google frequency of a nonword is a better predictor of response patterns in the lexical decision task than is base word frequency and that the effect of nonword frequency cannot be reduced to effects of base word frequency, component letter n-gram frequencies, nonword length, orthographic or semantic neighborhood density, or orthography-to-semantics consistency.

Second, recent advances in the field of distributional semantics have made it possible to compute distributed semantic representations for out-of-vocabulary words (i.e., words that are not in the training data) and, by extension, for nonwords. Specifically, `fastText`, which is an extension of the continuous skip-gram model in `word2vec` (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Chen, Corrado, & Dean, 2013), is able to compute semantic vectors for nonwords on the basis of semantic vectors of the component letter n-grams in a nonword. The semantic vectors for nonwords generated by `fastText` allow for the computation of indices of the lexical-distributional properties of semantics for nonwords. An example of a measure that proxies the semantics of nonwords is semantic neighborhood density, which taps into the number of words that are semantically similar to a (word or) nonword. A second measure of the semantic information associated with a nonword is orthography-to-semantics consistency, which gauges the semantic similarity between a word

and orthographically similar words (cf. Marelli & Amenta, 2018). Effects of both semantic neighborhood density (Buchanan, Westbury, & Burgess, 2001; Pexman & Hargreaves, 2008; Shaoul & Westbury, 2010) and orthography-to-semantics consistency (Marelli, Amenta, & Crepaldi, 2015; Marelli & Amenta, 2018) have been reported for words in the lexical decision task. Here, we investigate the effect of both measures for nonwords.

Specifically, we investigate the effects of frequency, semantic neighborhood density, and orthography-to-semantics consistency as well as well as the effects of length, orthographic neighborhood density, and mean bigram frequency for 18,547 words and 27,079 nonwords in the British Lexicon Project (henceforth BLP; Keuleers et al., 2012). We analyse the lexical decision latencies using a statistical technique from time-to-event analysis, the piece-wise exponential additive mixed model (henceforth PAMM; Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, et al., 2018). PAMMs fall under the statistical umbrella of time-to-event analysis. Time-to-event analysis techniques model the time until an event of interest occurs. Time-to-event analysis has a rich history in medicine and mechanical engineering, in which it is also known as survival analysis. Events of interests in these fields may be the death of a patient or the failure of a mechanical device. Here, we apply time-to-event analysis to the lexical decision data from the BLP. The event of interest in the lexical decision task is the response of a participant to a word or nonword stimulus. Rather than the response time itself, the dependent variable in a time-to-event analysis of lexical decision data is the probability of a response as it evolves over the response time window.

The advantage of time-to-event analysis in the context of lexical decision data is the ability to model predictor effects as a function of time - even though the response variable (i.e., the reaction time) is unidimensional in nature. Predictor effects thus may vary as a function of time in both a quantitative (i.e., "the effect of word frequency is most prominent during the early stages of the response window") and a qualitative fashion (i.e., "the effect of word frequency is facilitatory during the early stages of the response window, but inhibitory during the late stages of the response window"). As such, time-to-event analysis offers the opportunity to gain insight into the temporal development of language processing in the lexical decision task.

A number of recent studies have started to explore the potential of time-to-event analysis for the analysis of psycholinguistic data sets. Typically, the time-to-event analyses in these studies involve a comparison of survival functions (i.e., the number of stimuli for which the event of interest did not yet occur at time t) or hazard curves (i.e., the instantaneous probability of the event of interest at time t, provided that it did not occur prior to time t) for dichotomized versions of numerical predictors. In eye-movement research, for instance, the event of interest can be defined as the end of a fixation of the eye. Investigations of eye fixation durations during reading have shown divergences in the survival functions for high and low frequency words as early as 112 ms after fixation onset (Sheridan, Rayner, and Reingold (2013), see also Reingold, Reichle, Glaholt, and Sheridan (2012)) and divergences in the hazard curves for quantiles of the word frequency and word length distributions between 140 and 250 ms after fixation onset (Feng, 2009). Schmidtke, Matsuki, and Kuperman (2017) applied a similar analysis technique to investigate the relative timing of predictor effects for morphologically complex words (e.g., "goodness"; "good" + "ness") in the lexical decision task. Schmidtke and colleagues found an early divergence of the survival curves

for low frequency and high frequency words, followed by concurrent effects of dichotomized morphological and semantic predictors. Schmidtke et al. (2017) argue that this pattern of results challenges theoretical accounts of morphological processing that posit strictly serial processing of morphologically complex words, with form-based processing preceding semantic processing (cf. Taft, 2004).

The applications of time-to-event analysis to psycholinguistic data have uncovered interesting facts about the nature of lexical processing. The conversion of numerical variables into two-level or multi-level categorical variables, however, is less-than-desirable, as is fitting separate objective functions for each level of each lexical variable. Schmidtke (2016, p.160) acknowledges the problems associated with the "multiple models" approach and identifies a number of further shortcomings of the methodology described above, including the inability to account for subject and item related variance. He concludes that "Ultimately, more complex solutions to modelling survival rates of lexical decision responses will be able remedy many of these issues".

Luckily, "more complex solutions" exist in the statistical literature. A well-known and widely used model for time-to-event analysis that allows for an estimation of the influence of numerical variables on the time to the event of interest is the proportional hazards model proposed by Cox (1972). The Cox proportional hazards model assumes that the effects of predictors on the hazard function are constant over time. Extensions of the Cox model, however, have been developed to allow for time-varying predictor effects. Nilsson (2012) compared the fit of a general survival curve, the Cox proportional hazard model, and the extended Cox model in a time-to-event analysis of the duration of eye fixations during newspaper reading (as collected in the Dundee corpus, see Kennedy, 2003). An extended Cox model with six lexical predictors (including word frequency and word length) led to a substantial reduction in prediction error for held-out data as compared to both the general survival curve and a standard proportional hazards Cox model with the same set of predictors. The effects of the lexical predictors were most prominent between 175 and 225 ms after the onset of the fixation. Nilsson (2012) concluded that lexical variables have non-trivial effects on the time to the end of a fixation, and that the strength of these effects varies considerably as a function of time. An alternative implementation of a time-to-event analysis technique that allows for time-varying predictor effects is the additive hazards model proposed by Aalen (Aalen (1980), see also Aalen (1989, 1993); Scheike and Martinussen (2006)).

The extended Cox proportional hazards model and the Aalen additive hazards model allow predictor effects to vary in a non-linear manner as a function of time. By default, the effects of predictors themselves, however, are linear in both models. Workarounds that allow for non-linear predictor effects have been developed, but typically require the explicit specification of the functional form of a predictor effects, either through a (semi-)automated identification procedure (cf. Branders, Frénay, & Dupont, 2015; Perera & Tsokos, 2018) or with the use of domain-specific knowledge that is available prior to analysis. The statistical technique for time-to-event analysis adopted here, the PAMM, is an extension of the generalized additive mixed-effect model (GAMM; Wood, 2011, 2017). As such, it straightforwardly allows for non-linear predictor effects that develop in a non-linear manner over time. The functional form of non-linear effects in both the predictor dimension and the time dimension is estimated automatically through generalized cross validation or (restricted) maximum likelihood measures.

PAMMs offer detailed insight into the temporal development of non-linear predictor effects in response time analyses. As such, they have the potential to uncover information about the temporal dynamics of visual word recognition in the lexical decision task that can help further the development of models of visual word recognition in the lexical decision task. A detailed introduction to the PAMM is provided in the analysis section of this paper. For the interested reader, we provide the results of a more traditional multiple regression analysis in the discussion section of this paper. The qualitative and quantitative nature of predictor effects is similar in the PAMM analyses reported here and in the multiple linear regression models. The PAMM analyses, however, provides a richer window into the contribution of lexical-distributional predictors to the response times in the lexical decision task.

Methods

Materials

From the British Lexicon Project (henceforth BLP; Keuleers et al., 2012) we extracted average lexical decision latencies for all words for which at least 10 observations with a correct response were available and that had a minimum frequency of 0.1 per million in the SUBTLEX-US corpus (Brysbaert & New, 2009). This resulted in a set of 18,547 words. Furthermore, we extracted average response times for the 27,079 nonwords in the BLP for which at least 10 observations with a correct response were available.

The nonwords in the BLP were created using the pseudoword generator Wuggy (Keuleers & Brysbaert, 2010). Four criteria were followed for the creation of the nonwords: "(1) the nonword matched the syllabic and subsyllabic structure of the target word; (2) it differed from the target word in exactly one subsyllabic segment (onset, nucleus, or coda) for monosyllabic target words and in two subsyllabic segments for disyllabic target words; (3) the transition frequencies of the subsyllabic segments of the target word were matched as closely as possible; and (4) the morphological structure of the word was retained (e.g., if the word was a plural form, we tried to make a matching pseudoplural)" (Keuleers et al., 2012, p. 209). Despite the careful matching of nonwords to target words it is not the case that each nonword in the BLP can easily be linked to its target word by human readers. The first five nonwords in the data set, for instance, are "alcirans", "walfine", "doller", "invost", and "pinchtuck". Out of these five words the word that is most reminiscent of a real word is "doller"; which is highly similar to the real word "dollar". This similarity, however, is coincidental: given criterium (2) for the generation of nonwords, the target word for the nonword "doller" could not have been "dollar".

Design

The response variable under investigation is the average lexical decision latency across participants for the correct responses to the items in the set of words and nonwords described above. For both words and nonwords, we investigate the effects of six lexical predictors: frequency, length, mean bigram frequency, orthographic neighborhood density, semantic neighborhood density, and orthography-to-semantics consistency.

Frequencies for both words and nonwords were obtained semi-automatically through Google searches in the English language. Reported frequencies are the number of results estimated by the Google search engine (i.e., [xxx] in "About [xxx] results"). All frequencies

were collected between August 29, 2019 and September 2, 2019. Of the 27,079 nonwords under investigation, 1,504 had a Google frequency of 0. No less than 94.45% of the nonwords thus had a non-zero frequency on Google. Both word frequencies and nonword frequencies were log-transformed prior to analysis to remove a rightward skew from the frequency distributions. Henceforth, we therefore refer to these frequency measures as *(log) frequency*.

Previous studies furthermore reported effects of base word frequency. Base word frequency refers to the frequency of the real word a nonword is based on. The base words for the nonwords in the BLP are, unfortunately, not publicly available. To overcome a similar hurdle, Yap et al. (2015) approximated base word frequency through the average frequency of a nonword's orthographic neighbors. For the nonwords in the ELP, they found that higher values of this base word frequency measure corresponded to shorter response times. For the current data, however, the effect of a base word frequency measure based on the average frequency of a nonword's orthographic neighbors in a multiple linear regression model including the other predictors under investigation was relatively weak ($t = 2.602$, $p = 0.009$). By contrast, the effect of the Google frequency of a nonword in a similar model was overwhelming ($t = 49.721$, $p < 0.001$). Indeed, the effect of base word frequency has been less than consistent across studies. Whereas, Yap et al. (2015) and Ziegler et al. (2001) observed facilitatory effects of base word frequency, Andrews (1996) and Perea et al. (2005) found inhibitory effects. Consistent with the current findings, Allen et al. (1992) failed to observe an effect of base word frequency in either direction. Given these considerations, we omitted base word frequency from the analyses reported below.

For both words and nonwords, word length is defined as the length of the word in letters. The word length measures will be referred to as *length* throughout the rest of this paper. A number of options exist with respect to measures of the orthographic neighborhood density for words and nonwords. The orthographic neighborhood density measure that is perhaps most well-known is Coltheart's N (Coltheart et al., 1977), which defines the number of orthographic neighbors of a word as the number of words of the same length that differ by one letter from the target word (e.g.; neighbors of the word "bear" include "pear", "hear" and "beak", but not "ear"). For the current data, however, an alternative measure of orthographic neighborhood density, the average orthographic Levenshtein distance between a word and its 20 closest neighbors (henceforth *OLD20*; Yarkoni et al., 2008; Levenshtein, 1996) proved more predictive. The Levenshtein distance between two words is defined as the number of deletions, additions, or substitutions that are necessary to convert one word into the other. The Levenshtein distance between the words "bear" and "part", for instance, is 3 (one substitution, one deletion, and one addition).

The *OLD20* measures were calculated through the `old20()` function in the `vwr` package for R (Keuleers, 2013). For both words and nonwords the 20 closest neighbors were selected from the set of 18,547 real words under consideration. For the word "bear", the set of words under consideration contains over 20 Levenshtein neighbors (i.e., words at a Levenshtein distance of 1; "bar", "bead", "beak", "beam", "bean", "beard", "bears", "beat", "beau", "beer", "boar", "dear", "ear", "fear", "gear", "hear", "near", "pear", "rear", "sear", "tear", "wear", "year"). The value of *OLD20* for "bear", therefore, is 1. The nonword "fretto" has no orthographic neighbors at a Levenshtein distance of 1, 7 neighbors at a distance of 2 ("fresco", "fret", "frets", "ghetto", "grotto", "presto", and "pretty") and over 70 neighbors at a distance of 3. For the nonword "fresco", *OLD20* thus is $\frac{7*2+13*3}{20} = 2.65$.

As was the case for the word frequency distributions, the distribution of orthographic Levenshtein distance was characterized by a rightward skew. Hence, we applied a logarithmic transform to *OLD20* prior to analysis.

The fourth lexical predictor, *mean bigram frequency*, denotes the average frequency of the letter bigrams in a word. We calculated bigram frequencies on the basis of the Google word frequencies for the set of words under investigation. The frequencies of the letter bigrams in the word "bear", for instance, are 213.05 billion ("be"), 469.13 billion ("ea"), and 548.64 billion ("ar"). The value of *mean bigram frequency* for the word "bear" is the average of these frequencies, which is 410.27 billion. For the nonword "fretto", the frequencies of the component letter bigrams are 103.31 billion ("fr"), 808.58 billion ("re"), 245.38 billion ("et"), 95.83 billion ("tt") and 248.66 billion ("to"), for an average bigram frequency of 300.35 billion.

Finally, we included two measures that tap into the semantic information associated with a word or nonword in our analysis of the lexical decision data: semantic neighborhood density and orthography-to-phonology consistency. Both measures are calculated on the basis of distributed semantic representations. Traditionally, distributed semantic representations are most commonly obtained through count-based models. These models define word vectors on the basis of co-occurrence matrices that encode information about the frequency with which words occur in the same linguistic context (Landauer & Dumais, 1997; Lund & Burgess, 1996; Pennington, Socher, & Manning, 2014). Recently, however, prediction-based models have been used to generate semantic vectors as well. The continuous bag-of-words (CBOW) and continuous skip-gram models in `word2vec` (Mikolov, Sutskever, et al., 2013; Mikolov, Chen, et al., 2013) are perhaps the most well-known implementations of the prediction-based approach to distributional semantics. The word vectors obtained through these models maximize the predicted probabilities of all words in the input data given the context words (CBOW) or the predicted probabilities of the context words given each word in the input data (skip-gram).

Both count-based models and prediction-based models are able to generate high quality distributed semantic representations for words that are sufficiently frequent in the input data. The reliability of word vectors, however, rapidly decreases for words that appear in the input data a limited number of times. Furthermore, word vectors cannot be generated for words that are not in the input data. To overcome this limitation, Bojanowski, Grave, Joulin, and Mikolov (2017) proposed an extension of the `word2vec` skip-gram model that is based on an idea introduced by Schütze (1993). This extension of the skip-gram model is referred to as `fastText`.

The idea behind `fastText` is to take subword information into account. More specifically, the representation of a word consists of the full word, as well as all component letter 3-grams to 6-grams. The word "bear", for instance, is represented by the following sequences: "<bear>", "<be", "bea", "ear", "ar>", "<bea", "bear", "ear>", "<bear", and "bear>" (with "<" and ">" representing left and right word boundaries, respectively). Semantic vectors are calculated for each word and for each letter n-gram. Word vectors are defined as the sum of the semantic vectors for the sequences associated with words. The word vector for the word "bear", for instance, is the sum of the semantic vectors for the sequences "<bear>", "<be", "bea", "ear", "ar>", "<bea", "bear", "ear>", "<bear", and "bear>".

The inclusion of subword information allows `fastText` to generate higher quality word vectors for words with low frequencies in the input data, as well as distributed semantic representations for out-of-vocabulary words. One advantage of this is that reliable semantic representations can be obtained for morphologically rich languages such as Finnish. The inclusion of subword information allows the model to exploit the orthographic overlap between members of inflectional paradigms. As a result, robust word vectors can be obtained for low frequency inflectional variants that do not occur or occur only a limited number of times in the input data.

Here, we leverage the ability of `fastText` to generate word vectors for out-of-vocabulary words for a different purpose: the extraction of semantic vectors for nonwords. Assuming a nonword is not present in the input data, a `fastText` model does not contain a word vector corresponding to the whole word form of this nonword. Word vectors for the component letters n-grams that appear in the input data, however, are available. Word vectors for nonwords can be obtained by summing over the semantic vectors for the component letter n-grams that are present in the input data. A word vector for a nonword represents the semantic information that is associated with that nonword. We obtained word vectors for the words and nonwords under investigation from a `fastText` model trained on Wikipedia by Bojanowski et al. (2017).

The first semantic measure included in the analyses is semantic neighborhood density. A number of previous studies defined semantic neighborhood density as the number of words with a cosine similarity to the target word greater than a threshold value (see e.g., Pexman & Hargreaves, 2008; Shaoul & Westbury, 2010). Consistent with the findings of Buchanan et al. (2001), however, a semantic neighborhood density measure based on the average distance between the target word and its k closest semantic neighbors provided maximum explanatory power for the lexical decision data under investigation. Similar patterns of results were obtained across a wide range of values (3 to 50) for the parameter k. A value of 5, however, proved optimal in terms of predictive power. We therefore set k to 5 for the calculation of the semantic neighborhood density measures for both words and nonwords.

