Effects of phonological neighbourhood density and frequency in picture naming.

Solène Hameau[a, b], Britta Biedermann[a, c] Serje Robidoux [a, b], & Lyndsey Nickels[a, b]

[a] Department of Cognitive Science, Macquarie University, Sydney, Australia.

[b] Macquarie University Centre for Reading, Macquarie University, Sydney, Australia

[c] Curtin School of Allied Health, Faculty of Health Sciences, Curtin University, Perth, Australia.

Solène Hameau (corresponding author)

Department of Cognitive Science

Australian Hearing Hub, 16, University Avenue

Macquarie University, NSW, 2109

Australia

solene.hameau@mq.edu.au

Abstract

Speaking involves selecting a word among co-activated words in the lexicon. The factors determining which potentially co-activated words affect the production of spoken words remain underspecified. This research investigated the influence of words that sound similar to a target word (phonological neighbours) on the picture naming latency and accuracy of young English-speaking adults. Response time analyses showed a significant interaction between the frequency of the target and the frequency of those phonological neighbours that were higher in frequency than the target. Analysis of a published picture naming dataset gave similar results. The mechanisms underlying these results were explored using computational modelling. The critical interaction observed in the human data was successfully reproduced in analyses of the output of some versions of an interactive activation model. This model featured a relatively slow rise of activation in the phonological lexicon nodes, resulting in an increase in the effect of frequency. Overall, results show that phonological neighbourhood effects are tightly related to frequency effects.

161 words

Keywords

Phonological neighbourhood, spoken word production, frequency, computational modelling, lexical retrieval, picture naming

Funding:

This research was supported by a Cross-program Grant under the Australian Research Council Centre of Excellence in Cognition and its Disorders (CE110001021), an Australian Research Council Discovery Project grant (DP190101490), an international Macquarie

Research Excellence Scholarship (iMQRES) to SH, financial support from the Macquarie University Centre for Reading (MQCR) for SR, and an Australian Research Council Future Fellowship (FT120100102) to LN.

Declarations of Interest:

None.

Introduction

Every act of oral communication requires retrieval of a phonological word form: We need to select the word form that corresponds to the meaning we wish to convey, from a range of words, including those with a similar phonological form. One common metric to characterise form similarity between words in the lexicon is phonological neighbourhood density (PND). The phonological neighbourhood density of a given word refers to the number of words in the lexicon that only differ from that word by one phoneme, either substituted, added, or deleted (Luce, 1987). For example, under this definition, 'fat', 'kit', 'cab', 'at', 'scat' and 'cats' all count as phonological neighbours of 'cat'. This definition of neighbours is widely accepted and used in the speech production literature (e.g., Sadat et al., 2014; Vitevitch, 2002) and in programmes that allow the calculation of phonological neighbourhood density (PND, e.g., N-Watch (Davis, 2005) and Clearpond (Marian et al., 2012)). Consequently, this is the definition used in this paper[1]. Some words have many phonological neighbours (high PND, or 'dense' phonological neighbourhoods; e.g., cat, 50 neighbours; man, 51 neighbours), others have few (low PND, or 'sparse' neighbourhoods; e.g. inch, 5 neighbours; elk, 6 neighbours). Each of these neighbours has its own specific frequency value and hence a word's phonological neighbourhood can be of overall high or low frequency depending on the frequency of the neighbours (high or low phonological neighbourhood frequency (PNF)).

It is generally assumed that when a word is activated in the lexicon, the phonological neighbours of this word are also activated (e.g., Luce et al., 1990). This is true when we

[1] Other measures of neighbourhood density include the Clustering Coefficient (Watts & Strogatz, 1998), Neighbourhood spread (Andrews, 1997), or Phonological Levensthein distance (Suárez et al., 2011).

recognise spoken words, with consistent inhibitory effects of more dense phonological neighbourhoods (e.g., Luce & Pisoni, 1998). Despite less consistent effects in spoken word production (see below for review), some theories also hypothesise that phonological neighbours are activated in spoken word production and that this activation affects the lexical selection process (e.g., Chen & Mirman, 2012; Dell & Gordon, 2003). The more phonological neighbours that are active, the greater the influence is hypothesised to be. In models with interactivity between word nodes and phoneme nodes, phonological neighbours are generally assumed to be activated in spoken word production. For example, within the interactive activation account proposed by Dell and colleagues (1997), activation of phonological neighbours occurs by feedback from the phoneme level to the word level: activation flows back, not only to the target lexical item, but also to its phonological neighbours (see Figure 1). Then, if (many) neighbours are active, they in turn will further activate the target's phonemes, leading to facilitation of target production (e.g., Dell & Gordon, 2003). This model does not feature competition or inhibition within or between levels. However, even within this account, the effect of neighbours need not be facilitatory. If, for example, neighbours are strongly activated at the word level, approaching the level of activation of the target, then these neighbours might yield inhibitory effects. For example, if there is noise or damage to the system, such as weakened lexical connections following brain damage or in the case of healthy ageing, a phonological neighbour could be selected in the place of the target, therefore affecting accuracy. In addition, more time steps may be required for the target to reach a level of activation sufficiently superior to the level of activation of the phonological neighbours to be selected (see e.g., Gordon & Kurczek, 2013).

< Insert Figure 1 about here >

Chen and Mirman (2012) also modelled the influence of phonological neighbourhood density in a model where interactivity between levels is complemented with bi-directional inhibitory connections within the word level. The general principle here was that weakly active neighbours should exert facilitative effects while strongly active neighbours should be inhibitory. Chen and Mirman suggest that in spoken word production, phonological neighbours are weak neighbours and should, therefore, exert a facilitatory effect on response time, while semantic neighbours, on the other hand, are strong competitors and should therefore induce inhibitory effects.

It is, indeed, uncontroversial that producing a word to speak involves selecting the target word from a set of co-activated semantically related candidates. However, existing studies of "simple" picture naming do not find effects of the number of semantic competitors a word has in the lexicon. This can be seen through investigations of the influence of "semantic neighbourhood density" (the number of words that are close in meaning to a given target) that has not been shown to predict picture naming behaviour in unimpaired subjects (for a review, see Hameau et al., 2019). This is not to say that there is no competition between semantically related words in simple picture naming: the presence of semantic competition has been demonstrated in a range of paradigms and in a considerable number of studies (see Abdel Rahman and Melinger, 2009, for a review). However, because of these null findings, and in order not to lose the focus of the present study by considering this issue in detail, semantic neighbourhood density was not included as a predictor in the present study. [2]

[2] We also performed analyses including semantic neighbourhood density on the Australian picture naming response times and accuracy of the present study (unreported here but available upon request to the authors), and found no significant effect of this predictor, in line with previous findings.

Returning to focus on phonological neighbours, it seems clear from the literature, that current theories hypothesise that there may be a critical role of the degree of activation of both target and neighbours in the effects of neighbourhood on production. One of the most well-known causes of variation in lexical activation is word frequency. Lexical frequency can be represented either as different resting activation levels (e.g., Dell, 1988) where, by virtue of their higher resting levels of activation, higher frequency words are given a "head start" in the selection process, or as different connection weights between lexical and sublexical units (e.g., Chen & Mirman, 2012), where more frequent use results in stronger connection weights between a higher frequency word's lexical representation and its segments, compared to a lower frequency word. Both mechanisms result in faster and higher activation of phonemes leading to greater accuracy and shorter latencies in word production for higher frequency words. If the frequencies of the phonological neighbours of the target are also taken into account, one would expect that, first, the higher the frequency of phonological neighbours, the stronger their effects on target word selection; and second, the lower the frequency of a target, the stronger the effects of its phonological neighbours. Hence, maximal effects of phonological neighbours would be expected on words that are lower in frequency but have many phonological neighbours of higher frequency. However, the precise balance between overall facilitation and inhibition in these scenarios is unclear. The present study aims to shed light on these patterns through both behavioural and computational modelling experiments. We first review the current literature on phonological neighbourhood effects in spoken word production, and then return to this issue in more detail.

In spoken word production, the majority of research has focused on phonological neighbourhood density (PND), with less of a focus on effects of neighbourhood frequency. In general, there is a lower likelihood of phonological errors for words of high PND compared to low PND, in spontaneous speech (e.g., malapropisms in an English speech error corpus: Vitevitch, 1997), or in paradigms designed to induce speech errors experimentally (the SLIPs, or Spoonerisms of Laboratory-Induced Predisposition technique: Stemberger, 2004; Vitevitch, 2002). In naming to definition, high PND targets seem to elicit more correct responses and fewer tip-of-the-tongue states than low PND targets (e.g., Vitevitch & Sommers, 2003). However, in contrast to these facilitatory effects of PND in some tasks that incorporate spoken word production, there is not yet any clear consensus regarding the effects of PND on a particular task used to investigate spoken word production processes: picture naming. Findings in English picture naming differ with respect to the presence and the direction of any effect[3] (e.g., facilitation: Vitevitch, 2002; no effect: Vitevitch et al., 2004; inhibition: Newman & German, 2005), and effects seem to depend on the age of the participants (e.g., inhibitory effects on accuracy in children: Newman & German, 2002; no effects on accuracy in young adults: Vitevitch, 2002). The relevant literature is summarised in Table 1. In English speaking young adults, PND seems to exert either a facilitatory effect on latency (Vitevitch, 2002: Experiments 3, 4, and 5; see also Newman & Bernstein Ratner, 2007, for a marginally significant facilitatory effect) or no significant effect (Gordon & Kurczek, 2013; Vitevitch et al., 2004); while the effect on accuracy has also been either facilitatory (Newman & Bernstein Ratner, 2007) or non-

[3] Here, we focus on current findings in picture naming in the English language, acknowledging that these effects are not necessarily generalisable across languages, as shown for example by the mostly inhibitory effects of PND found in Spanish (Pérez, 2007; Sadat et al., 2014; Vitevitch & Stamer, 2006, but see Baus et al., 2008), in Dutch (Jescheniak & Levelt, 1994; Tabak et al., 2010), and even in words from an artificial lexicon (Frank et al., 2007).

significant (Gordon & Kurczek, 2013; Vitevitch, 2002: Experiments 3-5; Vitevitch et al., 2004: Experiment 3). A different pattern of results has been found in other age groups: in children, Bernstein Ratner and colleagues (2009) found no significant effect of PND on latencies despite facilitation on accuracy. In contrast, Arnold et al. (2005), found inhibitory effects for both children's latencies and their accuracy, and Newman and German (2002, 2005) observed a detrimental effect of high PND on accuracy. This was for three different PND measures: "standard" PND, and in the 2005 study, also for the number of phonological neighbours of higher frequency than the target, and frequency-weighted PND. Finally, in older adults, Gordon and Kurczek (2013) found inhibitory effects on latency (but not accuracy) of a measure that consisted of the residuals obtained by regressing PND on length (thereby removing the shared variance attributable to length).

<Insert Table 1 about here>

Turning to those studies that have investigated the effect of phonological neighbourhood frequency (PNF), the effects mostly seem to be facilitatory. In young adults (Newman & Bernstein Ratner, 2007) and in older adults (Vitevitch & Sommers, 2003), facilitatory effects of PNF were found on both accuracy and response latency, and on accuracy in children (Bernstein Ratner et al., 2009; Newman & German, 2002). However, these results need replication given the small number of available studies, in particular for the young adult group.

An important consideration here, is how phonological neighbourhood frequency is calculated. As noted above, it refers to the overall frequency of a word's neighbours. Some studies have used the average of the frequencies of each neighbour as a measure of neighbourhood frequency (e.g., Baus et al., 2008; Chan & Vitevitch, 2010; Vitevitch, 2002;

Vitevitch & Sommers, 2003), while others have used the summed frequency of the neighbours (e.g., Coady & Aslin, 2003; Mirman & Graziano, 2013). In the computational implementation of Levelt et al.'s (1999) theory, WEAVER++, the probability of lexical selection is determined by the Luce ratio, which refers to the activation of the target divided by the sum of the activation of the competitors and the target. Hence for this theory, what is important is the *sum* of the frequency of the phonological neighbours. In contrast, we are unaware of a theory that would predict the *average* frequency to be the relevant factor. Consequently, in the research presented here we used summed frequency as the measure of neighbourhood frequency. The use of summed PNF rather than average PNF, does result in a stronger confound with the *number* of phonological neighbours, compared to the use of average PNF (perhaps why, for example, Vitevitch and Luce (1998) refer to summed PNF as "frequency-weighted similarity neighborhood")[4]. However, it has the advantage of being less affected by the presence of neighbours that are potentially very low in frequency than a metric based on average frequency of neighbours.