The cosine similarity between two words w_1 and w_2 is the cosine of the angle between the corresponding semantic vectors $\vec{v}_1$ and $\vec{v}_2$, which is mathematically defined as the dot product of $\vec{v}_1$ and $\vec{v}_2$ divided by the product of the Euclidean norm of both vectors:

$$\cos(\theta) = \frac{\vec{v_1} \cdot \vec{v_2}}{\|\vec{v_1}\|\|\vec{v_2}\|} = \frac{\sum_{i=1}^{k} v_{1i} v_{2i}}{\sqrt{\sum_{i=1}^{k} v_{1i}^2} \sqrt{\sum_{i=1}^{k} v_{2i}^2}} \tag{1}$$

where k is the length of the word vectors, which was set to 300 in the `fastText` model used here.

Table 1 presents the closest semantic neighbors and their cosine similarity to the target words for two words ("bear" and "pepper") and two nonwords ("fretto" and "guntiors"). The semantic neighborhood density for the word "bear" is the average cosine similarity between "bear" and its 5 closest semantic neighbors "bears" (cosine similarity: 0.67), grizzly (0.63), paw (0.56), lion (0.55), and badger (0.54), which is 0.59. Similarly, the semantic neighborhood density for the word "pepper" is the average cosine similarity between "pepper" and its 5 closest semantic neighbors "peppers" (0.74), "ginger" (0.64), "garlic" (0.63),

Table 1
Closest semantic neighbors with cosine similarities for the words "bear" and "pepper" and the nonwords "fretto" and "guntiors".

word	neighbors
bear	bears (0.67), grizzly (0.63), paw (0.56), lion (0.55), badger (0.54), wolf (0.51), beaver (0.51), moose (0.50), teddy (0.50), bobcat (0.49), raccoon (0.48), bearing (0.48), elk (0.48), dog (0.47), paws (0.47)
pepper	peppers (0.74), ginger (0.64), garlic (0.63), onions (0.63), cloves (0.61), chilli (0.60), beans (0.60), peanut (0.59), peanuts (0.59), spicy (0.59), sauce (0.59), bean (0.59), onion (0.59), pickles (0.58), eggplants (0.58)
nonword	neighbors
fretto	fretting (0.60), frets (0.57), fret (0.57), tempo (0.50), cello (0.50), strings (0.49), strumming (0.48), oboe (0.48), bassoon (0.47), guitars (0.47), chords (0.47), woodwind (0.46), piano (0.46), harp (0.46), flutes (0.45)
guntiors	pushers (0.45), hassling (0.42), vandals (0.41), blindly (0.41), trolling (0.41), trolls (0.40), hounding (0.40), thread (0.40), butting (0.40), users (0.40), assholes (0.40), folks (0.39), bullies (0.39)

"onions" (0.63), and "cloves" (0.61), which is 0.65. The word "pepper" thus lives in a somewhat denser semantic neighborhood than the word "bear".

The closest semantic neighbors for the nonword "fretto" are inflectional variants of the word "fret", which, in addition to a constant state of anxiety, refers to a ridge on the fingerboard of stringed musical instruments. Unsurprisingly, therefore, other semantic neighbors of the nonword "fretto" include a variety of musical instruments and musical terminology. The semantic neighborhood density for the nonword "fretto" is the average cosine similarity between "fretto" and its 5 closest semantic neighbors "fretting" (0.60), "frets" (0.57), "fret" (0.57), "tempo" (0.50), and "cello" (0.50), which is 0.55. The semantic neighbors of the nonword "fretto" may give the impression that semantic neighborhood density is strongly correlated with component n-gram frequencies. It is therefore noteworthy that the correlation between semantic neighborhood density and mean bigram frequency is limited for both words ($r = 0.09$) and nonwords ($r = 0.13$).

The nonword "guntiors" lives in a less-than-pleasant semantic neighborhood, with neighbors such as "pushers" (people who sell illegal drugs), "vandals", "assholes", and "bullies". Provided that we "shall know a word by the company it keeps" (Firth, 1957), our first impression of "guntiors" thus is not the most positive. As before, the semantic neighborhood density for the nonword "guntiors" is the average cosine similarity between "guntiors" and its 5 closest semantic neighbors "pushers" (0.45), "hassling" (0.42), "vandals" (0.41), "blindly" (0.41), "trolling" (0.41), which is 0.42. Henceforth, we refer to the semantic neighborhood density measure through the acronym *SND*.

The second semantic variable under investigation is based on a measure of orthography-to-semantics consistency that was coined by Marelli et al. (2015) (cf. Marelli & Amenta, 2018). Marelli and Amenta (2018) reported effects of orthography-to-semantics consistency (henceforth OSC) on lexical decision latencies and word naming latencies. We define the OSC of a word w as the frequency-weighted average semantic similarity between a word and its k closest orthographic neighbors:

$$\text{OSC}_w = \frac{\sum_{i=1}^{k} \cos(\vec{w}, \vec{n_i})}{k} \tag{2}$$

where k is the number of orthographic neighbors of word w taken into consideration and $\cos(\vec{w}, \vec{n_i})$ is the cosine similarity between word w and word i (see Equation 1).

The set of orthographic neighbors of a word can be defined in a number of ways. Marelli and Amenta (2018) found that defining k as the set of words that embed word w provided optimal explanatory power for the lexical decision latencies for a random subset of $1,821$ words in the BLP. A majority of the nonwords in the BLP, however, is not embedded in real words. A similar definition of the set of orthographic neighbors thus is not feasible for the nonwords under investigation. Marelli and Amenta (2018) found a measure of OSC based on the k closest orthographic neighbors to have significant explanatory power for the lexical decision latencies in the BLP as well. Here, we therefore defined the set of orthographic neighbors of a word or nonword as the k words in the set of 18,547 under investigation with the shortest Levenshtein distance to the target word. We then calculated the OSC of a word on the basis of the frequency counts and semantic vectors described above.

A series of multiple regression models including the other predictors under investigation revealed similar patterns of results for a wide range of values for the parameter k. As was the case for the semantic neighborhood density measure, however, a value of 5 provided maximum explanatory power. We therefore set k to 5 for the calculation of the orthography-to-consistency measures for both words and nonwords. Note that Marelli and Amenta (2018) used a frequency-weighted version of OSC. For the current data, however, a frequency-weighted OSC measure proved substantially less powerful as compared to the OSC measure described above. Prior to analysis, we log-transformed the OSC measure to remove a rightward skew from the distribution of OSC values. Thus, we henceforth refer to the OSC measure as *(log) OSC*.

For each predictor, we removed outliers further than 2.5 standard deviations from the predictor mean prior to analysis. For real words, this led to the exclusion of 794 items (4.28% of the data). More precisely, we removed 136 outliers for *(log) frequency* (0.73%), 131 outliers for *length* (0.71%), 133 outliers for *(log) OLD20* (0.72%), 140 outliers for *mean bigram frequency* (0.75%), 219 outliers for *SND* (1.18%), and 136 outliers for *(log) OSC* (0.73%). For nonwords, the exclusion of predictor outliers resulted in the removal of 1,240 items prior to analysis (4.58% of the data). For *(log) frequency* we removed 9 predictor outliers (0.03%), whereas we removed 204 outliers for *length* (0.75%), 141 outliers for *(log) OLD20* (0.52%), 228 for *mean bigram frequency* (0.84%), 517 outliers for *SND* (1.91%) and 288 outliers for *(log) OSC* (1.06%), respectively. The results of PAMM models fit to the full data set were similar to the results of the PAMM analyses for words and nonwords reported here. The removal of predictor outliers thus had a marginal influence on the reported results.

Table 2 presents the range and adjusted range (after outlier removal) for the lexical predictors under investigation for both words and nonwords. Furthermore, it provides the mean, median, and standard deviation of each predictor after outlier removal. A comparison of the descriptive statistics of the lexical predictors for word and nonwords provides further information about the distributional properties of the stimuli. As can be seen in Table 2, the words and nonwords in the current data are closely matched for length, mean

Table 2
Summary of the predictors (log) frequency, length, mean bigram frequency, OLD20, SND, and (log) OSC for words and nonwords. Range is the original range of the predictor. Adjusted range is the range after removing predictor outliers. Mean, median and sd are the means, medians and standard deviations after outlier removal.

predictor	range	adj. range	mean	median	sd
words					
(log) frequency	11.56 - 23.95	12.76 - 23.16	17.92	17.85	2.02
length	2.00 - 13.00	3.00 - 10.00	6.36	6.00	1.53
mean bigram frequency	2.07 - 878.77	7.21 - 517.22	261.16	259.62	99.09
(log) OLD20	0.00 - 1.79	0.00 - 1.52	0.75	0.67	0.30
SND	0.32 - 0.93	0.42 - 0.82	0.62	0.62	0.08
(log) OSC	-2.73 - -0.20	-2.13 - -0.42	-1.26	-1.24	0.33
nonwords					
(log) frequency	0.00 - 20.56	0.00 - 18.82	8.15	8.83	4.25
length	2.00 - 13.00	3.00 - 10.00	6.61	7.00	1.51
mean bigram frequency	0.18 - 792.68	5.86 - 500.32	251.12	250.16	95.94
(log) OLD20	0.00 - 1.86	0.18 - 1.61	0.88	0.88	0.29
SND	0.28 - 0.76	0.30 - 0.62	0.45	0.45	0.06
(log) OSC	-3.51 - -0.38	-2.24 - -0.58	-1.41	-1.40	0.31

bigram frequency, orthographic neighborhood density, and - perhaps more surprisingly, for orthography-to-semantics consistency.

As expected, however, the average frequency of words (mean *(log) frequency*: 17.92) is higher than the average frequency of nonwords (mean *(log) frequency*: 8.15). Furthermore, the standard deviation is much larger for the nonword frequency distribution (4.25) than for the word frequency distribution (2.02). To gain more insight into the distributional differences between words and nonwords, the frequency distributions for words (solid line) and nonwords (dashed line) before outlier removal are plotted in the left panel of Figure 1. There is some overlap between both distributions. The Google frequency of the low frequency word "perjure" (*(log) frequency*: 12.58), for instance, is lower than that of the high frequency nonword "torb" (*(log) frequency*: 13.85). Nonetheless, the bulk of the probability mass is separated relatively clearly for words and nonwords.

Interestingly, the frequency distribution for nonwords shows a tendency towards multimodality. This could indicate that the nonwords in the BLP consist of different subsets with separate frequency distributions. This is intriguing, given the fact that all nonwords in the BLP were derived from real words following the same principles. As a result, the length, mean bigram frequency, and orthographic neighborhood density of nonwords closely resemble those of real words, as do the parts-of-speech tags of the nonwords. A visual inspection of the nonwords across the frequency range does not provide a clear reason to expect multimodality in the nonword frequency distribution either. The tendency towards multimodality in the nonword frequency distribution is modest and may not replicate to a different set of nonwords generated with the same procedure. Nonetheless, it is an interesting observation that we briefly return to in the discussion section of this paper.

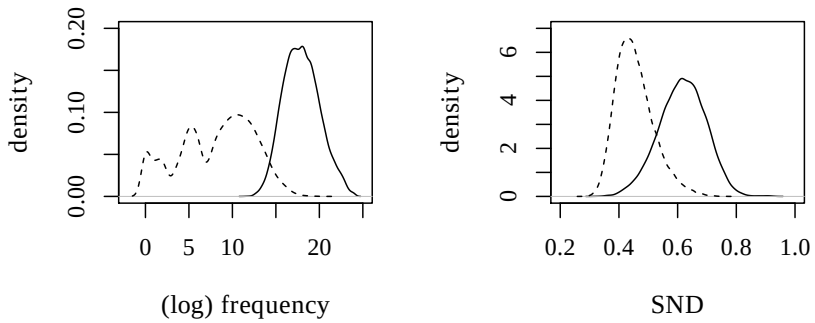

Figure 1. Probability density function of frequency (left panel) and semantic neighborhood density (right panel) for words (solid lines) and nonwords (dashed lines) before outlier removal.

On average, the semantic neighborhood density is higher for words 0.62 than for nonwords 0.45 as well. In addition, the standard deviation is somewhat higher for words 0.08 than for nonwords 0.06. The right panel of Figure 1 presents the distribution of *SND* for words (solid line) and nonwords (dashed line). Both distributions are close to normal. The distribution for nonwords, however, is shifted to the left as compared to the distribution for words. On average, nonwords thus live in sparser semantic neighborhoods than do words.

It is important to clarify that we do not wish to suggest that the location of a nonword in semantic space is stored and that this location is accessed when a nonword is encountered. Instead, we propose that the presentation of a nonword leads to inexorable activation patterns not only in the orthographic and phonological systems, but also in the semantic system (cf. Cassani, Chuang, & Baayen, 2019; Chuang et al., 2019). The semantic neighborhood density measure gauges how similar the activation pattern in the semantic system for a nonword is to the activation patterns in the semantic systems associated with real words. When we talk about the location of a nonword in semantic space, we use the word "location" as a convenient geographical metaphor for an impromptu activation pattern in the semantic system rather than as a reference to a pre-existing locus in semantic space.

Analysis

We analyzed the lexical decision latencies for words and non-words in the BLP with a novel technique for time-to-event analysis: the piece-wise exponential additive mixed model (henceforth PAMM; Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, et al., 2018). As noted above, time-to-event analyses model the time until an event of interest occurs. The event of interest in the lexical decision task is the "yes" or "no" response to a word or non-word stimulus. The dependent variable is the instantaneous probability of a response as it evolves over time.

The application of PAMMs to linguistic response time data is novel (cf. Hendrix, 2018; Hendrix & Sun, 2019; Hendrix, Ramscar, & Baayen, 2019). We therefore introduce the PAMM in more detail below. First, we present two functions of interest in time-to-event analysis: the survival function and the hazard function. Next, we discuss the data

pre-processing that is necessary prior to a PAMM analysis. We furthermore provide an introduction to the PAMM model and its conceptual and statistical properties. We end this section with a description of the PAMMs fitted to the lexical decision data.

Functions of interest in time-to-event analysis

The probability density function describes the relative likelihood of values for a continuous random variable. Visualizations of the probability density function provide information about the distributional properties of a response variable. The probability density function for the response times (i.e., lexical decision latencies) to the words (solid line) and nonwords (dashed line) under investigation is presented in the left panel of Figure 2. As can be seen in the left panel of Figure 2, the distributions of response times for both words and nonwords have a long right tail, as is common in linguistic response time distributions. As compared to the response time distribution for words, the response time distribution for nonwords is shifted to the right. This shift reflects the difference in average response time for words (648.22 ms) and nonwords (666.25 ms) in the data.

The integral of the probability density function is the cumulative distribution function $F(t)$. For a given time t, the cumulative distribution function describes the probability that the response time T is smaller than or equal to t:

$$F(t) = \int_{-\infty}^{t} f(x)dx = P(T \leq t) \tag{3}$$

The central function in time-to-event analysis, the survival function $S(t)$, is closely related to the cumulative distribution function $F(t)$:

$$S(t) = 1 - F(t) = P(T > t). \tag{4}$$

The survival function describes the probability of the time at which the event of interest occurs being greater than a given time t. For the response times in a lexical decision experiment, the survival function thus describes the probability that participants did not yet respond to a word or nonword stimulus at time t.

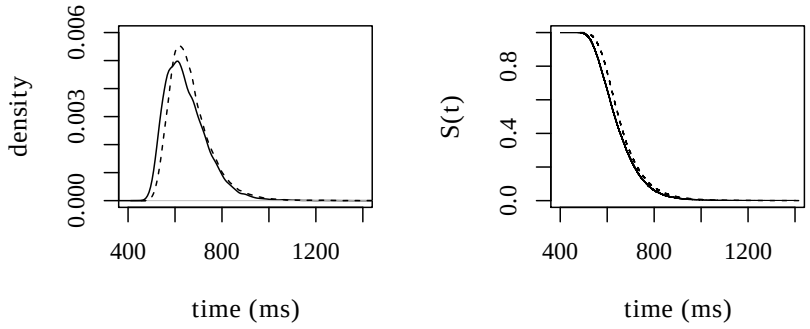

Figure 2. Probability density function (left panel) and survival function (right panel) for the words (solid lines) and nonwords (dashed lines) in the British Lexicon Project.

The survival function for the words (solid line) and nonwords (dashed line) under investigation is presented in the right panel of Figure 2. Before the first response comes in, the probability of "survival" is 1. As responses start coming in, the probability of survival decreases. The probability of "survival" beyond 1000 ms after stimulus onset is extremely low for both words (0.0041) and nonwords (0.0072). As before, the rightward shift of the survival function for nonwords as compared to the survival function for words reflects the fact that response times for nonwords are, on average, somewhat longer than response times for words.

The conceptual objective of time-to-event analysis is to estimate the time until an event of interest occurs. The mathematical properties of the survival function, however, are less than optimal for modeling purposes. Time-to-event analysis techniques therefore typically model the time until an event of interest occurs through a closely related function: the hazard function. The hazard function provides the instantaneous probability that the event of interest occurs at time t, given that it did not occur prior to time t. The hazard function is defined as:

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T \leq t + d \mid T \geq t)}{dt} = -\frac{d}{dt} \log(S(t)). \tag{5}$$

As demonstrated in the vignette for the `pammtools` package for the statistical software R (R Core Team, 2018), PAMMs provide accurate and stable estimates of the hazard function (Bender & Scheipl, 2018). Figure 3 presents the log-transformed hazard function with point-wise confidence intervals for the words (left panel) and nonwords (right panel) under investigation as estimated by a PAMM. For words, the instantaneous probability of a response rapidly increases between 500 and 560 ms after stimulus onset. For nonwords, the increase in instantaneous probability of a response occurs between 515 and 585 ms after stimulus onset. After the initial rapid increase, the instantaneous probability of a response remains high until 900 ms after stimulus onset, at which point in time 98.49% of the words and 97.71% of the nonwords have been responded to. The shape of the hazard functions for words and nonwords is typical for response times distributions in lexical decision experiments (cf. Hendrix, 2018).