While previous research has examined effects of PNF, we have argued that the prediction from current theories is that what is more likely to be important is not the "main" effect of PND or PNF, but the relative strength of activation of neighbours relative to the target. The influence of neighbours is predicted to be more influential on production of low frequency targets than on high frequency targets, and neighbours of higher frequency than the target to have stronger effects than neighbours in general. If this prediction is correct, then one

[4] Although in this paper we are reporting analyses pertaining to the summed frequency of phonological neighbours, the Supplementary Materials include a summary of supplementary analyses with average neighbourhood frequency measures, given their prevalence in the literature. The findings are mostly very similar between measures of summed and average frequency of a word's neighbours.

would expect effects of PND or PNF to vary, depending on the frequency of targets relative to the frequency of these items' phonological neighbours.

This idea of different effects of phonological neighbours depending on the relative levels of activation of a target and its phonological neighbours is not new. In auditory word recognition, the Neighbourhood Activation Model (Luce & Pisoni, 1998) has very similar predictions: Within this model, the Neighbourhood Probability Rule states that the absolute frequency of a given target word may have different effects on word recognition depending on the frequency of this target word's phonological neighbours. Luce and Pisoni predicted, for instance, that the words that would be the most difficult to recognise would be low frequency target words with neighbours that are high in frequency. Similarly, Newman and German (2002) directly targeted the frequency of the target and the frequency of its phonological neighbours in spoken word production by investigating the effect of the number of neighbours of higher frequency than the target (and found an inhibitory effect of these neighbours). However, no study has, to our knowledge, looked at the interaction between target frequency and phonological neighbourhood density or phonological neighbourhood frequency in picture naming (density or frequency of either all phonological neighbours, or of neighbours of higher frequency than the target only). This is a focus of the present study and will allow a better specification of the dynamics at play during spoken word production.

All the studies reviewed above but one (Gordon & Kurczek, 2013) used a factorial design, that is, controlled sets of stimuli with a dense/sparse neighbourhood or high frequency/low frequency neighbourhood condition. Because of the problems in precisely matching the item sets, this type of design usually leads to small numbers of items, resulting in a reduced number of trials. It can be as few as eight items (Arnold et al., 2005), and up to 72 (Newman & German, 2002), and is in contrast with, for example, Gordon and Kurczek (2013) who used 200 items in a continuous design (See Rabovsky and colleagues (2016), for a discussion regarding issues relating to the dichotomisation of continuous variables). Consequently, in the present study, we used a continuous design (i.e., without matching sets with manipulated variables) with a larger number of trials (more participants and more items) to increase power in the determination of which aspects of PND/ PNF are most critical in predicting picture naming behaviour.

Hence, in Experiment 1, we used simple picture naming as a tool to investigate the influence of several measures of PND and PNF on spoken word production in a group of Australian English speakers, using a large number of stimuli. We used linear mixed effect modelling to take into account individual variation induced by different participants and different items. In Experiment 2 we replicated our latency analysis with a published set of picture naming data in British English. Finally, in Experiment 3, we used computational modelling to explore the characteristics of the language system that can replicate the effects found across Experiment 1 and 2. Computational modelling is a powerful tool for theory building. By adjusting the parameters of a computer program that aims to simulate a certain behaviour, and comparing the outcomes of simulations that use different parameter settings, to the corresponding "real-life" behaviour, it is possible to test and refine theories. In Experiment 3, we ran a series of simulations in some versions of an interactive activation model (DRC-SEM, an extension of the Dual Route Computational (DRC) model of reading: Coltheart et al., 2001 that enables simulation of spoken word production from semantics) in order to explore the necessary features of the language production system required to simulate our behavioural effects.

Experiment 1: Picture naming in Australian English speakers

The goal of this experiment was, using a range of phonological neighbourhood measures, to determine:

a) how phonological neighbourhood measures predict picture naming behaviour (response time and accuracy) in English speaking young adults, while controlling for other variables that have proven to be influential in picture naming,
b) whether measures of phonological neighbourhood density and phonological neighbourhood frequency (including measures of neighbours of higher frequency than the target) interact with target frequency,
c) whether the strongest effects are observed with measures pertaining to neighbours that have higher frequency than the target.

Method

Participants

Forty monolingual English-speaking participants (29 females) were recruited from Macquarie University (Australia) and gave their written consent to participate in this study. All either received course credit, or were paid AU$15 for their participation. All had English as their native language and none was exposed to another language at home. They were aged between 18 and 36 years (mean 20.7, *SD* 3.64) and had normal or corrected-to-normal vision.

Stimuli

The stimuli presented for naming consisted of 386 black and white drawings taken from the International Picture Naming Project (IPNP) picture database (Székely et al., 2004). These stimuli were selected from the 520 IPNP items when they had a target name that consisted of a single word, and had name agreement of greater than 75% in English monolingual speakers as given in the IPNP database. Following the experiment, further items were excluded from the analyses, due to low (less than 50%) accuracy (24 items, 6% of total items), or because the dominant Australian English response for this target comprised two words (skipping rope, paper bag, coat hanger: 3 items, <1% of total items). This procedure resulted in a final set of 359 items for analysis. Separate analyses were performed on the 183 monosyllabic items within this analysed set, in order to verify that potential observed effects were maintained when only using monosyllabic words. This is because the computational modelling in Experiment 3 was conducted using only monosyllabic words (the only type of words in the DRC vocabulary). In addition, this allows better comparison with the reviewed studies in which targets were all monosyllabic (Arnold et al., 2005; Gordon & Kurczek, 2013; Newman & German, 2002; Vitevitch, 2002; Vitevitch et al., 2004; Vitevitch & Sommers, 2003).

Our control predictors included the properties of the words or trials that have commonly been shown to influence the speed of picture naming (e.g., Alario et al., 2004; Baayen & Milin, 2010; Perret & Bonin, 2018). These include the target attributes name agreement (consisting of average accuracy for each item), visual complexity, familiarity, age of acquisition, imageability, word length in phonemes and phonotactic probability, plus trial number. Experimental predictors included log word-form frequency[5] of the target, as well as four phonological neighbourhood measures: phonological neighbourhood density (PND) (consisting of the number of phonological neighbours of the target), higher frequency PND (corresponding to the number of phonological neighbours whose word frequency was higher than the frequency of the target word), summed phonological neighbourhood frequency (PNF) (the sum of the frequencies of all of the target word's phonological neighbours), and summed higher frequency PNF (the summed frequency of those phonological neighbours whose word frequency was higher than the frequency of the target). All measures are described in detail in Appendix A, and a list of stimuli is provided in the supplementary materials. Correlations between the different psycholinguistic variables are provided in Table 2.

<Insert Table 2 about here>

Procedure

DMDX (Forster & Forster, 2003) was used for presentation of the stimuli. Instructions were presented on the screen and explained further by the examiner: participants were asked to name each picture as quickly and accurately as possible, with a single word. Each trial started with a 200ms fixation cross, followed by a blank screen for 600ms, which was then followed by the target picture presented in the centre of the screen for 2000ms. Recording started upon stimulus presentation and continued for 2000ms after the picture disappeared. A new trial was initiated 1500ms after timeout. There were 10 practice items,

---

[5] We chose neighbour frequency counts based on the word-form rather than the lemma throughout, in line with previous studies (e.g., Newman & German, 2002; Vitevitch & Sommers, 2003), and applied this choice to target frequency as well. Another reason for this choice is that, with the one phoneme difference rule, neighbours can be the plural of the target word ("cats" is a neighbour of "cat"), therefore not corresponding to the lemma. We acknowledge that it would be equally sensible to consider lemmas and lemma frequency both for targets and phonological neighbours (which we implemented in a previous version of this manuscript, obtaining a very similar pattern of results).

followed by the 386 picture stimuli (that included the 359 items that were eventually analysed), organized in four blocks of approximately 96 items (11 minutes) each, separated by a short break. The order of stimuli was fully randomised, blocks were only created in order to allow for a few breaks in the experiment. The experiment lasted approximately an hour.

Analysis

Data were analysed in R (R Core Team, 2014) using generalised linear mixed effects modelling to assess the specific respective influence of each PND/PNF variable, allowing the effect to vary across participants and items, while removing the variance associated with control predictors.

Vocal response latencies were manually adjusted using CheckVocal (Protopapas, 2007). A response was coded as correct when it corresponded exactly to the target word, with no self-correction or dysfluency. Acceptable alternative responses (e.g., 'refrigerator' for 'fridge' or 'bathtub' for 'bath') were discarded from both latency and accuracy analyses, given that the phonological neighbourhood values would differ for these responses. Incorrect responses included visual errors, semantic errors, and no responses (or timeout), and were removed from response time (RT) analyses.

For each participant, response times that were above three standard deviations from the participant's mean (1.22% of all correct trials) were replaced by the actual value of three standard deviations above the participant's mean. This ensured that longer response times that expressed the difficulty of some items for a given participant were not removed from the data, but their replacement ensured a more normal distribution. Raw response latencies were used as dependent variable. In order to satisfy normality assumptions, a generalised linear mixed effect model was fitted, assuming an inverse Gaussian distribution (with an identity link function). This procedure was preferred to the use of a linear mixed effects model with transformed RTs, as transformation can distort the ratio scale properties of measured variables, which comes with a risk of misinterpretation (Lo & Andrews, 2015)[6]. All independent variables were standardised across the set of 359 items for each participant, to have a mean of zero and standard deviation of 1, in order to allow for clearer interpretation of the size of the effects and of interactions.

Addressing multicollinearity

Inspection of the pairwise correlations between all predictors revealed that several of the item-related predictors were correlated, especially and unsurprisingly all the PND/PNF variables (see Table 2). Correlations between explanatory variables can create multicollinearity and inflate standard errors between multicollinear variables. To avoid multicollinearity between PND/PNF measures, we chose to run separate analyses with each individual PND/PNF variable. Moreover, while some authors have residualised predictors in order to overcome multicollinearity between length and PND (e.g., Gordon & Kurczek, 2013; Sadat et al., 2014), Wurm and Fisicaro (2014) do not recommend using this procedure, as it may lead to interpretation errors. As regressions already allow us to determine the effect of one predictor independent of the variance shared with control variables, we did not residualise our predictors. Multicollinearity was monitored by looking at Variance Inflation Factors (VIFs), that are indicators of multicollinearity: Depending on the authors, VIFs above 2.5 (e.g., Allison, 2012) or 5 (e.g., Hutcheson & Sofroniou, 1999) are

[6] We thank Steve Lupker, who reviewed our initially submitted manuscript, for the suggestion of using GLM with untransformed RTs. We had initially run LME analyses with transformed RTs and the pattern of results (effects of phonological neighbourhood variables) was very similar. A description of these analyses and the associated results are available upon request to the first author of this manuscript.

a sign of potentially problematic multicollinearity. In another attempt to reduce potential model instability due to multicollinearity, it was decided to remove age of acquisition and familiarity from the set of control predictors[7]. This was motivated by the fact that, in our analyses the inclusion of these two variables led to a reversal of the effect of target log frequency most likely due to multicollinearity as age of acquisition and familiarity are strongly correlated with log frequency (around r = .5). Note that the pattern of results concerning effects of neighbourhood variables was very similar in the analyses including age of acquisition and familiarity to those reported here (details available from the authors on request). Potentially problematic patterns of multicollinearity between frequency and, for instance, age of acquisition do not seem uncommon: for example, in some spoken picture naming studies, frequency effects have been shown to lose significance when age of acquisition is taken into account (see, for example, Bonin et al., 2002). Age of acquisition and frequency are both thought to affect "the lexical system" (Alario et al., 2004) and although concept familiarity (as opposed to word familiarity) is thought to affect earlier stages of processing, in reality it seems that concept familiarity ratings are confounded with the subjective frequency of words (with correlations as strong as r = .73: Boukadi et al., 2016; see also Chedid et al., 2019). Therefore, target word frequency, age of acquisition and (concept) familiarity may be measuring similar constructs and it is therefore not essential to include them all in our analyses. Indeed, removing other variables that are conceptually similar to frequency could be argued to clarify the actual relationship between frequency and the dependent variable.