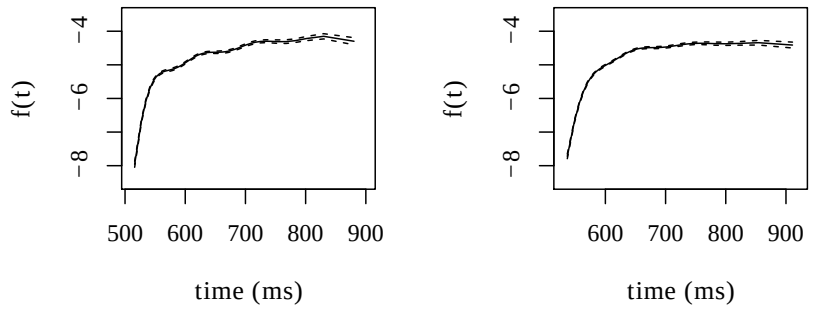

Figure 3. Log-transformed instantaneous hazard function ($f(t)$) with point-wise confidence intervals for the words (left panel) and nonwords (right panel) in the British Lexicon Project as modeled through a piece-wise exponential additive mixed model.

Data pre-processing

The response variable in the lexical decision data from the BLP is the average reaction time for a word across participants. In standard linear or non-linear regression models the dependent variable is this reaction time. By contrast, the dependent variable in piece-wise exponential additive mixed models (PAMMs) is whether or not a stimulus was responded to at time t. To be able to fit a PAMM to the lexical decision data a transformation of the data is therefore required. More specifically, the data need to be presented to the model in a format that Bender and Scheipl (2018) refer to as piece-wise exponential data. When the piece-wise exponential data format is used piece-wise exponential models are essentially Poisson regression models (M. Friedman, 1982). The piece-wise exponential data format therefore allows for the estimation of piece-wise exponential models in regression frameworks such as the generalized additive mixed model (GAMM).

The piece-wise exponential data format splits the time each stimulus is "at risk" of being responded to into J intervals. The intervals $(k_{j-1}, k_j]$, $j = 1 \dots J$ are defined by the cut points $\kappa_0 < \dots < \kappa_J$. For each interval j, the hazard function is assumed to be constant. This assumption explains the name piece-wise exponential model: the (log-transformed) hazard function of an exponential distribution is constant over time. Technically, the hazard function in piece-wise exponential models is thus defined as:

$$\lambda(t) = \lambda(t_j), \quad \forall\, t \in (k_{j-1}, k_j] \tag{6}$$

where t_j typically equals k_j in the PAMM (i.e., the hazard function is estimated for the end-points of each interval).

The choice of the cut points and the number of intervals is arbitrary. The arbitrary selection of cut points and intervals was a limitation for previous implementations of piece-wise exponential models (PEMs) in the context of the generalized linear model. Too few cut points would lead to crude and inaccurate model estimates. Conversely, too many cut points would result in overfitting and unstable estimates (Demarqui, Loschi, & Colosimo, 2008). It is important to note that piece-wise exponential models do not suffer from this limitation in the context of the generalized additive mixed model (GAMM). The implementation of GAMMs in the `mgcv` package for R prevent overfitting through penalization of wiggliness (see Wood, 2011, 2017, for details). As a result, PAMMs generate accurate and stable estimates of the hazard function as long as a sufficiently large number of cut points is used.

One option for defining cut points is to evenly spread out the cut points over the time at which stimuli are "at risk" of being responded to. Here, we instead opted for cut points at the quantiles of the response time distributions, excluding cut points at the extreme ends of the response time distributions. For words, we did not include cut points prior to 500 ms, because a mere 94 words (0.53%) were responded to earlier than 500 ms after stimulus onset. Similarly, we did not include cut points after 975 ms, as no more than 102 words (0.57%) had average reaction times greater than 975 ms. For nonwords, no cut points were included prior to 515 ms (89 nonwords; 0.34%) or after 1060 ms (101 nonwords; 0.39%). For the remaining part of the response time distributions for words (500 - 975 ms) and nonwords (515-1060 ms), we included cut points at 51 quantiles (0 to 1 in steps of 0.02) of the response time distribution. Cut points based on quantiles of the response time distributions have the advantage that more accurate model estimates are obtained for

dense areas of the distributions.

We transformed the data to the piece-wise exponential data format with the `split_data()` function of the `pammtools` package (Bender & Scheipl, 2018). An example of the representation of the lexical decision data for the word "bear" and the nonword "fretto" in the piece-wise exponential data format is presented in Table 3. For each stimulus, the piece-wise exponential data format contains a separate row for each interval. For each row, the start (*tstart*) and end point (*tend*) of the interval are defined. To estimate the development of the hazard function over time, the end point of the intervals is included as a predictor in PAMMs. The column status contains a binary variable that indicates whether (1) or not (0) the word was responded to in that interval. The status, henceforth referred to as δ, is the dependent variable in PAMMs.

Table 3
Piece-wise exponential data format for the word "bear" and the nonword "fretto".

word	tstart	tend	interval	offset	status
bear	0.00	500.19	(0.00,500.19]	6.21	0
bear	500.19	519.16	(500.19,519.16]	2.94	0
bear	519.16	529.39	(519.16,529.39]	2.33	0
...	...	...	...	...	...
bear	558.24	563.03	(558.24,563.03]	1.56	0
bear	563.03	567.03	(563.03,567.03]	1.39	0
bear	567.03	571.15	(567.03,571.15]	0.83	1
nonword	tstart	tend	interval	offset	status
fretto	0.00	515.12	(0.00,515.12]	6.24	0
fretto	515.12	539.21	(515.12,539.21]	3.18	0
fretto	539.21	550.12	(539.21,550.12]	2.39	0
...	...	...	...	...	...
fretto	656.27	660.42	(656.27,660.42]	1.42	0
fretto	660.42	664.87	(660.42,664.87]	1.49	0
fretto	664.87	669.00	(664.87,669.00]	0.17	1

Finally, the piece-wise exponential data format includes an offset (*offset*). For intervals in which a stimulus was not responded to, the offset is equivalent to the log-transformed duration of the interval. The first interval in which "bear" is at risk has a duration of 500 ms. The word "bear" was not responded to in this interval. The offset, therefore, is $\log(500) = 6.21$. For intervals in which a stimulus was responded to, the offset is the log-transformed duration of the period during which the stimulus was "at risk" of being responded to in the current interval. The word "bear" was responded to in the (567.03,571.15] interval, at 569.33 ms after stimulus onset. In this interval, it was at risk of being responded to from 567.03 ms to 569.33 ms, for a duration of 2.30 ms. The offset, therefore, is $\log(2.56) = 0.94$. The offset provides the PAMM with information about the exact response time for each stimulus.

It is noteworthy that in the context of PAMMs stimuli with extreme response times do not need to be excluded prior to analysis. The first interval starts at 0 ms after stimulus onset. Words or nonwords with average response times before the first cut point therefore remain part of the analysis. The exact response times for words or nonwords with average response times after the last cut point are not available to the model. The fact that these

stimuli were not responded to prior to the last cut point, however, is. The absence of a response prior to the last cut point provides the model with valuable information about the nature of words or nonwords that are particularly hard to respond to.

Piece-wise exponential additive mixed models (PAMMs)

The piece-wise exponential additive mixed model (PAMM; Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, et al., 2018) is a semi-parametric extension of the piece-wise exponential model (PEM; M. Friedman, 1982) in the framework of generalized additive mixed models (GAMM; Wood, 2011, 2017). The PAMM is able to fit a large class of models for time-to-event analysis with the full flexibility of GAMMs. In this section, we introduce the PAMM in more detail. This introduction is based on the description of the PAMM in Bender and Scheipl (2018), Bender, Groll, and Scheipl (2018), and Bender, Scheipl, et al. (2018). For more information, we refer the interested reader to these papers.

Above, we introduced the piece-wise exponential data format, which splits the data into a number of intervals J. For each stimulus i, the response status δ_{ij} is encoded for each interval $j \in 1, \ldots, J$. Given the predictor values $\mathbf{x}_i$, the PEM defines the hazard function $\lambda(t|\mathbf{x}_i)$ at all time points t in the interval $j := (\kappa_{j-1}, \kappa_j]$ as:

$$\lambda(t|\mathbf{x}_i) = \lambda_j \exp(\mathbf{x}_i^\top \boldsymbol{\beta}), \quad \forall\, t \in (\kappa_{j-1}, \kappa_j] \tag{7}$$

where λ_j is the baseline hazard for time interval j and the vector $\boldsymbol{\beta}$ contains the regression coefficients for the predictors with values $\mathbf{x}_i$ for stimulus i.

The hazard function is fitted in a piece-wise fashion for each interval. Fitting the hazard function in a piece-wise fashion allows for flexible time-to-event models, while avoiding a number of technical estimation problems. The predictor term $\mathbf{x}_i^\top \boldsymbol{\beta}$, however, is constant. Furthermore, the functional shape of predictor effects within an interval is restricted (i.e., predictor effects are linear). The PAMM overcomes these limitations of the PEM and allows for non-linear predictor effects that vary non-linearly as a function of time. In addition, the PAMM allows for the inclusion of random effects. Given the predictors values $\mathbf{x}_i$ for stimulus i, the PAMM defines the hazard function $\lambda(t|\mathbf{x}_i)$ at all time points t in the interval $j := (\kappa_{j-1}, \kappa_j]$ as:

$$\lambda(t|\mathbf{x}_i) = \lambda_0(t_j) \exp\Big(\sum_{k=1}^{p} f_k(x_{i,k}, t_j) + b_{\ell_i} \Big), \quad \forall\, t \in (\kappa_{j-1}, \kappa_j] \tag{8}$$

where $\lambda_0(t_j)$ is the baseline hazard for time interval j, $f_k(x_{i,k}, t_j)$ are smooth functions for predictor $k \in 1, \ldots, p$ for each time point t in the interval j, and b_{ℓ_i} are random intercepts associated with group $\ell \in 1, \ldots, L$ to which stimulus i belongs.

The smooth functions $f_k(x_{i,k}, t_j)$ are estimated through the weighted sum of a set of basis functions (cf. Baayen, Vasishth, Kliegl, & Bates, 2017). Together, these basis functions allow PAMMS to model non-linear predictor effects without a pre-defined functional shape. Practically, non-linear, non-linear time-varying predictor effects can be fitted through tensor product interactions between time and the predictor (see Wood, 2011, for an introduction to tensor product interactions in the context of GAMMs). Note that although predictor effects are modeled through smooth functions, the estimates of these effects remains piece-wise constant, as is the case for the baseline hazard $\lambda_0(t_j)$.

In Equation 8, random effects are limited to the random intercepts b_{ℓ_i}. As noted by Bender and Scheipl (2018), more complex random effect structures for nested or crossed groups can be accommodated in the GAMM framework as well. In the analyses reported here, however, no random effect terms are included. We therefore refrain from a more detailed discussion of random effects in the PAMM. Similarly, we do not discuss the option to include effects of time-varying predictors (i.e., predictors with predictor values that vary as a function of time) in the PAMM, as all predictors in the current analyses are constant over time.

The formulation of the PAMM in Equation 8 is multiplicative in nature. Using $\log(ab) = \log(a) + \log(b)$, we can reformulate Equation 8 as an additive model as follows:

$$\log(\lambda(t|\mathbf{x}_i)) = \log \lambda_0(t_j) + \sum_{k=1}^{p} f_k(x_{i,k}, t_j) + b_{\ell_i}, \quad \forall\, t \in (\kappa_{j-1}, \kappa_j]. \tag{9}$$

By default, the implementation of the PAMM in the `mgcv` package works with the additive representation of the PAMM in Equation 9. The reported effects of the predictors on the hazard function are therefore on the log-scale. If so desired, however, model estimates can be converted to hazard rates on the original scale or to survival probabilities using the post-processing functions provided by the `pammtools` package (Bender & Scheipl, 2018).

Fitted models

We fitted piece-wise exponential additive mixed models (PAMMs) with a Poisson distribution (`family = poisson()`) to the representation of the response times for words and nonwords in the piece-wise exponential data format described above using the `mgcv` package for R (Wood, 2011, 2017). The objective of the PAMMs reported here is to model the response status δ (0 or 1) as a function of time and the lexical predictors *(log) frequency*, *length*, *mean bigram frequency*, *(log) OLD20*, *SND*, and *(log) OSC*. We estimated the baseline hazard through a smooth of time (i.e., the end-point of the intervals). Time-constant predictor effects were fit through smooths for the lexical predictors, whereas we allowed for time-varying predictor effects by including tensor product interactions between time and predictor in the model, as modeled through `ti()` terms (see Wood, 2017, for more details). To ensure that the results of the models remained interpretable, we limited predictor smooths (`k = 4`) and time by predictor interactions (`k = c(4,4)`) to fourth order nonlinearities. No restrictions were placed on the smooth for time. Although it is technically possible to model three-way interactions between time and two predictors in the PAMM, we did not include such three-way interactions in the reported models to retain interpretability of the results.[1]

We conclude this section with a note on dependencies between predictors. The presence of linear dependencies between predictors in a linear regression model is referred to as multicollinearity. Multicollinearity is problematic, because it can lead to estimates of predictor effects that are qualitatively and quantitatively inaccurate (L. Friedman & Wall, 2005;

[1] A post-hoc analysis, however, revealed a significant three-way interaction between time, *(log) frequency*, and *(log) OSC* for words ($\chi^2 = 38.859$, $p < 0.001$) and - to a lesser extent - for nonwords ($\chi^2 = 6.446$, $p = 0.056$). During the later stages of the response time window, the reported effects of *(log) frequency* are most prominent (words) or exclusively present (nonwords) for high values of *(log) OSC*.

Wurm & Fisicaro, 2014; Tomaschek, Hendrix, & Baayen, 2018). Linear dependencies between predictors can be described using pairwise correlations. Strong pairwise correlations between predictors can be problematic, because they increase the likelihood of suppression (i.e., a change in the sign of a predictor effect, see Wurm & Fisicaro, 2014). In the situation where $r_{YX1} > r_{YX2} > 0$, the sign of the coefficient for X_2 in a linear regression model that regresses Y on X_1 and X_2 changes when the pairwise correlation r_{X1X2} between predictors X_1 and X_2 exceeds r_{YX2}/r_{YX1} (L. Friedman & Wall, 2005).

Pairwise Pearson correlations between the predictors in the current analyses are shown in Table 4. For words, pairwise correlations between predictors are modest, with the exception of the pairwise correlations between *length* and *(log) OLD20* ($r = 0.782$). For nonwords, the situation is more complicated. In addition to a strong correlation between *length* and *(log) OLD20* ($r = 0.787$), strong pairwise correlations exist between *(log) frequency* and *length* ($r = -0.785$) and between *(log) frequency* and *(log) OLD20* ($r = -0.820$). The Google frequency of a nonword thus is strongly correlated with its length and the density of its orthographic neighborhood. More frequent nonwords are shorter and have more orthographic neighbors. The pairwise correlation between *SND* and *(log) OSC* is substantial as well ($r = 0.637$). Furthermore, medium strength correlations exist between *length* on the one hand and *SND* ($r = 0.357$) and *(log) OSC* ($r = 0.319$) on the other hand. Estimates of the collinearity across the full set of predictors confirm that potentially harmful collinearity is present in the data. The condition number κ (Belsley, Kuh, & Welsch, 1980) is high for the data set for words ($\kappa = 42.813$), as well as for the data set for nonwords ($\kappa = 50.026$).

The analogue of multicollinearity in non-linear regression models is concurvity (Buja, Hastie, & Tibshirani, 1989). Concurvity refers to the presence of non-linear dependencies

Table 4
Pairwise Pearson correlations of lexical-distributional variables for words and nonwords.

	(log) frequency	*length*	*mean bigram frequency*	*(log) OLD20*	*SND*	*(log) OSC*
words						
(log) frequency	-	-0.325	0.047	-0.280	0.020	-0.216
length	-0.325	-	0.250	0.782	0.222	0.215
mean bigram frequency	0.047	0.250	-	-0.068	0.093	0.043
(log) OLD20	-0.280	0.782	-0.068	-	0.091	-0.009
SND	0.020	0.222	0.093	0.091	-	0.285
(log) OSC	-0.216	0.215	0.043	-0.009	0.285	-
nonwords						
(log) frequency	-	-0.785	-0.017	-0.820	-0.115	-0.106
length	-0.785	-	0.298	0.787	0.357	0.319
mean bigram frequency	-0.017	0.298	-	-0.043	0.141	0.124
(log) OLD20	-0.820	0.787	-0.043	-	0.077	-0.014
SND	-0.115	0.357	0.141	0.077	-	0.637
(log) OSC	-0.106	0.319	0.124	-0.014	0.637	-

between predictors. In the context of generalized additive models, concurvity is present when a smooth or tensor product interaction can be captured by other smooths or tensor product interactions in the model (Amodio, Aria, & D'Ambrosio, 2014). As is the case for multicollinearity in linear regression models, concurvity can lead to inaccurate estimates of predictor effects. Furthermore, model estimates may be unstable in the presence of concurvity. To ensure that the smooth and tensor product interaction estimates reported here are robust, we therefore inspect the amount of concurvity in PAMMs.