[7] The removal of age of acquisition and familiarity followed the advice of an anonymous reviewer. Results of analyses including age of acquisition and familiarity are available upon request to the authors.

Generalised linear mixed effects models were run to assess the contribution of our variables of interest, and were performed using the lme4 software package (Bates et al., 2015), p-values were obtained using lmerTest (Kuznetsova et al., 2017). We first searched for the best "base" model (prior to the inclusion of PND/PNF variables). The initial model included by-item and by-participant random intercepts in order to account for the random variation corresponding to specific words or speakers, and control variables (trial order, visual complexity, imageability, length in phonemes, phonotactic probability and name agreement) as fixed factors. Next, to define the "base model", we used model comparison to remove those fixed factors that did not significantly improve the model fit. Once this base model was hence defined, the relevant measures of phonological neighbourhood (PND, PNF, higher frequency PND, or summed higher frequency PNF), as well as target log frequency, were added as main effects, in separate models, together with an interaction between the phonological neighbourhood variable and target log frequency. Given that the different measures of PND/PNF are not compared within the same model, we were able to assess each measure's predictive value compared to the other measures by inspection of these separate models' respective Akaike Information Criterion (AIC). The lower the AIC, the better the fit for that particular model and hence the better the predictive value of the PND/PNF measure in that model.

Results

A summary of all the results of every model run in the study is available in the supplementary materials.

Response time analyses

Full set of 359 items

First, average response time was significantly negatively correlated to average accuracy for each item (r = -.665, p < .001). This shows that participants did not sacrifice response time at the expense of accuracy nor vice-versa. The inclusion of imageability as a predictor along with the other control predictors did not improve the fit of a model incorporating the 265 items that had imageability values ($X^2(1)$ = 0.82, p = .366). Imageability was then removed from the predictors, leading to the further analyses being performed on the full set of 359 items. Consistent with previous literature on the predictors of picture naming latencies, words with higher name agreement predicted faster reaction times ($X^2(1)$ = 93.58, p <.001). The other predictors did not significantly improve the model (visual complexity: $X^2(1)$ = 2.54, p = .111); number of phonemes: $X^2(1)$ = 1.14, p = .286; phonotactic probability: $X^2(1)$ = 0.01, p = .905; trial number: $X^2(1)$ = 1.09, p = .296). Hence, the final "base" model included name agreement as a control fixed factor. In order to account for the random variation induced by specific participants and items, we included participant and item as random factors (with random intercepts).

To determine how measures of PND/PNF affected performance, we examined each independently when introduced into the base model as a main effect, together with an interaction term with target log frequency (see Table 3). The model including higher frequency PND and its interaction with target log frequency did not converge, hence the model without the interaction term is reported here.

<Insert Table 3 about here>

No PND or PNF measure showed any effect on response time, with the exception of a significant interaction between summed higher frequency PNF and target log frequency. This interaction is illustrated in Figure 2. For words of lower log frequency, the resulting effect of high summed higher frequency PNF tended to be inhibitory, while the effect was facilitatory for targets of higher log frequency. The main effect of frequency did not reach significance in any of the analyses.

<Insert Figure 2 about here>

Subset of 183 monosyllabic items

Response time analyses for the 183 monosyllabic items included the same control predictors as in the analyses of the full set of 359 items, and again all independent variables were standardised. When PND/PNF variables were added separately in the model as main effects, a significant interaction between target log frequency and summed higher frequency PNF was observed, similar to the analysis of the full set (see Table 3). No other significant effects of PND/PNF were observed, and similar to the analyses on the full set, the effect of target log frequency did not reach significance on any of the analyses.

Accuracy analyses

Accuracy analyses were performed using generalised linear mixed effect models, using the lme4 software package (Bates et al., 2015). Accuracy was coded as error = 0, correct = 1, and overall, accuracy was 87.35%. For accuracy analyses, we considered the same set of fixed predictors as for RT analyses, and also included by-item and by-participant random intercepts. Once a set of control variables was defined, the relevant measures of phonological neighbourhood density or phonological neighbourhood frequency were added separately as main effects in distinct models, together with an interaction with target log frequency. All independent variables were standardised, in the same fashion as in response time analyses.

Full set of 359 items

The inclusion of imageability in a model run on the 265 items with imageability values significantly improved the model ($X^2(1) = 6.07$, $p = .014$). Hence, imageability was kept as a control variable, leading to analyses being performed on 265 items instead of 359. Amongst the remaining control variables, only length in phonemes made a significant contribution to the model's fit ($X^2(1) = 6.54$, $p =.011$): longer words were less likely to be named correctly. The other predictors were not found to improve the model significantly (trial order: $X^2(1) = 1.40$, $p = .237$; visual complexity: $X^2(1) = 0.33$, $p = .565$; phonotactic probability: $X^2(1) = 1.06$, $p = .302$). Consequently, the accuracy "base" model included imageability and length in phonemes as fixed effects, and by-item and by-participant random intercepts.

Next, each PND/PNF measure was added individually to that base model as a main effect, together with an interaction with target log frequency. No effect of any PND/PNF measure or interaction reached significance (see Table 3). Target log frequency only reached significance in the higher frequency PND model (ß = 1.36, $z(0.15) = 2.07$, $p = .039$): more frequent words were more likely to be correctly named. Note that models run on the full set of 359 items without imageability (and/or without length in phonemes) led to a very similar pattern of results.