The `mgcv` package provides the `concurvity()` function, which estimates the degree of concurvity in a generalized additive model (Wood, 2017). The concurvity index for a model term ranges between 0 and 1. A concurvity index of 1 indicates that a model term can entirely be captured by the other terms in the model. Table 5 presents the concurvity indices for the smooths and tensor product interactions in the PAMMs fitted to the piecewise exponential data for the lexical decision latencies for words and nonwords (i.e, the "estimate" measure provided by the `concurvity()` function of the `mgcv` package).

As can be seen in Table 5 some concurvity is present in the PAMMs fitted to the lexical decision data. Consistent with the collinearity between *length* and *(log) OLD20*, a substantial part of the main effect smooths and tensor product interactions for *length* (main effect smooth: 0.738, tensor product interaction with *time*: 0.525) and *(log) OLD20* (main effect smooth: 0.699, tensor product interaction with *time*: 0.553) in the model for words can be captured by the other terms in the model. The concurvity estimates for *length* (main effect smooth: 0.823, tensor product interaction with *time*: 0.636) and *(log) OLD20* (main effect smooth: 0.796, tensor product interaction with *time*: 0.659) were relatively high for the PAMM fit to the nonwords as well.

Table 5
Concurvity estimates for the smooths and tensor product interactions in the PAMM*s fitted to the lexical decision data for words and nonwords.*

model term	words	nonwords
parametric terms		
intercept	0.125	0.067
smooths		
time	0.254	0.182
(log) frequency	0.335	0.691
length	0.738	0.823
mean bigram frequency	0.288	0.331
OLD20	0.699	0.796
SND	0.167	0.476
(log) OSC	0.260	0.478
tensor product interactions		
time by *(log) frequency*	0.294	0.505
time by *length*	0.525	0.636
time by *mean bigram frequency*	0.257	0.270
time by *OLD20*	0.553	0.659
time by *SND*	0.120	0.385
time by *(log) OSC*	0.218	0.399

Concurvity for *(log) frequency* was present in the model fit to the nonwords as well (main effect smooth: 0.691, tensor product interaction with *time*: 0.505). The concurvity for *(log) frequency*, however, was more moderate as compared to the collinearity described above. The effect of *(log) frequency* can therefore be separated somewhat more easily from the effects of the other predictors in a PAMM as compared to a multiple linear regression model. The same holds true for the effects of the semantic predictors. The concurvity estimates for both *SND* (main effect smooth: 0.476, tensor product interaction with *time*: 0.385) and *(log) OSC* (main effect smooth: 0.478, tensor product interaction with *time*: 0.399) were modest.

While some concurvity is present in both models, the concurvity indices for the PAMMs fit to the words and the nonwords are unlikely to lead to unstable model estimates or uninterpretable effects. Nonetheless, we verified the robustness of the reported results through two post-hoc analyses. First, in response to a request by a reviewer, we fit PAMMs without *(log) OLD20* to both data sets. For words, this led to a strong reduction in the concurvity for *length* (main effect smooth: 0.246, tensor product interaction with *time*: 0.159. For nonwords, we also observed a reduction in concurvity, albeit a less dramatic one. Moderate concurvity remained present for *length* (main effect smooth: 0.708, tensor product interaction with *time*: 0.487) and *(log) frequency* (main effect smooth: 0.628, tensor product interaction with *time*: 0.452), presumably due to the strong correlation between both predictors. The effects of *(log) frequency*, *SND*, and *(log) OSC* in the PAMMs in which we omitted *(log) OLD20* as a predictor were qualitatively and quantitatively similar to the effects in the full models reported below.

Second, we applied a principal components analysis (henceforth PCA) with varimax rotation to the set of predictors for both words and nonwords. The varimax rotation provided a set of orthogonal rotated components that mapped onto individual predictors. Crucially, the predictors of interest in the nonword data set were uniquely represented by dedicated principal components. The fourth rotated component represented *SND* (*RC4*, loading: 0.934, all other loadings ≤ 0.334), whereas the second rotated component encoded *(log) OSC* (loading: 0.937, all other loadings ≤ 0.338). The frequency of a nonword was mapped onto the fifth rotated component (loading: 0.432, all other loadings ≤ 0.112). We entered the rotated components into a PAMM fit to the lexical decision data. The effects of *(log) frequency*, *SND*, and *(log) OSC* were qualitatively similar to the results reported below. The post-hoc analyses thus indicate that the effects reported here are statistically robust, despite the presence of substantial collinearity in the data sets.

Results

Overall model fit

The results for the PAMM fit to the lexical decision data for words are presented in Table 6. Table 7 presents the results for the PAMM fit to the lexical decision data for nonwords. For the parametric terms in each model, we provide the β estimates and the corresponding standard errors, z-values, and p-values. For the smooth terms the reference degrees of freedom, the estimated degrees of freedom (EDF), the χ^2 value and the p-value are provided. The smooth terms in a generalized additive model (GAM) are allotted degrees of freedom by the basis functions (cf. Baayen et al., 2017). Under the penalization performed

Table 6
Results for a PAMM *fit to the lexical decision latencies for words. Provided are β coefficients, standard errors (S.E.) and z-values for parametric terms, and estimated degrees of freedom (edf), reference degrees of freedom (ref. df) and χ^2-values for smooth terms*

parametric terms	β	S.E.	z-value	p-value
Intercept	-6.171	0.027	-228.087	< 0.001
smooth terms	edf	ref. df	χ^2-value	p-value
s(time)	8.992	9.000	5774.851	< 0.001
s(*(log) frequency*)	2.917	2.983	4276.916	< 0.001
ti(time, *(log) frequency*)	7.657	8.363	1486.745	< 0.001
s(*length*)	2.800	2.959	8.495	0.023
s(time, *length*)	5.252	6.096	151.522	< 0.001
s(*mean bigram frequency*)	2.780	2.959	138.911	< 0.001
ti(time, *mean bigram frequency*)	6.859	8.015	17.785	0.022
s(*OLD20*)	2.990	2.999	91.937	< 0.001
ti(time, *OLD20*)	6.943	7.780	49.026	< 0.001
s(*SND*)	2.402	2.758	290.530	< 0.001
ti(time, *SND*)	2.135	2.734	58.160	< 0.001
s(*(log) OSC*)	1.459	1.765	207.936	< 0.001
ti(time, *(log) OSC*)	2.802	3.249	28.523	< 0.001

by a GAM, however, not all available degrees of freedom are (necessarily) used. The number of estimated degrees of freedom is a measure of the degrees of freedom that are actually used by a smooth term, and therefore of the degree of non-linearity of an effect (Sóskuthy, 2017; Baayen et al., 2017). Below, we discuss the results for each of the parametric and non-parametric terms in the both models in a sequential fashion.

Table 7
Results for a PAMM *fit to the lexical decision latencies for nonwords. Provided are β coefficients, standard errors (S.E.) and z-values for parametric terms, and estimated degrees of freedom (edf), reference degrees of freedom (ref. df) and χ^2-values for smooth terms*

parametric terms	β	S.E.	z-value	p-value
Intercept	-5.804	0.014	-417.128	< 0.001
smooth terms	edf	ref. df	χ^2-value	p-value
s(time)	8.993	9.000	9834.507	< 0.001
s(*(log) frequency*)	2.947	2.997	2300.981	< 0.001
ti(time, *(log) frequency*)	8.104	8.753	288.363	< 0.001
s(*length*)	1.028	1.054	3340.491	< 0.001
s(time, *length*)	7.735	8.443	1286.548	< 0.001
s(*mean bigram frequency*)	2.038	2.458	40.891	< 0.001
ti(time, *mean bigram frequency*)	1.038	1.075	4.793	0.033
s(*OLD20*)	2.994	3.000	793.965	< 0.001
ti(time, *OLD20*)	7.943	8.635	607.660	< 0.001
s(*SND*)	2.411	2.753	120.505	< 0.001
ti(time, *SND*)	4.602	5.699	21.917	0.001
s(*(log) OSC*)	2.387	2.726	329.374	< 0.001
ti(time, *(log) OSC*)	8.106	8.689	70.197	< 0.001

Baseline hazard

Although all model terms contribute to the estimate of the baseline hazard in a PAMM, two model terms were specifically included for the estimation of the baseline hazard: the parametric intercept of the model, and the non-parametric smooth for time. The intercept of the model is highly significant for both words (z = -228.087, $p = < 0.001$) and nonwords (z = -417.128, $p = < 0.001$), as is the smooth for time (words: $\chi^2 = 5774.851$, $p = < 0.001$; nonwords: $\chi^2 = 9834.507$, $p = < 0.001$). The estimated (log) baseline hazard for words (left panel) and nonwords (right panel) is presented in Figure 4. The estimate of the baseline hazard is highly similar to the estimate of the baseline hazard in a PAMM without lexical predictors presented above. We therefore refrain from further discussion of the baseline hazard here.

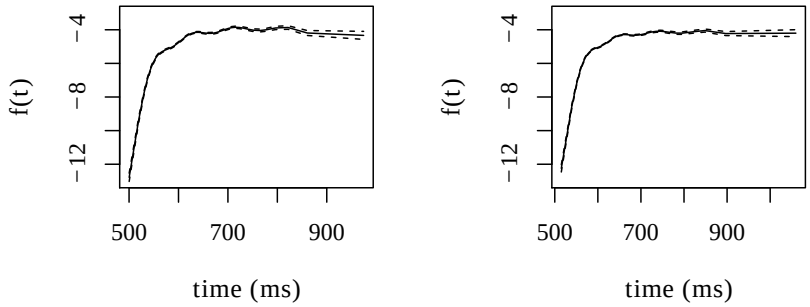

Figure 4. Estimated baseline hazard for words (left panel) and nonwords (right panel) with point-wise confidence intervals.

Frequency

For words, we found a highly significant main effect *(log) frequency* ($\chi^2 = 4276.916$, $p = < 0.001$), as well as a highly significant interaction between time and *(log) frequency* ($\chi^2 = 1486.745$, $p = < 0.001$). The partial main effect of *(log) frequency* (left panel) and the partial effect of the interaction between time and *(log) frequency* (right panel) are presented in Figure 5. Partial effect plots visualize the estimated adjustment to the (log) baseline hazard as a function of a smooth term or tensor product. Note that the graphical representation of the partial interaction is not exact. As all time-varying effects in a PAMM are piece-wise constant, all estimated time by predictor interactions are step functions in the time dimension. For the number of intervals used here (51), however, the difference is negligible (Bender & Scheipl, 2018).

The left panel of Figure 5 shows that (log) hazard rates are higher for high frequency words. Conceptually, the main effect of *(log) frequency* indicates that overall the instantaneous probability of a response is higher for high frequency words than it is for low frequency words. The main effect of *(log) frequency*, however, is modulated by the interaction that is shown in the right panel of Figure 5. Warmer colors in the right panel of Figure 5 indicate higher (log) hazard rates. The partial effect of the interaction between time and *(log) frequency* suggests that the increase in (log) hazard rates for high frequency

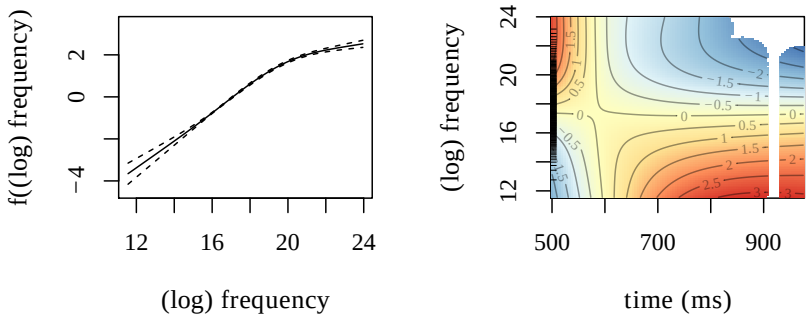

Figure 5. Partial main effect of (log) frequency (left panel) and partial interaction between time and (log) frequency (right panel) for words. Warmer colors indicate higher (log) hazard rates.

words is particularly prominent during the early stages of the response window. Later on, the facilitatory main effect of word frequency is offset by an opposite effect in the partial interaction between time and *(log) frequency*.

The interpretation of the overall effect of *(log) frequency* from the partial effect plots in Figure 5 is less than straightforward, because it requires a joint evaluation of the partial main effect and the partial interaction with time. Henceforth, we therefore visualize predictor effects on the basis of the sum of the partial main effect of the predictor and the partial interaction between the predictor and time. The left panel of Figure 6 presents the sum of the partial main effect of *(log) frequency* and the partial interaction between time and *(log) frequency*, which can be interpreted as the time-sensitive estimated adjustment to the (log) baseline hazard as a function of *(log) frequency*. The white areas in Figure 6 correspond to areas of the plot for which no data were available.

The current effect of *(log) frequency* is in line with the classical word frequency effect in lexical decision (Forster & Chambers, 1973; Murray & Forster, 2004; Balota et al., 2004; Keuleers et al., 2012). Traditional statistical techniques provide an overall estimate of the effect of frequency that lead to conclusions such as "high frequency words are responded to faster than low frequency words". The PAMM analysis presented here reveals information about the temporal development of the frequency effect. As can be seen in the left panel of Figure 6, the effect of *(log) frequency* is almost exclusively present for the early stages of the response time window. For words that are not responded to at 600 ms after stimulus onset, *(log) frequency* no longer is a strong predictor of response times, or, more technically, of the instantaneous probability of a response. The PAMM model thus provides insight into the time course of the word frequency effect that would not have been available through traditional analyses of the data.

The right panel of Figure 6 presents the effect of *(log) frequency* for nonwords. Both the main effect of *(log) frequency* ($\chi^2 = 2300.981$, $p = < 0.001$) and the interaction between time and *(log) frequency* ($\chi^2 = 288.363$, $p = < 0.001$) were highly significant. As was the case for words, the effect of nonword frequency was most prominent during the early stages of lexical processing, with a somewhat smaller effect size for nonwords than for words. The qualitative nature of the frequency effects for words and nonwords, however, was different.

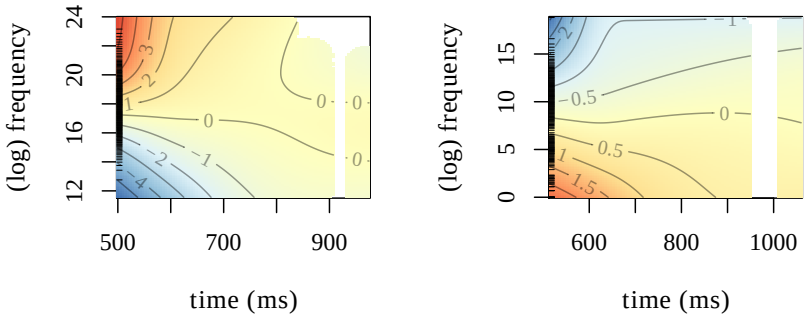

Figure 6. Effect of (log) frequency for words (left panel) and nonwords (right panel). Warmer colors indicate higher (log) hazard rates.

Whereas (log) hazard rates were higher for high frequency words, (log) hazard rates are lower for high frequency nonwords.

The opposite effects of frequency for words and nonwords might seem surprising at first sight. Opposite effects of predictors for words and nonwords, however, are commonly observed and originate from the nature of the lexical decision task. The more word-like a nonword, the harder it is to reject it as a potential word and respond "no" (Andrews, 1989, 1992, 1997; Sears, Lupker, & Hino, 1999). The lower hazard rates associated with high frequency nonwords therefore are not indicative of lower levels of activation in the mental lexicon. To the contrary, the decreased hazard rates for high frequency nonwords suggest difficulties in response selection due to higher levels of activation in the mental lexicon. The increased activation for nonwords presumably does not correspond to an increase in "local activation" of dedicated lexical representations, as no dedicated lexical representations should exist for nonwords. Instead, the current results suggest that the visual presentation of high frequency nonwords gives rise to an increased amount of total activation in the mental lexicon as compared to the visual presentation of low frequency nonwords. Previous studies have referred to the total amount of activation in the mental lexicon as "global activation" (Yap et al., 2015; Norris, 2006; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Grainger & Jacobs, 1996). Nonetheless, the interpretation of the frequency effect for nonwords is less-than-straightforward. Why are activation patterns in the mental lexicon different for high frequency nonwords as compared to low frequency nonwords? We attempt to shed further light on the effect of nonword frequency in the discussion section of this paper.

To establish the temporal onset of predictor effects, we calculated two sigma (95%) confidence intervals around the contour surfaces. The temporal onset of a predictor effect is defined as the first point in time at which 0 is not within this two sigma confidence interval for at least one value of a predictor (see Hendrix, Bolger, & Baayen, 2017). For both words and nonwords, the first point in time at which 0 was not within the two sigma confidence interval for all values of *log (frequency)* coincided with the onset of the analysis window (i.e., 500 ms after stimulus onset for words and 515 ms after stimulus onset for nonwords). This confirms that the effects of both word frequency and nonword frequency arise early. Both effects remain significant until the end of the analysis window (i.e., 975 ms after stimulus

onset for words and 1060 ms after stimulus onset for nonwords). As noted above, however, the effect size of both frequency effects decreases as a function of time. For both words and nonwords, the effect of *(log) frequency* is most prominent for the early parts of the response time distribution.

Length

We observed a significant main effect of *length* ($\chi^2 = 8.495$, $p = 0.023$) and a significant interaction between time and *length* ($\chi^2 = 151.522$, $p = < 0.001$) for words. The effect of *length* for words is presented in the left panel of Figure 7. The onset of the effect of *length* coincides with the onset of the analysis window (500 ms after stimulus onset), whereas the offset of the effect of *length* coincides with the offset of the analysis window (975 ms after stimulus onset). The semi-transparent area in the right panel of Figure 7 correspond to time points at which the effect of *length* was not significant (i.e., at which 0 was in the two sigma confidence interval for all values of *length*; from 788 ms after stimulus onset to 814 ms after stimulus onset)

The functional form of the word length effect in lexical decision is not entirely undisputed. A number of studies report inhibitory effects, with longer response times for longer words (O'Regan & Jacobs, 1992; Hudson & Bergman, 1985). Other studies, however, failed to observe word length effects (Frederiksen & Kroll, 1976; Richardson, 1976). More recently, (Baayen, 2005) documented a U-shaped effect of word length. Consistent with the U-shaped effect observed by Baayen (2005), New et al. (2006) reported longer response times for short words (3-5 letters) as well as for long words (8-13 letters) as compared to words of intermediate length (5-8 letters).