Subset of 183 monosyllabic items

For monosyllabic items, the base model had the same fixed predictors as in the analysis of the larger set of items: imageability and length in phonemes, and the same random factors (by-item and by-participant random intercepts). This led to analyses being performed on the 169 items with imageability values only, instead of the full subset of 183 monosyllabic items. Again, all independent variables were standardised. When standardised PND/PNF predictors were added individually to the base model as main effects along with target log frequency, we found no effect of any neighbourhood measure, and no significant interaction between these measures and target log frequency. Full model outputs are reported in the supplementary materials, but in Table 3, the results of the models are reported. In this set of analyses, the effect of target log frequency was significant (more frequent words were more likely to be named accurately). Note that accuracy analyses run on the whole set of 183 monosyllabic items and without imageability (and/or without length in phonemes) led to a very similar pattern of results.

Discussion

In this experiment, we investigated the influence of several different measures of phonological neighbourhood density and phonological neighbourhood frequency on picture naming while taking into account the influence of some other common predictors of picture naming behaviour, and accounting for the variance induced both by specific participants and specific items. More specifically, we were interested in how neighbourhood effects varied depending on the frequency of the target and the frequency of its neighbours, hypothesising strongest effects of neighbours on lower frequency targets, and strongest effects for phonological neighbours of higher frequency than the target word compared to phonological neighbours as a whole. The measures relating to the full number of phonological neighbours of a given word and their frequency (PND and PNF) did not have significant effects in any analysis[8]. There were also no effects of the number of neighbours of higher frequency than the target (higher frequency PND). However, critically, there was an interaction between target log frequency and the summed frequency of those neighbours higher in frequency than the target (summed higher frequency PNF) in

[8] Note that a previous version of this manuscript (available upon request to the authors) included target length (number of phonemes) as a predictor in all analyses (including latency analyses), even if it did not successfully predict response time and accuracy. Since length is correlated with PND/PNF measures (especially PND), it could be that its inclusion changes the pattern of significance for phonological neighbourhood measures. However, the analyses including length led to the same patterns of significance (no effects of PND and PNF, some effects of higher frequency neighbourhood measures as an interaction of these variables with target log frequency).

response time analyses, showing that indeed, neighbours that are higher in frequency than the target have strongest effects, and that the effects of these neighbours depended on the frequency of the target. Inspection of the AIC of each individual model showed that the fit was best for those models that included measures of summed higher frequency PNF, followed by those with higher frequency PND, then (for all sets of analyses except accuracy analyses on the full set of items) PNF, finally PND. There was no effect of any PND/PNF measure on accuracy.

The lack of a main effect of PND on *latency* is consistent with Vitevitch et al. (2004) and Gordon & Kurczek (2013), although the latter study used a residualised measure of PND, but inconsistent with Vitevitch's often cited earlier study (Vitevitch, 2002: Experiments 3, 4, and 5). The absence of an effect of PND on *accuracy* is consistent with several studies (Gordon & Kurczek, 2013, Vitevitch, 2002: Experiment 4; Vitevitch et al., 2004: Experiment 3) but inconsistent with others that found facilitation from high PND on accuracy (Newman & Bernstein Ratner, 2007; Vitevitch, 2002: Experiment 3). The absence of a main effect of PNF (of neighbours of all frequencies) for this type of task and participant sample is inconsistent with the facilitatory effect found by Newman and Bernstein Ratner (2007) for both accuracy and response time.

The two studies showing different results as the present study (Newman & Bernstein Ratner, 2007; Vitevitch, 2002) used factorial designs with relatively small sets of stimuli (44 to 48 items in total; versus the 359 in our study), that were typically monosyllabic words matched for at least word frequency, and fewer participants (24 to 34; versus 40 participants in our study) (see Table 1, earlier). In addition to the fact that our study is consequently better powered, we believe that our design is more likely to enable detection of “true” effects of the experimental variables. First, with the advent of mixed modelling techniques we were able to take into account the variability induced by both specific participants and specific items, as opposed to comparing average performance across participants and items (as is the case in Newman & Bernstein Ratner, 2007 and in Vitevitch, 2002). Such averaging does not take advantage of all the data and important individual differences between participants may be overlooked. Second, our study attempted to take into account a greater number of predictors of picture naming compared to earlier studies, and particularly variables that are thought to affect earlier processing stages in picture naming: specifically, visual complexity, imageability and name agreement. In comparison, none of these studies included either of these measures (note, though, that visual complexity and imageability did not consistently predict performance in the present study). The lack of control of potentially confounding variables can produce spurious results, therefore increasing the likelihood of a Type 1 error (i.e., finding effects of PND measures that would not be observed if these variables were controlled for). Third, in these two studies, PND/PNF measures were dichotomised so that performance on a set of low and a set of high neighbourhood density/frequency could be compared. Dichotomising a continuous variable results in a loss of information, lower measurement precision, and usually a considerable loss of power in subsequent analyses (see e.g., MacCallum et al., 2002), hence keeping the experimental variable as a continuous measure in a large set of items with a range of values of this experimental variable is more informative and precise, and increases power. In addition, our item set is more diverse than Vitevitch's (2002) item set. Vitevitch (2002) used exclusively 3-phoneme-long CVC words, and Newman and Bernstein Ratner used monosyllabic words only in their PND experiment, and one- to two-syllables long in their PNF experiment. Our item set contains words of different lengths and CV structures, hence more representative of the diversity of words that are actually used to speak. Furthermore, the dichotomising procedure requires the matching of other variables, which can result in retaining unusual materials, again perhaps less representative of everyday speech (e.g., Cutler, 1981, cited in Ellis et al., 1996).

Vitevitch (2002) does not provide the list of items used in his experiments, but we know that these items consisted exclusively of short words (three phoneme-long). It is a well-known fact that shorter words tend to be more frequent than longer words (e.g., Zipf, 1935). In our analyses, we found different effects of a phonological neighbourhood measure depending on the frequency of the target (as seen in the interaction between target log frequency and the frequency of neighbours of higher frequency than the target). More specifically, we found inhibitory effects on lower log frequency targets and facilitatory effects on higher log frequency targets. It is hence possible that, depending on their overall frequency, datasets generate either inhibitory or facilitatory effects of (higher frequency) phonological neighbours. If Vitevitch's (2002) item sets are of high frequency (being short words), they might be more likely to generate facilitatory effects of phonological neighbours, which is what was found (note that we only found effects of neighbours of higher frequency than the target, not of all neighbours like Vitevitch).