The effect of *length* observed here fits well with the U-shaped effect of word length reported by Baayen (2005) and New et al. (2006). As can be seen in the left panel of Figure 7, the effect of word length is inhibitory for the left part of the response time distribution. Early on, the probability of an instantaneous response thus is lower for longer words. Later, the effect of word length reverses, with a higher probability of an instantaneous response for long words than for short words. Recently, Hendrix (2018) observed similar bifurcated effects of word length in PAMM analyses of the lexical decision latencies in the English Lexicon Project (Balota et al., 2007), the Dutch Lexicon Project (Keuleers et al., 2010), and the French Lexicon Project (Ferrand et al., 2010).

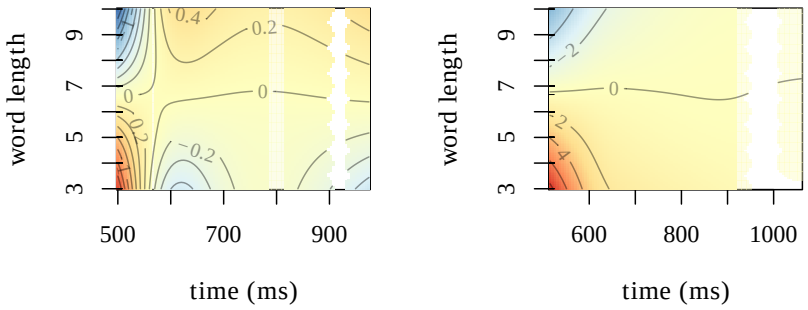

Figure 7. Effect of word length for words (left panel) and nonwords (right panel).

The early inhibitory effect of word length is likely to reflect the visual information uptake process. The visual information uptake process takes longer for longer words. As a result, early responses are less likely for long words. The later facilitatory effect of word length is theoretically more interesting. An explanation for this effect is that long words contain more (sub)lexical information that helps the reader identify the word than shorter words (Baayen, Milin, Filipović Durdević, Hendrix, & Marelli, 2011; Ramscar, Hendrix, Shaoul, Milin, & Baayen, 2014). As a result, longer words are easier to respond to than shorter words once the visual information uptake process has been completed. Words of medium length thus seem to reflect a "sweet spot" in lexical-distributional space where the visual information uptake process can be completed in a limited period of time, while the amount of information provided by the component letters is sufficient for rapid identification.

For nonwords, we observed a significant main effect of *length* ($\chi^2 = 3340.491$, $p = < 0.001$) and a significant interaction between time and *length* ($\chi^2 = 1286.548$, $p = < 0.001$) as well. The effect of length for nonwords is presented in the right panel of Figure 7. The effect of length for nonwords is most prominent during the early stages of the response time window, but remains significant until 921 ms after stimulus onset. Whereas the effect of length for words is not entirely undisputed, the effect of length for nonwords is unequivocal. Previous studies have consistently reported longer response times for longer nonwords (Yap et al., 2015; Balota et al., 2004; Whaley, 1978). The current findings are in line with these studies.

The temporal bifurcation that characterized the effect of length for words is not present for nonwords. The effect of length for words, however, suggests that the effect of length for nonwords may be a composite of an early and a later effect in the same direction. As was the case for words, we expect the visual information uptake process to take longer for longer nonwords. After the completion of the visual information uptake process, however, the increased amount of (sub)lexical information in long nonwords results in a higher activation of more real words in the mental lexicon. As a result, it is harder to verify that a long nonword is not an existing word.

Mean bigram frequency

For words, both the main effect of *mean bigram frequency* ($\chi^2 = 138.911$, $p = < 0.001$) and the interaction between time and *mean bigram frequency* ($\chi^2 = 17.785$, $p = 0.022$) were significant. The effect of *mean bigram frequency* is presented in the left panel of Figure 8 and first reaches significance at 530 ms after stimulus onset. The last point in time at which we observed a significant effect of *mean bigram frequency* is 788 ms after stimulus onset.

Recently, Baayen et al. (2011) and Milin, Feldman, Ramscar, Hendrix, and Baayen (2017) observed inhibitory effects of bigram frequency in the lexical decision task, with longer response latencies for words with more frequent bigrams. These studies interpreted the effect of mean bigram frequency as a behavioral manifestation of the principles of discrimination learning. The more frequent the bigrams in a word, the more words they appear in and the less information they provide about the identity of the current word. The reduced information provided by high frequency bigrams thus results in longer reaction times (see also Ramscar et al., 2014). The effect of *mean bigram frequency* observed here is consistent with the effect of bigram frequency reported by Baayen et al. (2011) and Milin et

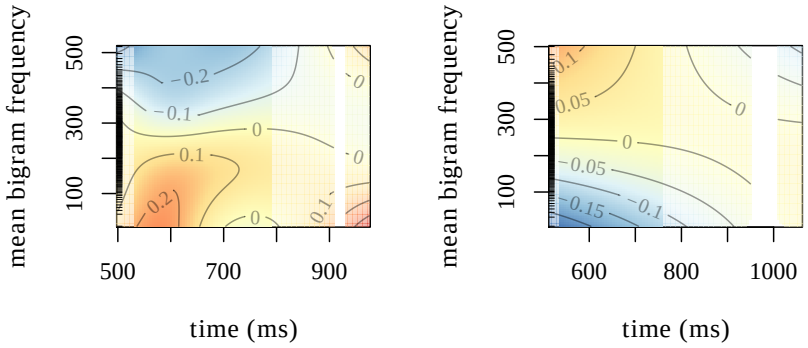

Figure 8. Effect of mean bigram frequency (in billions) for words (left panel) and nonwords (right panel).

al. (2017): the higher the average bigram frequency, the lower the instantaneous probability of a response between 530 and 788 after stimulus onset.

The PAMM analysis for nonwords revealed a significant main effect *mean bigram frequency* ($\chi^2 = 40.891$, $p = < 0.001$) and a significant interaction between time and *mean bigram frequency* ($\chi^2 = 4.793$, $p = 0.033$) as well. The effect of *mean bigram frequency* for words is presented in the right panel of Figure 8. Instantaneous hazard rates are higher for nonwords that consist of more frequency bigrams from 532 ms until 760 ms after stimulus onset.

As was the case for the effect of word frequency and the late effect of word length, the effect of average bigram frequency is opposite in nature for words and nonwords. High frequency bigrams make it harder to respond "yes" to words. By contrast, the effect of *mean bigram frequency* suggests that "no" decisions are easier for nonwords that consist of high frequency bigrams. High frequency bigrams result in low activation of a large number of words, whereas low frequency bigrams result in high activation of a small number of words (Ramscar et al., 2014). The facilitatory effect of mean bigram frequency for nonwords thus indicates that strong activation of a few words makes it harder to respond "no" to a nonword as compared to moderate activation of a large number of words.

Orthographic neighborhood density

We gauged the effect of orthographic neighborhood density through the *OLD20* measure, with lower values of *OLD20* corresponding to denser orthographic neighborhoods. For words, we found a significant main effect of *OLD20* ($\chi^2 = 91.937$, $p = < 0.001$), as well as a significant interaction between time and *OLD20* ($\chi^2 = 49.026$, $p = < 0.001$). The effect of *OLD20* is first significant at 511 ms after stimulus onset and remains significant until the end of the analysis window (975 ms after stimulus onset).

The effect of *OLD20* for words is presented in the left panel of Figure 9. Although there is a hint of an inverse U-shaped effect at the start of the analysis window, the overall nature of the effect of *OLD20* is inhibitory. The probability of an instantaneous response thus is higher for low values of *OLD20* (i.e., for words from dense orthographic neighborhoods). The current effect of *OLD20* is consistent with the facilitatory effects

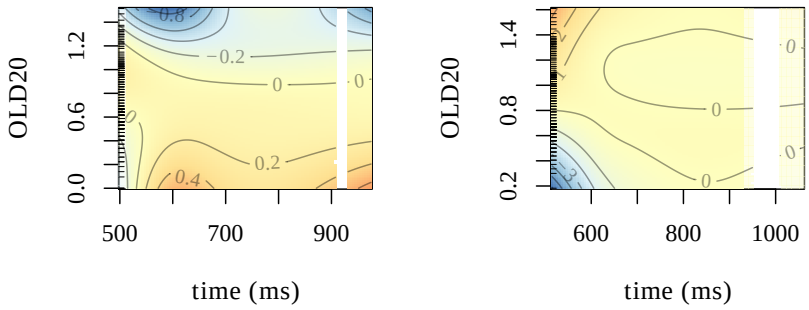

Figure 9. Effect of orthographic neighborhood density (*(log) OLD20*) for words (left panel) and nonwords (right panel).

of orthographic neighborhood density reported in lexical decision studies that adopt more traditional analysis techniques (Yarkoni et al., 2008; Keuleers et al., 2010; Andrews, 1989, 1992, 1997; Forster & Shen, 1996).

The right panel of Figure 9 presents the effect of *OLD20* for nonwords. The effect size of the effect of *OLD20* is larger for nonwords than for words. Correspondingly, both the main effect of *OLD20* ($\chi^2 = 793.965$, $p = < 0.001$) and the interaction between time and *OLD20* ($\chi^2 = 607.660$, $p = < 0.001$) are highly significant. As was the case for the effect of *OLD20* for words, the effect of *OLD20* for nonwords is more prominent during the early stages of the analysis window, reaching significance from the onset of the analysis window (515 ms after stimulus onset) until 930 ms after stimulus onset.

Consistent with the results of previous studies that investigated the effect of orthographic neighborhood density on lexical decision latencies for nonwords (Yap et al., 2015; Balota et al., 2004; Carreiras et al., 1997; Forster & Shen, 1996; Andrews, 1989; Coltheart et al., 1977), the effect of *OLD20* for nonwords is facilitatory in nature, with lower hazard rates for low values of *OLD20* (i.e., for nonwords from dense orthographic neighborhoods). The inhibitory effect of orthographic neighborhood density for nonwords provides further evidence for the idea that it is harder to correctly reject nonwords that are more word-like (Andrews, 1989, 1992, 1997; Sears et al., 1999) in the sense that they generate more global activation in the mental lexicon (Yap et al., 2015; Norris, 2006; Coltheart et al., 2001; Grainger & Jacobs, 1996).

Semantic neighborhood density

We calculated semantic neighborhood density estimates for both words and nonwords on the basis of a `fastText` model trained on Wikipedia by Bojanowski et al. (2017). The effect of the resulting predictor *SND* (i.e., semantic neighborhood density) for words is presented in the left panel of Figure 10. Both the main effect of *SND* ($\chi^2 = 290.530$, $p = < 0.001$) and the interaction of *SND* with time ($\chi^2 = 58.160$, $p = < 0.001$) were highly significant. The effect of *SND* was present from the start of the analysis window (500 ms after stimulus onset) and remained significant until 795 ms after stimulus onset. Previous studies reported that a denser semantic neighborhood allows for faster visual word

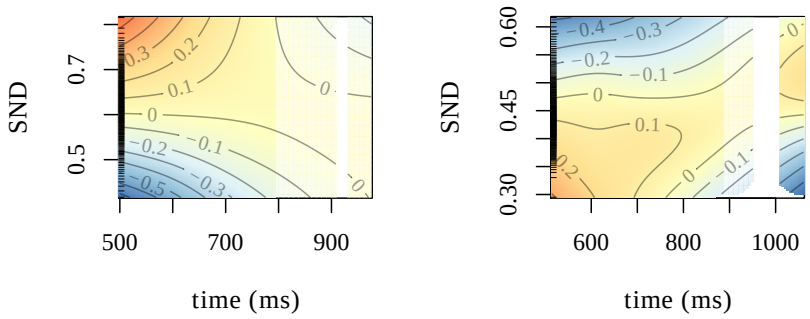

Figure 10. Effect of semantic neighborhood density (*SND*) for words (left panel) and nonwords (right panel).

recognition (Buchanan et al., 2001; Pexman & Hargreaves, 2008; Shaoul & Westbury, 2010). The current effect of *SND* is in line with these findings: the instantaneous probability of a response is higher for words from dense semantic neighborhoods (i.e., for high values of *SND*).

The current study is the first to investigate the effect of semantic neighborhood density for nonwords. As was the case for words, the main effect of *SND* was highly significant for nonwords ($\chi^2 = 120.505$, $p = < 0.001$). Furthermore, the PAMM analysis revealed a significant interaction between time and *SND* ($\chi^2 = 120.505$, $p = < 0.001$). The effect of *SND* for nonwords reached significance from the start of the analysis window (515 ms) until 888 ms after stimulus onset. Consistent with the pattern of results for word frequency, mean bigram frequency, and orthographic neighborhood density, the effects of *SND* for words and nonwords thus are in the opposite direction. Whereas the instantaneous probability of a response is higher for words from dense semantic neighborhoods, hazard rates are lower for nonwords from dense semantic neighborhoods. Again, this pattern of results fits well with the idea that higher levels of activation make it harder to respond "no" to a nonword in the lexical decision task.

At the end of the analysis window - from 1006 to 1060 ms after stimulus onset - the effect of *SND* reverses, with a higher instantaneous probability of a response for nonwords from dense semantic neighborhoods. The reversal of the effect of semantic neighborhood density, however, was not present in a principal components analysis of the nonword lexical decision data. The robustness of the late facilitatory effect of *SND*, therefore, is questionable. Hence, we refrain for further discussion of this effect here.

Orthography-to-semantics consistency

The second semantic measure under investigation, *(log) OSC*, taps into the consistency of the mapping between orthography and semantics. Higher values of *(log) OSC* indicate a greater semantic similarity between a word or nonword and its orthographic neighbors. For words, we observed a significant main effect of *(log) OSC* ($\chi^2 = 207.936$, $p = < 0.001$) and a significant interaction of *(log) OSC* with time ... ($\chi^2 = 28.523$, $p = < 0.001$). Figure 11 visualizes the effect of *(log) OSC* for words, which reaches significance

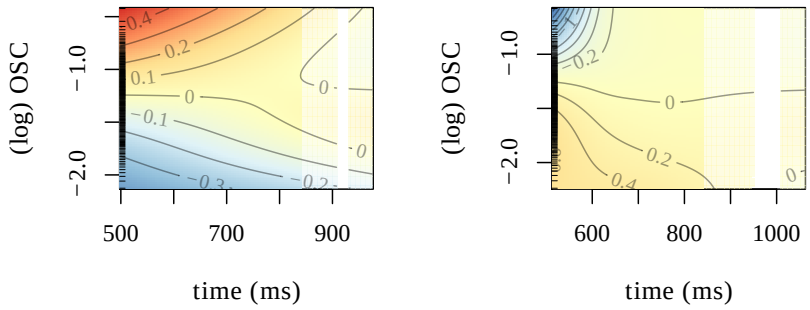

Figure 11. Effect of orthography-to-semantic consistency (*(log) OSC*) for words (left panel) and nonwords (right panel).

from the start of the analysis window (500 after stimulus onset) until 842 ms after stimulus onset. Consistent with the facilitatory effects of orthography-to-semantics consistency on lexical decision latencies reported in previous studies (Marelli & Amenta, 2018; Marelli et al., 2015; Jared, Jouravlev, & Joanisse, 2017), instantaneous hazard rates are higher for words with more consistency orthography-to-semantics mappings during this interval.

We observed a significant effect of *(log) OSC* for nonwords as well. The PAMM analysis for nonwords revealed a significant main effect of *(log) OSC* ($\chi^2 = 329.374$, $p = < 0.001$) and a significant interaction between time and *(log) OSC* ($\chi^2 = 329.374$, $p = < 0.001$). As can be seen in the right panel of Figure 11, the effect of orthography-to-semantics consistency for nonwords is inhibitory in nature, with a higher instantaneous probability of a response for nonwords with a more consistent orthography-to-semantics mapping from the start of the analysis window (515 ms after stimulus onset) until 840 ms after stimulus onset. As was the case for the effects reported above, the qualitative nature of the effect of orthography-to-semantics consistency for nonwords thus is opposite to the qualitative nature of the effect of orthography-to-semantics consistency for words.

Discussion

Nonword reading is commonly assumed to be uninformative about lexical processing. Recently, however, Yap et al. (2015) reported the results of an analysis of the lexical decision latencies for nonwords in the English Lexicon Project (ELP; Balota et al., 2007). These results shed new light on the mechanisms that drive visual word recognition. Here, we reported the results of an analysis of the lexical decision latencies for both words and nonwords in the British Lexicon Project (BLP; Keuleers et al., 2012) using a novel statistical technique from time-to-event analysis: the piece-wise exponential additive mixed model (PAMM; Bender & Scheipl, 2018; Bender, Groll, & Scheipl, 2018; Bender, Scheipl, et al., 2018). We document a novel effect of the frequency of nonwords on Google. Furthermore, the PAMM analysis revealed effects of two predictors related to semantic properties of nonwords: semantic neighborhood density and orthography-to-semantics consistency. The findings reported here provide interesting new insights into the processes that drive visual word recognition; not only for nonwords, but also for words. Below, we discuss the implica-

tions of the effects reported here for our understanding of lexical processing and for existing models of visual word recognition. First, however, we establish the advantages of using a PAMM for the analysis of lexical decision latencies through a comparison of the results of the PAMM analyses reported here with the results of more traditional multiple linear regression models.