On the other hand, Newman & Bernstein Ratner (2007) provide a list of the stimuli used in their experiments. Two item sets are relevant here: one that was used to investigate effects of PND, the other to investigate effects of PNF. Of their PND set (44 items), our item set included 29 items (14 low and 15 high PND items, for a total of 1,160 trials (1,083 correct trials)). Namely, our items do not include the verbs that were part of Newman and Bernstein-Ratner's stimuli. The authors actually found no effect of PND on response time when taking out these verb stimuli. We ran our response time analyses on these 29 items and also found no effect of PND ($X^2(1) = 1.52$, $p = .218$), suggesting that the effect was at least partly driven by these verb stimuli. In addition to no PND effect, we found a significant interaction between log frequency of the target and the summed frequency of neighbours of higher frequency than the target ($X^2(1) = 5.62$, $p = .018$), therefore maintaining the effects we found in our full set. Of Newman and Bernstein-Ratner's PNF set of 44 stimuli, our item set included 33 items (20 low and 13 high PNF items). While there was a numerical advantage for response time on high PNF items in this overlapping set of items (806ms for high PNF vs 830ms on average for low PNF items), our analyses did not show an effect of PNF ($X^2(1) = 0.17$, $p = .682$). We did find an interaction between target log frequency and the summed frequency of neighbours of higher frequency than the target ($X^2(1) = 4.37$, $p = .037$), again maintaining the effects we found on the full set.

Our most notable finding was an interaction between summed higher frequency PNF and the log frequency of the target word in response time analyses. The higher the frequency of those neighbours of higher frequency than the target, the slower the RTs for the lowest frequency targets, but also the faster the naming for targets of high log frequency. A parallel can be drawn between these results and the predictions of the Neighbourhood Activation Model in spoken word recognition where neighbours of higher frequency than the target have an inhibitory effect on targets of low frequency (e.g., Luce & Pisoni, 1998). Since these findings are novel, and the effects are relatively small, it is important that we investigate their replicability. Consequently, in Experiment 2, we used the same predictors and analysis procedure in order to determine whether we could replicate these findings using another picture naming dataset (Johnston et al., 2010) and another speaker group – British English speakers (rather than Australian English).

Experiment 2: Picture naming with British English speakers

Method

This experiment uses the picture naming data from Johnston et al. (2010).

Participants

Johnston et al. (2010) report response times from 25 native English speakers, 21 female (so 84%) aged 18 to 27 years (mean age 20.08 years), who had lived continuously in the United Kingdom. This sample appears similar to the Australian sample in Study 1 with respect to age, but with a higher proportion of female participants.

Stimuli

Participants named 539 black-and-white line drawings, mostly selected from Szekely et al. (2004). In our analysis, the item set was reduced to 412 items that corresponded to a single word according to the available name agreement, had a frequency count in CELEX, whose phonotactic probability could be calculated with Phonological Corpus Tools, and had accuracy greater than 50%. Within this final set of words, 315 (about 76%) overlapped with the stimuli that were analysed in Experiment 1 with Australian speakers, but all 412 words were analysed. A summary of the properties of all words in this set, along with those of words used in Experiment 1, is available in Appendix A (and the list of items for all the experiments is provided in Appendix B).

Procedure

The procedure was similar to Experiment 1 procedure (see Johnston et al., 2010, for detail): the pictures were shown in a random order to the participants for naming (after a 500ms blank screen followed by a 500ms fixation cross, the picture was displayed for a maximum of two seconds), without any pre-familiarisation with the pictures or their names.

Analysis

Responses were analysed in a similar way to Experiment 1. Data in Johnston et al. (2010) is available at the trial level (response time for each correct response, for each of the 25 participants). Johnston et al. considered a response correct if it corresponded to the dominant name determined by earlier ratings. Overall, accuracy was 87.17%. At the trial level, Johnston et al.'s dataset only provides response time for correct trials. However, according to the authors, 4.7% of the whole data (i.e., 4.7% of the total of 539 pictures in their set) had no response time, due not to an incorrect response, but to instances in which the voice key was triggered by an environmental sound, or failed to trigger, when a dysfluent response was given, or when the participant took longer than two seconds to begin responding. It is probable therefore that there are some trials that would normally have been excluded from accuracy analyses because they did not qualify as a failure to retrieve the correct target word, but rather as technical glitches. Consequently, while we did carry out accuracy analyses, these are not reported in the main text but can be found in the supplementary materials. In addition to the pre-processing described by the authors (exclusion of instances in which the voice key was mis-triggered or in which the participant took more than 2 seconds to respond), we removed a further seven data points that were below three standard deviations below the participant's mean (0.08% of the data): these were all below 100ms. In addition, 157 data points (1.75%) that were over three standard deviations above the mean were replaced by the actual value of three standard deviations above the participant's mean.

As for Experiment 1, generalised linear mixed effects models were used to assess the contribution of our variables of interest on response time and accuracy, using the lme4 software package (Bates et al., 2015) and p-values were obtained using lmerTest (Kuznetsova et al., 2017). Similar to Experiment 1, analyses were performed on the full set of 412 items but also separately on the 210 monosyllabic words within this set. We included the same predictors that were retained in the response time analyses of the Australian data for better comparability across analyses. Our base model hence included by-item and by-participant random intercepts, and in addition, name agreement was included but needed to be specific to the pictures used in Johnston et al. (2010) and was therefore based on average accuracy for each item. Next, we added each PND/PNF variable separately to that base model as a main effect, along with an interaction with target log frequency. All independent variables were standardised as in Experiment 1.

Results

Response time analysis

Full set of 412 items

First, as expected, higher name agreement predicted faster responses (in the base model: ß = -91.73, $z(6.43) = -14.28$, $p < .001$). Next, when adding the PND/PNF variables and target log frequency, as in Experiment 1, a significant interaction between higher frequency PNF and the target's log frequency was observed (Table 4, Model 4). This interaction is illustrated in Figure 3. The summed frequency of neighbours of higher frequency than the target had an inhibitory effect on response time for lower log frequency targets, but for targets of higher frequency, it had a facilitatory effect. The interaction between target log frequency and higher frequency PND was also significant (See Table 4, Model 2) a finding that was not observed in the Australian dataset analyses. Finally, there was a significant facilitatory effect of target log frequency in all response time models performed on the full set of 412 items.