Piece-wise exponential additive mixed models for response time data

The piece-wise exponential additive mixed model (PAMM) is an extension of the piece-wise exponential model (PEM; M. Friedman, 1982) that leverages the wealth of statistical possibilities offered by generalized additive mixed models (GAMM; Wood, 2011, 2017). The PAMM offers insight into the non-linear development of non-linear predictor effects over the response time window. The development of the (PAMM) is recent. The application of PAMMs to response time data in general, and to behavioral measures from psycholinguistic experiments, therefore, is less than widespread. The current study is part of an initial exploration of the potential of PAMMs for uncovering information about lexical processing from reaction times in the lexical decision task that is not available through traditional analysis techniques (cf Hendrix, 2018; Hendrix & Sun, 2019; Hendrix et al., 2019).

To establish the benefits of a PAMM analysis of the lexical decision latencies in the BLP as compared to a traditional multiple linear regression analysis, we fit multiple linear regression models to the reaction times for both words and nonwords. The dependent variable in both models was the average reaction time for a word or nonword in the BLP. We applied inverse transforms to the reaction times to remove a rightward skew from the reaction time distributions. The predictors were identical to the predictors in the PAMM analyses reported above: *(log) frequency*, *length*, *mean bigram frequency*, *(log) OLD20*, *SND*, and *(log) OSC*. As was the case for the PAMMs reported above, we removed predictor outliers further than 2.5 standard deviations from the predictor mean prior to analysis. Furthermore, we excluded reaction times further than 2.5 standard deviations from the reaction time mean. This resulted to the exclusion of 0.93% of the data for words and 1.42% of the data for nonwords.

The results for the linear regression model fit to the lexical decision latencies for words are presented in Table 8. All predictor effects reached significance. Consistent with the effects of the PAMM analysis, the effects of *(log) frequency* ($t = -104.925$, $p < 0.000$), *SND* ($t = -19.296$, $p < 0.000$), and *(log) OSC* ($t = -16.987$, $p < 0.000$) were facilitatory,

Table 8
Results for a linear multiple regression model fit to the lexical decision latencies for words. Provided are β coefficients, standard errors (S.E.), t-values and p-values

	β	S.E.	t-value	p-value
Intercept	-0.000	0.000	-26.133	< 0.001
(log) frequency	-0.000	0.000	-104.925	< 0.001
length	-0.000	0.000	-3.644	0.000
mean bigram frequency	0.000	0.000	12.606	< 0.001
(log) OLD20	0.000	0.000	9.135	< 0.001
SND	-0.000	0.000	-19.296	< 0.001
(log) OSC	-0.000	0.000	-16.987	< 0.001

whereas the effects of *mean bigram frequency* ($t = 12.606$, $p < 0.000$) and *(log) OLD20* ($t = 9.135$, $p < 0.000$) were inhibitory in nature. The PAMM for words revealed a temporally bifurcated effect of *length*, with an early inhibitory effect followed by a later, prolonged facilitatory effect. The later facilitatory effect dominates the early inhibitory effect in the response times, as indicated by a (relatively weak) facilitatory effect of *length* in the multiple regression model ($t = -3.644$, $p = 0.000$).

Table 9 presents the results of the linear regression model fit to the nonword data. As was the case for words, all predictor effects reached significance. Again, the sign of the predictor effects was consistent with the qualitative nature of the predictor effects in the PAMM analysis reported above. The effects of *(log) frequency* ($t = 49.721$, $p < 0.000$), *length* ($t = 58.524$, $p < 0.000$), *SND* ($t = 12.770$, $p < 0.000$), and *(log) OSC* ($t = 17.860$, $p < 0.000$) were inhibitory, whereas the effects of *mean bigram frequency* ($t = -4.708$, $p < 0.000$) and *(log) OLD20* ($t = -17.577$, $p < 0.000$) were facilitatory in nature. The qualitative nature of the effects of the predictors thus was opposite for words and for non-words, which, once more, is consistent with the results of the PAMM analyses reported above.

The multiple linear regression models fit to the reaction times for words and nonwords demonstrate that the results of the PAMM analyses are statistically robust. Unlike the multiple linear regression models, however, the PAMM analyses provide detailed information about the temporal development of predictor effects over the response time window. The PAMMs fit to the word and the nonword data, for instance, revealed that the effects of the lexical-distributional predictors on the instantaneous probability of a response were consistently more prominent during the early stages of the response time window than during the late stages of the response time window. The effect size of the word frequency effects, for instance, decreased substantially as a function of time. The same holds true for the effects of word length, mean bigram frequency, orthographic neighborhood density, semantic neighborhood density, and orthography-to-semantics consistency. The results of the PAMM analyses thus indicate that the predictors under investigation allow for relatively rich insight into the processes that influence the probability of an instantaneous response during the early stages of the response time window, but that these predictors provide little information about the instantaneous probability of a response during later stages of the response time window. The predictors included here dominate the experimental literature and the concomitant development of models of visual word recognition. The PAMM analyses reported here therefore suggest that our understanding of the processes

Table 9
Results for a linear multiple regression model fit to the lexical decision latencies for non-words. Provided are β coefficients, standard errors (S.E.), t-values and p-values

	β	S.E.	t-value	p-value
Intercept	-0.002	0.000	-139.488	< 0.001
(log) frequency	0.000	0.000	49.721	< 0.001
length	0.000	0.000	58.524	< 0.001
mean bigram frequency	-0.000	0.000	-4.708	< 0.001
(log) OLD20	-0.000	0.000	-17.577	< 0.001
SND	0.000	0.000	12.770	< 0.001
(log) OSC	0.000	0.000	17.860	< 0.001

that influence the decision making process when a participant is unable to respond to a word early on, therefore, is limited. This offers interesting opportunities for future research, in which PAMMs could help identify the factors that drive the decision making process during the later stages of the response time window.

The effect of word length for words further demonstrates the advantage of the PAMM over traditional analyses techniques. The PAMM analysis indicates that the inhibitory effect of word length is a composite of an early inhibitory effect and later facilitatory effect that is temporally more widespread (cf. Hendrix, 2018). We speculated that the early inhibitory effect of word length may arise due to the increased costs of visual information uptake for long words, whereas the later facilitatory effect may reflect the increased information provided by longer words. As noted above, the effect of word length in the multiple regression model is reduced to a facilitatory effect with a relatively small effect size. The early inhibitory effect of word length thus is masked in a traditional regression analysis of the response times. The ability of the PAMM to model predictor effects that develop in a non-linear manner over time therefore helped uncover information about the effect of word length that would not have been available through a traditional regression analysis of the data.

Here, we focused on non-linear main effects of predictors and non-linear interactions of predictors with time. Non-linear, non-linearly time-varying effects, however, are but one of the statistical opportunities offered by the PAMM. While we restricted ourselves to the interplay of a time and a single predictor here, the PAMM caters for the investigation of the temporal development of non-linear interactions between predictors as well. In addition, predictors need not be constant over time. Predictors with time-sensitive predictor values can be included in a straightforward manner in the piece-wise exponential data format used by the PAMM. Furthermore, random intercepts and more complex random effect structures for nested or crossed stimulus properties are available (cf. Wood & Scheipl, 2017). The PAMM therefore offers a rich set of statistical tools that can help researchers gain a more thorough understanding of response time data from linguistic experiments.

We end our discussion of the PAMM with a precautionary note. The PAMM helps provide insight into the temporal development of predictor effects on the instantaneous probability of a response. An early effect on the instantaneous probability of a response, however, does not translate in a one-to-one manner to an early effect on lexical processing. The effects of word frequency and semantic neighborhood density on the instantaneous probability of a response, for instance, are both significant when the first responses start to come in. This does not imply, however, that the temporal onset of effects of word frequency and semantic neighborhood density on lexical processing are identical (nor does it exclude this possibility). It is important, therefore, to carefully consider the types of conclusions that can and that cannot be drawn when interpreting the results of a PAMM.

Models of visual word recognition

The presentation of a word or nonword stimulus results in the activation of various types of lexical information in the mental lexicon, either in the form of different activation levels of symbolic lexical units or as activation patterns over subsymbolic units. A “yes” (i.e., the presented stimulus is a word) or “no” (i.e., the presented stimulus is a nonword) response in the lexical decision task is made on the basis of the activated lexical information.

Recent evidence from the neuroscience literature suggests that there is a functional and neurobiological separation between response learning and response selection (see Grindrod, Bilenko, Myers, & Blumstein, 2008; Milin et al., 2017; Baayen, Hendrix, & Ramscar, 2013; Hendrix, 2016, for a discussion of this issue in the context of the lexical decision and word naming paradigms). Whereas response learning and, in the context of the lexical decision task, the activation of lexical information when a linguistic stimulus is presented take place in the temporal and parietal lobes of the cortex, the frontal lobe of the cortex is responsible for the selection of the appropriate response (Botvinick, Cohen, & Carter, 2004; Novick, Trueswell, & Thompson-Schill, 2010; Yeung, Botvinick, & Cohen, 2004). The idea that there is a functional separation between the activation of lexical information and response selection is either explicitly or implicitly embodied in most computational models of visual word recognition.

Evidence for a functional separation between lexical activation and response selection in the context of the lexical decision task is provided by Holcomb, Grainger, and O'Rourke (2002). Holcomb et al. (2002) report the results of a neurobiological investigation of the opposite effect of orthographic neighborhood density for words and nonwords through the registration of event-related potentials (ERPs) during the lexical decision task. The response times in Holcomb et al. (2002) showed the habitual facilitatory effect of orthographic neighborhood density for words and inhibitory effect of orthographic neighborhood density for nonwords (cf. Yap et al., 2015; Balota et al., 2004; Carreiras et al., 1997; Forster & Shen, 1996; Andrews, 1989; Coltheart et al., 1977). The ERPs, however, revealed qualitative a qualitatively similar pattern of results for words and nonwords. Both words and nonwords from dense orthographic neighborhoods gave rise to larger N400s as compared to words and nonwords from sparse orthographic neighborhoods. This pattern of results suggest that lexical activation is guided by the same principles for words and nonwords, and that the opposite pattern of results for words and nonwords arises during response selection. The same core phenomenon in lexical activation can thus lead to qualitatively different effects on behavioral measures due to task-specific response selection mechanisms (see Holcomb et al., 2002, p. 939). Before we discuss the implications of the effects reported here for our understanding of lexical processing, we introduce the proposed mechanisms for lexical activation and response selection in a few of the most influential models of visual word recognition.

The multiple-read out model (henceforth MROM) proposed by Grainger and Jacobs (1996) is an interactive activation model within the more general tradition of connectionist network models (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982). Response selection in the MROM is based on both local activation (i.e., the activation of the dedicated lexical representations corresponding to individual words) and global activation (i.e., the activation in the mental lexicon as a whole). A "yes" response is made when the activation of a word reaches a threshold value. A "no" response is made when no word reaches threshold activation before the deadline. The deadline in the MROM is variable, and is proportional to and determined by the amount of global activation at the early stages of lexical processing. The MROM successfully captures a number of benchmark effects in the lexical decision literature, such as the inhibitory effect of orthographic neighborhood density for nonwords. The model has been criticized, however, for the fact that the variable deadline mechanism is included for pragmatic rather than theoretical reasons (i.e., to

explain the effect of orthographic neighborhood density, see Norris, 2006). The same criticism applies to another interactive activation model of visual word recognition that adopts the variable deadline approach for the lexical decision task: the dual-route cascaded model (DRC; Coltheart et al., 2001).

A second influential model of visual word recognition is the Bayesian Reader (Norris, 2006, 2009). The premise of the Bayesian Reader is that readers behave like optimal decision makers in the context of an input (i.e., the presentation of a stimulus) and prior probabilities of words (i.e., word frequencies). Response selection in the lexical decision task is based on the posterior probability that the input was generated by a word and the posterior probability that the input was generated by a nonword. These posterior probabilities are a function of the relative likelihood that the input was generated by a word or nonword, which depends on the prior probability of each individual word and nonword and the likelihood of the input given each individual word and nonword. Global activation thus is model-intrinsic in the Bayesian Reader. As noted by (Norris, 2009, p. 210), there is no functional separation between lexical activation and response selection: "the duration of lexical processing and the duration of decision processing are one and the same thing". The same holds true for the REM-LD model proposed by Wagenmakers et al. (2004), which is a Bayesian model specifically designed for word and nonword recognition in the lexical decision task that differs from the Bayesian reader with respect to its input representations and the way in which posterior probabilities for words and nonwords are computed.

The Naive Discriminative Reader (NDR; Baayen et al., 2011) is a two-layer symbolic network model of visual word recognition that is formulated on the principles of discrimination learning. Baayen et al. (2011) reported excellent predictive power of the NDR for (the effect of lexical predictors on) response times in the lexical decision task. Effects of orthographic neighborhood density, for instance, were captured through the activation of the target word (i.e., the word presented on the screen) as a result of competition during learning. What the NDR does not do, however, is generate actual responses. The reason for this is that the authors of the NDR subscribe to the idea of a functional separation between lexical activation and response selection (Baayen et al., 2013). To generate responses in the NDR, it is therefore necessary to adopt a separate response selection mechanism that operates on the output of the discrimination learning network.

The need for an independent response selection mechanism inspired the development of two stand-alone modules that select a response on the basis of the lexical activations generated by a model for visual word recognition. Ratcliff, Gomez, and McKoon (2004) explored the potential of the diffusion model proposed by Ratcliff (1978) in the context of lexical decision. The drift rate of the diffusion model for lexical decision depends on the output of a lexical activation model. The diffusion model assumes that lexical processing takes place during a fixed period of time prior to the onset of the diffusion process. Reaction times thus solely depend on the diffusion process. Dufau, Grainger, and Ziegler (2012) proposed a leaky competing accumulator (LCA) response module for lexical decision. The LCA model consists of two nodes : a "yes" node and a "no" node. The "yes" node and the "no" node communicate through inhibitory connections. The input to the "yes" node is a measure of lexical activation obtained from an independent model, whereas the input to the "no" node is a constant minus the same measure of lexical activation. Lexical decision latencies are generated on the basis of this input through a leaky, competitive mechanism.

The models of visual word recognition for lexical decision discussed here differ with respect to the mechanisms proposed for lexical activation and response selection, as well as with respect to the functional architecture that integrates lexical activation and response selection. The notion of word-likeness (Andrews, 1989, 1992, 1997; Sears et al., 1999), however, is an important concept across models of visual word processing in the lexical decision task in the context of nonword processing, be it as a pragmatic construct to account for behavioral data or as an emergent property of the adopted learning mechanism. The word-likeness of a nonword is gauged through the amount of global activation in the mental lexicon. The different models of visual word recognition agree that greater levels of global activation should result in longer response times for nonwords (see, however Perea et al., 2005, for an effect of base word frequency that is in the opposite direction), either as an inherent property of lexical activation or as a result of the decision making process. The more word-like a nonword, the harder it is to (correctly) respond "no" in the lexical decision task. The exact lexical-distributional properties of a nonword that contribute to its word-likeness, however, remain a topic of ongoing research.

Frequency effect for nonwords

Previous studies approximated the frequency of a nonword through base word frequency measures. Base word frequency is defined either as the frequency of the real word a nonword is based on or the frequency of a nonwords' orthographic neighbors. The qualitative nature of the effect of base word frequency in these studies, however, has been less-than-consistent. Whereas some studies reported facilitatory effects of base word frequency (Yap et al., 2015; Ziegler et al., 2001), others documented inhibitory effects (Andrews, 1996; Perea et al., 2005) or null effects (Allen et al., 1992). Here, we obtained nonword frequencies through Google searches for the $10,000$ nonwords under investigation. For the current data, the correlation between the Google nonword frequencies and a base word frequency measure similar to the one used by Yap et al. (2015) is $r = 0.464$. At least for the data used here, base word frequency thus is a fairly crude approximation of nonword frequency. Furthermore, for the current data, the effect size of a base word frequency measure based on the average frequency of a nonword's orthographic neighbors in a multiple linear regression model of the nonword lexical decision data was much smaller ($t = 2.602$, $p = 0.009$) than the effect of the Google frequency of a nonword ($t = 49.721$, $p < 0.001$) in a similar model. This is not to say that the effect of base word frequency is not theoretically interesting. For the lexical decision data for the nonwords in the BLP, however, the nonword frequency counts obtained from Google provide superior explanatory power.

How should the nonword frequency effect observed here be interpreted? Providing an answer to this question, it turns out, is less-than-trivial. One possibility is that some participants may have previously encountered a subset of the nonwords in the BLP, and that this subset of nonwords drives the nonword frequency effect. The tendency towards multimodality in the nonword frequency distribution (see Figure 1) would be in line with such an interpretation of the nonword frequency effect. Manual inspection of the nonwords in the BLP indicated that the lexical status of a minority of the nonwords under investigation could indeed be considered questionable. Participants may have experienced the nonword "doller", for instance, as a miss-spelling of the word "dollar". Although it is not technically a word, the nonword "liker" is easily interpreted as "a person who likes (something)".

Urban Dictionary, a crowdsourced online dictionary for slang words, defines the nonword "mesty" as a portmanteau of "messy" and "nasty". Nonwords that may previously have been encountered by participants, however, form a small minority of the nonwords in the BLP. The effect of nonword frequency reported here thus is unlikely to be driven by previous experience with a subset of the nonwords under investigation. Indeed, the reported effect of nonword frequency remained highly significant in a PAMM analysis that included nonwords with a Google frequency smaller to or equal to 100 only (21.80% of the nonword data; main effect of (log) frequency: $\chi^2 = 386.445$, $p < 0.001$, interaction of time with *(log) frequency*: $\chi^2 = 74.428$, $p < 0.001$). The qualitative nature of the effect of *(log) frequency* in this PAMM analysis was highly similar to the qualitative nature of the effect of nonword frequency reported above.

A second option is that the nonword frequency effect captures the effect of a latent predictor that is highly correlated with nonword frequency. Nonword frequency is highly correlated with length ($r = -0.785$) and orthographic neighborhood density (as measured through the average Levenshtein distance between a word and it's 20 closest orthographic neighbors; $r = -0.820$). The nonword frequency effect could therefore be a latent effect of visual complexity or orthographic neighborhood density. To exclude this possibility, we carried out two additional analyses. First, we investigated the role of non-linear correlations between predictors through concurvity estimates for the fitted PAMMs. The moderate concurvity that was present in the PAMM fit to the nonword lexical decision data is unlikely to have led to unstable model estimates or uninterpretable predictor effects. The concurvity estimates for *(log) frequency* (main effect smooth: 0.691, tensor product interaction with time: 0.505) indicated that a substantial part of the non-linear time-varying effect of *(log) frequency* cannot alternatively be captured by length, orthographic neighborhood density, or the other predictors under investigation. Second, we carried out a principal components analysis with varimax rotation. This principal components analysis allowed us to separate *(log) frequency* from *length* and *(log) OLD20* to a decent extent. At 0.432, the loading of *(log) frequency* on the corresponding rotated component was moderate. The loadings of *length* and *(log) OLD20* on the rotated component that corresponded to *(log) frequency* were lower than or equal to 0.122, as were the loadings of all other predictors entered into the analyses. The rotated component thus captured a decent proportion of the variance related to *(log) frequency*, while being orthogonal or near-orthogonal to *length* and *(log) OLD20*. The effect of the rotated component corresponding to *(log) frequency* in a PAMM analysis of the lexical decision data for nonwords was highly similar to the effect of nonword frequency reported above.

The possibility remains, of course, that the nonword frequency effect is a latent effect of a lexical-distributional variable that was not included in the current analyses. Primary candidates for such lexical-distributional variables are frequencies of component letter n-gram sequences. Here, we included *mean bigram frequency* as a proxy of the component letter bigram sequences. The *(log) frequency* and *mean bigram frequency* measures were nearly entirely orthogonal ($r = -0.017$). Nonetheless, it is theoretically possible that the frequencies other letter n-gram sequences strongly correlate with *(log) frequency*. We therefore calculated mean letter unigram, mean letter trigram, mean letter quadgram, and mean letter pentagram frequencies as well. The correlations of mean unigram frequency ($r = -0.069$), mean trigram frequency ($r = -0.179$), mean quadgram frequency ($r = -0.224$), and mean

pentagram frequency ($r = -0.114$) with the frequency of the nonword as a whole, however, were weak. The effect of nonword frequency thus is unlikely to be a latent effect of the frequency of component letter n-gram sequences. As noted by Harald Baayen during the review process, with the exclusion of base word frequency, word length, orthographic neighborhood density, and component letter n-gram frequency as potential confounds, "there is no obvious mediating variable that can be held responsible for the nonword frequency effect".

The effect of nonword frequency thus remains an enigma. On the one hand, the effect of nonword frequency is qualitatively similar to the effect of word frequency, albeit opposite in nature due to task demands (i.e., nonwords require a "no" response, whereas word require a "yes" response). One the other hand, however, it is unlikely that an interpretation of the word frequency effect and the nonword frequency effects along the same lines is correct. Although models of visual word recognition disagree on the exact cognitive architecture of the processes that drive the word frequency effect, the consensus is that the more often a participant encountered a word in the past, the easier it is to access that word. As noted above, the effect of nonword frequency reported here was present across the entire nonword frequency range. Participants almost certainly did not encounter low frequency nonwords before. Hence, it is unlikely that previous experience with (a subset of the) nonwords in the BLP drives the effect of nonword frequency observed here. Presumably, the mechanisms that underlie the nonword frequency effect, therefore, are distinct from the mechanisms that underlie the word frequency effect.

The PAMM analyses revealed a highly significant effect of nonword frequency with a considerable effect size. In a multiple linear regression model fit to the nonword lexical decision latencies, nonword frequency was the second strongest predictor of the lexical decision latencies for nonwords in the BLP (see above), after word length. To further demonstrate the prominence of the nonword frequency effect we conducted random forest analyses of the lexical decision latencies for the words and nonwords in the BLP. Using the *ranger* package for R (Wright & Ziegler, 2017), we fit random forests with 500 trees to the data, setting the parameter for the number of predictors that are considered in each tree to 2. The random forest for words explained 35.80% of the variance in the data, whereas the random forest for nonwords explained 30.65% of the variance in the data. Following the recommendation of Nicodemus, Malley, Strobl, and Ziegler (2010), we established the contribution of the lexical-distributional variables through unscaled permutation-based variable importances. We divided variable importances by the sum of all variable importances to obtain the relative influence of predictors in the random forest models.

The relative influences of the lexical-distributional variables for words (left panel) and nonwords (right panel) in the random forest models are presented in Figure 12. For words, *frequency* accounts for a majority of the predictive power of the random forest (relative influence: 0.651). The contributions of *length* (relative influence: 0.111), *mean bigram frequency* (relative influence: 0.024), *OLD20* (relative influence: 0.104), *SND* (relative influence: 0.055), and *OSC* (relative influence: 0.055) are more modest. For nonwords, the predictor with the greatest explanatory power is *length* (relative influence: 0.385). Consistent with the results of the multiple regression model, *frequency* is the second strongest predictor for the lexical decision latencies for nonwords in the random forest model (relative influence: 0.342). Again, the predictive power of nonword frequency is considerable greater

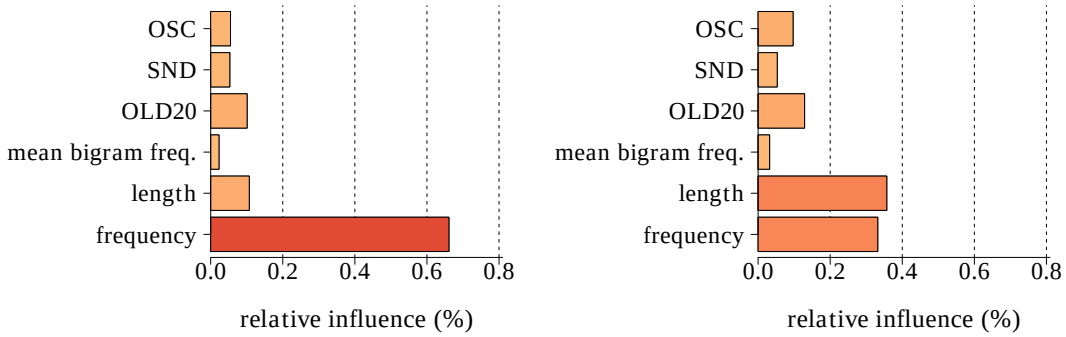

Figure 12. Relative influence of the lexical-distributional variables *frequency*, *length*, *mean bigram frequency*, *OLD20*, *SND*, and *OSC* in a random forests fit to the lexical decision data for words (left panel) and nonwords (right panel).

than that of *mean bigram frequency* (relative influence: 0.035), *OLD20* (relative influence: 0.127), *SND* (relative influence: 0.051), and *OSC* (relative influence: 0.060).

The prominence of the nonword frequency effect across the PAMM, the multiple linear regression, and the random forest analyses of the data establishes nonword frequency as an important predictor for response times to nonwords in the lexical decision task. The current results fit well with the frequency effect for nonwords in the word naming task that was recently reported by Hendrix et al. (2019). In this study, nonword frequency emerged as the strongest predictor of naming latencies in a re-analysis of the nonword naming data collected by McCann and Besner (1987), with a higher t-value for nonword frequency than for length or orthographic neighborhood density in a multiple linear regression model of the data. The frequency with which nonwords are used thus captures an essential aspect of nonword processing, even through individual participants are highly unlikely to have encountered these nonwords before. The questions posed by the effect of nonword frequency thus are not only interesting, but also highly relevant. The answers to these questions have the potential to provide important new insights into nonword processing and into the processes that underlie visual word recognition.

The nonword frequency effect poses an interesting challenge to models of visual word recognition. The multiple read-out model (MROM; Grainger & Jacobs, 1996) accounts for the effect of word frequency by assuming higher levels of local activation (i.e., activation of the lexical representation of a word) for high frequency words as compared to low frequency words. As noted above, however, participants are highly unlikely to have previous experience with most of the nonwords in the BLP. Lexical representations cannot exist for words that have not previously been encountered. Increased local activation, therefore, is a less-than-persuasive explanation for the nonword frequency effects observed here. The frequency effect in the Bayesian reader arises due to the fact that the posterior probability of a given word is a function not only of the perceptual evidence for that word, but also of the prior probability of the word (i.e., its frequency). All nonwords in a set of nonwords that a given participant has no prior experience with, however, should be equiprobable for that participant. The frequency effect for nonwords thus cannot straightforwardly be explained as a consequence of increased prior probabilities for high frequency nonwords.

In the Naive Discriminative Reader (NDR; Baayen et al., 2011) frequency effects arise as a result of stronger associations between input features (i.e., letters and letter combinations) and lexical representations. These stronger associations are a consequence of the increased experience with high frequency words. As noted above, however, lexical representations cannot exist for nonwords, nor do participants have previous experience with nonwords.

The mechanisms that are responsible for the word frequency effect in the MROM, the Bayesian Reader, and the NDR therefore do not offer a convincing explanation for the nonword frequency effect observed here. The fact that the effect of nonword frequency cannot be explained through the cognitive architectures responsible for the word frequency effect in models of visual word recognition indicates that existing interpretations of the word frequency effect should be carefully reconsidered as well. Apparently, it is possible for frequency effects to emerge in the absence of experience with individual words. While it is highly likely that previous experience with words contributes to word frequency effects, it may therefore provide an incomplete explanation of these effects. The mechanisms that drive the enigmatic effect of nonword frequency reported here may contribute to word frequency effects as well. Further research into the nature of the nonword frequency effect reported here thus is pivotal for a comprehensive understanding lexical processing not only for nonwords, but also for words.

Semantic effects for nonwords

We furthermore reported effects of semantic neighborhood density, as well as of orthography-to-semantics consistency. The semantic vectors that underlie both predictors were extracted from a `fastText` model trained on Wikipedia by Bojanowski et al. (2017). As noted above, `fastText` is an extension of the `word2vec` skip-gram model that takes subword information into account. Semantic vectors for words are defined as the sum of the semantic vector for the word itself and the semantic vectors for its component letter 3-grams to 6-grams. For nonwords, no semantic vectors for the nonword as a whole are available. Semantic vectors for nonwords, therefore, are defined as the sum of the semantic vectors for its component letter n-grams. These semantic vectors for nonwords represent the location of nonwords in the same multi-dimensional semantic space that describes the semantic properties of words.

The use of semantic vectors derived from a `fastText` model for the analysis of psycholinguistic data is novel. To establish the influence of using substring information to generate semantic vectors, we compared the current semantic measures *SND* and *(log) OSC* to identical measures calculated on the basis of semantic vectors from a standard `word2vec` skip-gram model, which was trained on Wikipedia as well (Yamada, Asai, Shindo, Takeda, & Takefuji, 2018). The correlation of the *SND* measure used here with an identical *SND* measure calculated on the basis of the semantic vectors in the `word2vec` model was 0.842. The correlation of the *(log) OSC* measure with an identical *(log) OSC* measure calculated on the basis of the semantic vectors in the `word2vec` model was 0.869. The high correlations indicate that the semantic vectors for words generated by a `fastText` model are highly similar to the semantic vectors generated by a more traditional `word2vec` model. Multiple regression models of the lexical decision data for words including *(log) frequency*, *length*, *(log) old20*, and *mean bigram frequency* as predictors, furthermore indicated that the effects of *SND* and *(log) OSC* were similar for the measures based on the semantic

vectors from the `fastText` model (*SND*: $t = -19.296$; *(log) OSC*: $t = -16.987$) and the semantic vectors from the `word2vec` model (*SND*: $t = -18.748$; *(log) OLD*: $t = -19.507$). The semantic effects for words reported here thus do not depend on the use of substring information for the semantic vectors generated by `fastText`.

For nonwords, the semantic vectors in the `fastText` model rely exclusively on the use of substring information. Bowers, Davis, and Hanley (2005) provide evidence for the relevance of letter substrings for activation patterns in the semantic system. Bowers et al. (2005) reported the results of a semantic categorization task, in which target words either contained subsets (e.g., target word "hatch", subset "hat") or were part of supersets (e.g., target word " bee", superset "beer"). Target words were categorized in both a congruent condition and an incongruent condition. In the congruent condition the correct response was identical for the target word and the subset or superset ("Does hatch refer to a human body part?"), whereas in the incongruent condition the correct response for the target word and the subset or superset were different ("Does hatch refer to a piece of clothing?"). Responses were slower in the incongruent condition for target words that contained subsets, as well as for target words that were parts of supersets. Both subsets and supersets thus were processed to the semantic level, which indicates that activation patterns in the semantic system are sensitive to orthographic information below the word level.

Semantic nonword priming studies furthermore indicated that the semantic system is activated not only through word reading, but also through nonword reading. For nonwords that were derived from real words through letter transposition (e.g. the nonword "therad" was derived from the base word "thread"), for instance, White (1986), reported shorter naming latencies for nonwords that were preceded by a semantically related prime (e.g., prime "needle", target "therad"). Semantic priming effects were observed for nonwords that were derived from real words through letter substitution as well. Rosson (1983), for instance, found semantic priming effects for prime-target pairs such as "famb" (base word: "lamb") - "sheep". Similarly, Bourassa and Besner (1998) observed semantic priming effects for prime-target pairs such as "deg" (base word: "dog") - "cat", albeit only at short prime durations. Deacon, Dynowska, Ritter, and Grose-Fifer (2004) extended the findings of these studies in an ERP experiment. Deacon et al. (2004) reported semantic priming effects that were similar to the semantic priming effects for real words for nonwords that were derived from real words through single ("tolip", base word: "tulip") or double ('contle', base word: "candle") letter substitution. These results of these studies suggest that the activation of semantic information through the visual presentation of a nonword is automatic, and inevitable.

The question that remains, then, is how to interpret the effects of *SND* and *(log) OSC* for nonwords. For words, the situation is relatively clear. The denser the semantic neighborhood of a word, the shorter the response times in visual word recognition studies (Buchanan et al., 2001; Pexman & Hargreaves, 2008; Shaoul & Westbury, 2010). Similarly, a greater consistency of the orthography-to-phonology mapping corresponds to shorter lexical decision latencies (Marelli & Amenta, 2018; Marelli et al., 2015; Jared et al., 2017). As noted by Marelli and Amenta (2018, p. 1493), there is a clear distinction between both measures and the concepts that underlie them: "OSC captures semantic information that is tightly entangled with the word orthography and has an effect on lexical access that is independent from the one associated with the sheer semantic neighborhood.". Indeed, at $r = 0.285$, the correlation between *SND* and *(log) OSC* for words is weak.

The semantic vectors for nonwords, however, exclusively depend on component letter n-grams sequences. A possibility, therefore, is that the *SND* measure is tightly tied to the orthography of nonwords and may therefore taps into orthography-to-semantics consistency, rather than into semantic neighborhood density. The correlation of $r = 0.637$ between *SND* and *(log) OSC* indicates suggests that this idea is reasonable. A principal components analysis with varimax rotation, however, allowed us to map both predictors onto orthogonal rotated components. The loading of *SND* on the corresponding component was 0.934, whereas the loading of *(log) OSC* on this component was 0.334. Similarly, the loading of *(log) OSC* on its corresponding component was 0.937, whereas the loading of *SND* on this component was 0.338. The effects of both rotated components in a PAMM fit to the nonword lexical decision data were qualitatively similar to the effects of *SND* and *(log) OSC* reported above. The results of the principal components analysis suggest that the effect of *SND* for nonwords does not seem to be a latent effect of *(log) OSC*. Furthermore, the principal components analysis indicates that the effects of *SND* and *(log) OSC* cannot be reduced to effects of word frequency, word length, mean bigram frequency, or orthographic neighborhood density.

Nonetheless, as pointed out by a reviewer, the current results do not provide conclusive evidence about the extent to which the *SND* and *(log) OSC* measures for nonwords tap into distinct concepts. The results of the principal components analysis indicate that the current *SND* and *(log) OSC* measures have distinct effects on the lexical decision latencies for nonwords in the BLP. There are at least two explanations for this, both of which may be partially or entirely responsible for the observed pattern of results. The first is that, as is the case for words, the *SND* and *(log) OSC* measures tap into distinct concepts. The second is that the distinct effects of both measures observed here are a result of the different manner in which orthographic similarity is operationalized in both measures. The semantic vectors for nonwords that underlie the *SND* measure are the sum of the semantic vectors for the component letter n-grams. Orthographic neighbors in the context of the *SND* measure thus are words that share component letter n-grams. By contrast, for the *(log) OSC* measure, orthographic neighbors were defined as the 5 words with the shortest Levenshtein distance to a nonword. The possibility remains, therefore, that the *(log) OSC* and *SND* measures used here both tap into orthography-to-semantics consistency and that the distinct effects of both measures are (at least partially) due to the manner in which these measures were calculated. Further research is necessary to provide further insight into this issue.

The effects of semantic neighborhood density and orthography-to-semantics consistency indicate that the orthographic presentation of a nonword leads to activation patterns in the semantic system, much like the orthographic presentation of a real word does (cf. Deacon et al., 2004; Cassani et al., 2019; Chuang et al., 2019). The more similar the activation patterns in the semantic system for a nonword are to the activation patterns in the semantic system for real words, the more word-like a nonword is. Consequently, it is harder to respond "no" to a non-word with a high semantic neighborhood density (Andrews, 1989, 1992, 1997; Sears et al., 1999). Similarly, nonwords that generate similar activation patterns in the semantic system as orthographically similar real words tend to lead to higher levels of activation in the semantic system. This, too, makes it harder to reject these nonwords as potential real words, and leads to longer response times in the lexical decision task. As was the case for the effect of frequency, the inhibitory effects of semantic neighborhood

density and orthography-to-phonology consistency, therefore, reflect difficulties in response selection as a consequence of activation patterns in the semantic system that resemble the activation patterns in the semantic system for real words.

Above, we noted that the notion of word-likeness is an important concept in the context of nonword processing across models of visual word recognition. Typically, it is assumed that greater levels of global activation indicate a higher degree of word-likeness. An inspection of the semantic vectors for words and nonwords used here, however, revealed that the average standard deviation of the semantic vectors for the words (0.250) and nonwords (0.244) is similar. The average sum of the non-negative loadings on the dimensions of semantic space is similar for words (30.365) and nonwords (29.963) as well. Assuming that the semantic vectors extracted from `fastText` provide an accurate estimation of the (amount of) semantic information associated with words and nonword, the nonwords in the BLP would thus be nearly as word-like as the words in the BLP. To the extent that word-likeness is a useful concept in the context of nonword processing, an interpretation of word-likeness in terms of the amount of global activation may therefore not be optimal. At least in the semantic domain, word-likeness is perhaps better thought of as the similarity of the activation patterns for a nonword to the activation patterns for similar real words.

Currently, none of the models of visual word recognition in the lexical decision task discussed above take the role of semantics into account, although it should be noted that Cassani et al. (2019) and Chuang et al. (2019) started exploring semantic effects for nonwords in the context of linear discrimination learning (Baayen, Chuang, Shafaei-Bajestan, & Blevins, 2019; Baayen, Chuang, & Blevins, 2018). The effects of the measures of semantic neighborhood density and orthography-to-semantics consistency reported here indicate the semantic information influences lexical processing in visual word recognition, as gauged through the visual lexical decision task. The fact that these effects are present not only for words, but also for nonwords indicate that these effects do not crucially depend on the presence of lexical representations. Rather, the semantic effects reported here fit well with the idea that semantic information is represented in a distributed fashion, and that this information is activated in an automatic fashion, independent of the lexical status of the linguistic input. The semantic effects reported here thus pose a challenge for the extension of existing models of visual word recognition. Furthermore, the current results highlight the importance of the development of semantic measures that can be calculated not only for words, but also for nonwords. The current semantic neighborhood density and orthography-to-semantics consistency measures calculated on the basis of `fastText` models are a promising step in this direction.

Conclusions

We reported the results of an investigation of the average lexical decision latencies for 18,547 words and 10,000 nonwords in the British Lexicon Project (BLP; Keuleers et al., 2012). Response times in the lexical decision task are often though to reflect the activation of individual lexical representations in the mental lexicon. Nonword processing, in this view, is uninteresting due to absence of lexical representations for nonwords. Previous studies, however, provide support for the idea that lexical activation arises independent of the lexical status of a presented stimulus. A number of studies, for instance, have reported effects of orthographic neighborhood density for nonwords (Yap et al., 2015; Balota et al.,

2004; Carreiras et al., 1997; Forster & Shen, 1996; Andrews, 1989; Coltheart et al., 1977). The nature of the neighborhood density effect for nonwords is opposite to the nature of the neighborhood density effect for words (Yarkoni et al., 2008; Keuleers et al., 2010; Andrews, 1989, 1992, 1997; Forster & Shen, 1996), presumably due to the opposite nature of the task for words and nonwords in the lexical decision task (i.e, a "yes" response versus a "no" response). The results reported here are consistent with the opposite effects of orthographic neighborhood density for words and nonwords.

We furthermore reported two types of novel effects for nonword reading in the lexical decision task. These effects provide further evidence for lexical activation in the mental lexicon as a result of the visual presentation of a nonword. First, the current study is the first to report a true nonword frequency effect in the lexical decision task. Previous studies approximated the frequency of a nonword through measures of the frequency of either the real word it was derived from or its orthographic neighbors. These approximations, however, are relatively crude, which may explain the inconsistent results for such measures in previous work (Yap et al., 2015; Ziegler et al., 2001; Andrews, 1996; Perea et al., 2005; Allen et al., 1992). Here, we obtained true nonword frequencies through Google searches. The analysis of the nonword data revealed a robust inhibitory effect of nonword frequency with a considerable effect size. Despite our best efforts, we were unable to reduce the effect of nonword frequency to the effects of one or more other lexical-distributional variables. We furthermore established that the effect of nonword frequency is unlikely to reflect previous experience with (a subset of) the nonwords under investigation. For now, we cannot offer a convincing explanation for the frequency effect for nonwords. Understanding the processes that drive the enigmatic nonword frequency effect is an interesting topic for further research, and promises to shed further light on the lexical mechanisms that drive visual word recognition not only for words, but also for nonwords.

Second, we observed hitherto unobserved effects of semantic measures for nonwords. The semantic measures under investigation were semantic neighborhood density and orthography-to-semantics consistency. Both measures were calculated on the basis of semantic vectors extracted from a semantic space that was generated using `fastText` (Bojanowski et al., 2017). `fastText` generates semantic vectors for words as well as for component letter n-grams. Semantic vectors for nonwords can be computed through the addition of the semantic vectors for its component letter n-grams. For the lexical decision latencies from the BLP, the analysis of the data revealed facilitatory effects of semantic neighborhood density (cf. Shaoul & Westbury, 2010; Pexman & Hargreaves, 2008; Buchanan et al., 2001) and orthography-to-semantics consistency (Marelli & Amenta, 2018; Marelli et al., 2015; Jared et al., 2017) for words, and novel, inhibitory effects of both predictors for nonwords. The semantic effects for nonwords indicate that lexical representations are not a prerequisite for the activation of information in the semantic system. The effects of frequency, semantic neighborhood density, and orthography-to-semantics consistency reported here pose interesting questions for the further development of models of visual word recognition.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), 387774888.

References

Aalen, O. O. (1980). A model for non-parametric regression analysis of counting processes. In W. Klonecki, A. Kozek, & J. Rosinski (Eds.), *Lecture notes in statistics-2: Mathematical statistics and probability theory* (pp. 1–25). New York: Springer-Verlag.

Aalen, O. O. (1989). A linear regression model for the analysis of life time. *Statistical Methods*, *8*, 907–925.

Aalen, O. O. (1993). Further results on the non-parametric linear regression model in survival analysis. *Statistical Methods*, *12*, 1569–1588.

Allen, P. A., McNeal, M., & Kvak, D. (1992). Perhaps the lexicon is coded as a function of word frequency. *Journal of Memory and Language*, *31*, 826–844.

Amodio, S., Aria, M., & D'Ambrosio, A. (2014). On concurvity in nonlinear and nonparametric regression models. *Statistica*, *74*(1), 85–98.

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

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

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language*, *35*, 775–800.

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review*, *4*(4), 439–461.

Baayen, R. H. (2005). Data mining at the intersection of psychology and linguistics. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 69–83). Hillsdale, New Jersey: Erlbaum.

Baayen, R. H., Chuang, Y. Y., & Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The Mental Lexicon*, *13*(2), 230–268.

Baayen, R. H., Chuang, Y. Y., Shafaei-Bajestan, E., & Blevins, J. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity*.

Baayen, R. H., Hendrix, P., & Ramscar, M. (2013). Sidestepping the combinatorial explosion: An explanation of n-gram frequency effects based on naive discriminative learning. *Language and Speech*, *56*, 329-347.

Baayen, R. H., Milin, P., Filipović Durdević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*, 438–482.

Baayen, R. H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows. Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, 206–234.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*, 283–316.

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*, 445–459.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics. Identifying influential data and sources of collinearity.* New York: Wiley.

Bender, A., Groll, A., & Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, *18*, 299–321.

Bender, A., & Scheipl, F. (2018). pammtools: Piece-wise exponential additive mixed modeling tools. Retrieved from `https://arxiv.org/pdf/1806.01042.pdf`

Bender, A., Scheipl, F., Hartl, W., Day, A. G., & Küchenhoff, H. (2018). Penalized estimation of complex, non-linear exposure-lag-response associations. *Biostatistics*. Retrieved from `https://doi.org/10.1093/biostatistics/kxy003`

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Science*, *8*, 539–546.

Bourassa, D. C., & Besner, D. (1998). When do nonwords activate semantics? Implications for models of visual word recognition. *Memory and Cognition*, *26*(1), 61–74.

Bowers, J. S., Davis, C. J., & Hanley, D. A. (2005). Automatic semantic activation of embedded words: Is there a "hat" in "that"? *Journal of Memory and Language*, *52*(1), 131–143.

Branders, S., Frénay, B., & Dupont, P. (2015). Survival analysis with Cox regression and random non-linear projections. In *Proceedings of the 23th European Symposium on Artificial Neural Networks* (pp. 119–124).

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.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin and Review*, *8*, 531–544.

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 453–510.

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 857–871.

Cassani, G., Chuang, Y. Y., & Baayen, R. H. (2019). On the semantics of non-words and their lexical category. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *Electronic publication ahead of print*, 1–49.

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning Memory and Cognition*, *27*(1), 225–235.

Chuang, Y. Y., Vollmer, M. L., Shafaei-Bajestan, E., Gahl, S., Hendrix, P., & Baayen, R. H. (2019). On the processing of nonwords in word naming and auditory lexical decision. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (p. 1233-1237).

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornick (Ed.), *Attention and Performance* (Vol. VI, p. 535-556). Hillsdale, New Jersey: 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.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, *34*, 187–220.

Deacon, D., Dynowska, A., Ritter, W., & Grose-Fifer, J. (2004). Repetition and semantic priming of nonwords: Implications for theories of N400 and word recognition. *Psychophysiology*, *41*(1), 60–74.

Demarqui, F. N., Loschi, R. H., & Colosimo, E. A. (2008). Estimating the grid of time-points for the piecewise exponential model. *Computational and graphical statistics*, *14*(3), 333-356.

Dufau, S., Grainger, J., & Ziegler, J. C. (2012). How to say "no" to a nonword: A leaky competing accumulator model of lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 1117–1128.

Feng, G. (2009). Time course and hazard function: A distributional analysis of fixation duration in reading. *Journal of Eye Movement Research*, *3*(2), 1–22.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., & Méot, A. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, *42*(2), 488–496.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–1955. In *Studies in linguistic analysis* (p. 1-32). Oxford: Blackwell.

Forster, K. I. (1998). The pros and cons of masked priming. *Journal of Psycholinguistic Research*, *27*, 203–233.

Forster, K. I., & Chambers, S. M. (1973). Lexical access and naming time. *Journal of Memory and Language*, *12*(6), 627–635.

Forster, K. I., & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory effects in lexical decision and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 696–713.

Frederiksen, J. R., & Kroll, J. F. (1976). Spelling and sound: Approaches to the internal lexicon. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 361–379.

Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple regression. *The American Statistician*, *59*, 127–136.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, *10*(1), 101–113.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*, 518–565.

Grindrod, C. M., Bilenko, N. Y., Myers, E. B., & Blumstein, S. E. (2008). The role of the left inferior frontal gyrus in implicit semantic competition and selection: An event-related fMRI study. *Brain Research*, *1229*, 167-178.

Hendrix, P. (2016). *Experimental explorations of a discrimination learning approach to language processing* (Unpublished doctoral dissertation). Eberhard Karl's Universität,

Tübingen.

Hendrix, P. (2018). A cross-linguistic investigation of response time distributions in lexical decision. *Poster presentation at 24th Architectures and Mechanisms for Language Processing Conference (AMLaP)*.

Hendrix, P., Bolger, P., & Baayen, R. H. (2017). Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(1), 128–149.

Hendrix, P., Ramscar, M., & Baayen, H. (2019). NDRa: A single route model of response times in the reading aloud task based on discriminative learning. *PLOS ONE*, *14*(7).

Hendrix, P., & Sun, C. C. (2019). A time-to-event analysis of auditory lexical decision data. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia* (p. 1218-1222).

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience*, *14*, 938–950.

Hudson, P. T. W., & Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language*, *24*, 46–58.

Jared, D., Jouravlev, O., & Joanisse, M. F. (2017). The effect of semantic transparency on the processing of morphologically derived words: Evidence from decision latencies and event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(3), 422–450.

Kennedy, A. (2003). *The dundee corpus.* [CD-ROM].

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=vwr` (R package version 0.3.0)

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633.

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*(174).

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.

Landauer, T., & Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Levenshtein, V. I. (1996). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, and Computers*, *28*(2), 203–208.

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (OSC) estimates for 15,017 English words. *Behavior Research Methods*, *50*(4), 1482–1495.

Marelli, M., Amenta, S., & Crepaldi, D. (2015). Semantic transparency in free stems: The

effect of Orthography–Semantics Consistency on word recognition. *Quarterly Journal of Experimental Psychology*, *68*, 1571–1583.

McCann, R., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 14–24.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review*, *88*, 375-407.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*, 227–234.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, *abs/1301.3781*. Retrieved from `http://arxiv.org/abs/1301.3781`

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119.

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PLOS one*, *12*(2), e0171935.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, *111*(3), 721-756.

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 and Review*, *13*(1), 45–52.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(110).

Nilsson, M. (2012). *Computational models of eye-movements in reading: A data-driven approach to the eye-mind link* (Doctoral dissertation). Uppsala Universitet, Uppsala, Sweden.

Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, *116*(1), 207–219.

Novick, J., Trueswell, J., & Thompson-Schill, S. (2010). Broca's area and language processing: Evidence for the cognitive control connection. *Language and Linguistics Compass*, *4*(10), 906-924.

O'Regan, J. K., & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 185–197.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Retrieved from `http://www.aclweb.org/anthology/D14-1162`

Perea, M., Rosa, E., & Gómez, C. (2005). The frequency effect for pseudowords in the lexical decision task. *Perception and Psychophysics*, *67*, 301–314.

Perera, M., & Tsokos, C. (2018). A statistical model with non-linear effects and non-proportional hazards for breast cancer survival analysis. *Advances in Breast Cancer Research*, *7*, 65–89.

Pexman, P. M., & Hargreaves, I. S. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin and Review*, *15*(1), 161–167.

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org`

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, R. H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in Cognitive Science*, *6*, 5–42.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159–182.

Reingold, E. M., Reichle, E. D., Glaholt, M. G., & Sheridan, H. (2012). Direct lexical control of eye movements in reading: Evidence from a survival analysis of fixation durations. *Cognitive Psychology*, *65*, 177–206.

Richardson, J. T. E. (1976). The effects of stimulus attributes on latency of word recognition. *British Journal of Psychology*, *67*, 315–325.

Rosson, M. B. (1983). From SOFA to LOUCH: Lexical contributions to pseudoword pronunciation. *Memory and Cognition*, *11*(2), 152–160.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*, 60–94.

Scheike, T. H., & Martinussen, T. (2006). *Dynamic regression models for survival data.* New York: Springer.

Schmidtke, D. (2016). *Semantic processing of morphologically complex words: Experimental studies in visual word recognition* (Doctoral dissertation). McMaster University, Hamilton, Canada.

Schmidtke, D., Matsuki, K., & Kuperman, V. (2017). Surviving blind decomposition: A distributional analysis of the time-course of complex word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1793–1820.

Schütze, H. (1993). Word space. In *Advances in neural information processing systems 5* (pp. 895–902). Morgan Kaufmann.

Sears, C. R., Lupker, S. J., & Hino, Y. (1999). Orthographic neighborhood effects in perceptual identification and semantic categorization tasks: A test of the multiple read-out model. *Perception and Psychophysics*, *61*(8), 1537–1554.

Shaoul, C., & Westbury, C. (2010). Exploring lexical co-occurrence space using HiDEx. *Behavior Research Methods*, *42*, 393–413.

Sheridan, H., Rayner, K., & Reingold, E. M. (2013). Unsegmented text delays word identification: Evidence from a survival analysis of fixation durations. *Visual Cognition*, *21*(1), 38–60.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. Retrieved from `https://arxiv.org/pdf/1703.05339.pdf`

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *Quarterly Journal of Experimental Psychology*, *57*(4), 745–765.

Tomaschek, F., Hendrix, P., & Baayen, R. H. (2018). Strategies for managing collinearity in multivariate linguistic data. *Journal of Phonetics*, *71*, 249-267.

Wagenmakers, E. J., Steyvers, M., Raaijmakers, J. G., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*(3), 332–367.

Whaley, C. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*, 143–154.

White, H. (1986). Semantic priming of nonwords in lexical decision. *The American Journal of Psychology*, *99*(4), 479–485.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73*(1), 3–36.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC.

Wood, S. N., & Scheipl, F. (2017). gamm4: Generalized Additive Mixed Models using 'mgcv' and 'lme4' [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=gamm4` (R package version 0.2-5)

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*, 1-17.

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analysis does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48.

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Takefuji, Y. (2018). Wikipedia2Vec: An optimized tool for learning embeddings of words and entities from Wikipedia. *arXiv preprint 1812.06280*.

Yap, M. J., Sibley, D. E., Balota, D. A., Ratcliff, R., & Rueckl, J. (2015). Responding to nonwords in the lexical decision task: Insights from the English Lexicon Project. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *41*(3), 597–613.

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

Yeung, N., Botvinick, M. M., & Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychological Review*, *111*(4), 931-959.

Ziegler, J. C., Jacobs, A. M., & Klüppel, D. (2001). Pseudohomophone effects in lexical decision: Still a challenge for current word recognition models. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 547–559.