<Insert Table 4 about here>

<Insert Figure 3 about here>

Subset of 210 monosyllabic items

Similar to the analyses on the full set, in the subset of monosyllabic items, name agreement had a facilitatory effect on latencies (in the base model: ß = -91.74, $z(8.94) = -10.26$, $p < .001$). When PND/PNF variables were added separately to the model as main effects along with the interaction with target log frequency, a significant interaction between target log frequency and summed higher frequency PNF was observed, in the same direction as in the whole set and consistent with Experiment 1 findings. In addition, an interaction between target log frequency and higher frequency PND was observed, in the same direction (see Table 4). Finally, there was a significant facilitatory effect of target log frequency in all response time models performed on the set of 210 monosyllabic items.

Discussion

Experiment 2 was a replication using a published picture naming dataset of the methods of the Experiment 1 response time analysis, therefore incorporating a (partially) different item set and different speakers.

A similar pattern was observed for response times in British English picture naming from Johnston et al. (2010) to that found in Experiment 1 with the Australian English picture naming data: an interaction between the log frequency of the target, and the summed frequency of neighbours of higher frequency than the target. For targets that were low in frequency, having high summed higher frequency PNF resulted in inhibitory effects, while on targets that had high frequency, the effect tended to be facilitatory. This was the case for analyses using the full sets of items and on monosyllabic sets, hence on four (partially overlapping) datasets. Additionally, in the British dataset, a similar interaction was observed between target word log frequency and the number of neighbours of higher frequency than the target.

Overall, the fact that, across Experiments 1 and 2, it was only those measures based on neighbours of higher frequency than the target (higher frequency PND and summed higher frequency PNF) that showed significant effects suggests that neighbours that are lower in frequency than the target most likely have a negligible effect on target retrieval. We examined this claim explicitly in the Australian dataset by investigating the effect on picture naming of the number of neighbours of lower frequency than the target and of their summed frequency (see supplementary materials). No effect of any of these variables was observed nor any interaction between any of these variables and target log frequency (all $p > .05$).

Critically, our results suggest that phonological neighbourhood effects are modulated by target word log frequency: when the target is low in frequency with increasing frequency of higher frequency neighbours, naming is slower, but the effect is reversed when the target is high in log frequency.

In order to develop a clearer understanding of how theories can account for observed patterns, in the next section we will examine whether the effects of phonological neighbourhood that were consistently found in Experiments 1 and 2 across response time analyses (i.e., an interaction between target frequency and summed higher frequency PNF) can be successfully simulated using a computational model, and if so, which parameters are important to do so. Although the results from the present behavioural investigations do not conclusively rule out an effect of the number of higher frequency neighbours (higher frequency PND), at this time, we do not believe that the evidence for this effect is strong enough to justify its inclusion as a benchmark for model success. Therefore, in the next section, we do not consider higher frequency PND in the simulations. Previous computational simulations (Chen & Mirman, 2012; Dell & Gordon, 2003) produced facilitatory effects of phonological neighbourhood density on word production but did not include consideration of target or neighbour frequency. They also used extremely small vocabularies (five words or two words). Consequently, here we examine effects of phonological neighbourhood in an interactive activation computational model with a much larger and more realistic vocabulary, that also implements frequency – DRC-SEM, an adaptation of the Dual Route Cascaded Model (DRC: Coltheart et al., 2001).

Experiment 3: Computational modelling

This experiment aimed to explore the characteristics of the language production system that allow the effects of neighbourhood found in both Experiments 1 and 2 to emerge, using an adaptation of the DRC model (Coltheart et al., 2001). The DRC model of reading has been demonstrated to simulate aspects of human reading behaviour that are relevant to the present study. For instance, the frequency effect on DRC reading speed is remarkably similar to the frequency effect seen in human reading data. In simulating lexical decision "YES" latencies, DRC was able to capture an interaction between frequency and (orthographic) neighbourhood density: high neighbourhood density speeds lexical decision responses to target words but only for low frequency targets.

In order to simulate picture naming, we used an experimental version of the DRC model which allows activation of phonological forms from semantics rather than via print. Coltheart and colleagues (1999) implemented a minimal semantic system consisting in three colour concepts, in order to model the Stroop effect (naming the ink colour of colour words). Building on this rudimentary model, DRC-SEM was constructed with a semantic system that contained one semantic unit for each of around 8,000 of the words in DRC's vocabulary[9] (see Ceccherini, 2015 for description). The semantic system also has (unimplemented) inhibitory within-level links, the ability to add noise or decay, and imageability scaling whereby simulation of potential imageability effects can be examined for the subset of nodes that have associated imageability values. Picture naming can be simulated with DRC-SEM by switching on the unit in the Semantic System that corresponds to the picture. This in turn activates the corresponding unit in the phonological (output) lexicon (or corresponding word node), leading next to the activation of the corresponding phonemes in the phoneme system. The naming response is produced when activation of the phoneme units reaches a pre-established threshold. The time needed to reach this threshold (or number of cycles) constitutes the model's reaction time.

Communication between levels of representation in the DRC-SEM model is both cascaded and interactive. Critical for simulation of phonological neighbourhood effects in naming is that phonological neighbours of a target can be activated in the phonological (output) lexicon by feedback from the phoneme level to the phonological lexicon. Additionally, as frequency is implemented as different resting levels of units in the phonological lexicon, activation levels of the units that represent high-frequency words grow faster than those of units that represent low-frequency words. Hence, DRC-SEM should allow both the activation of phonological neighbours in production and differential levels of activation of units in the phonological lexicon depending on their frequency. Apart from activation between levels, units within levels can also interact through lateral inhibition. For example, an activated unit in the phonological lexicon will inhibit other activated units at that level.

[9] DRC-SEM is based on DRC 1.2.1 and is available for download at *https://maxcoltheart.wordpress.com/drc/* with full detail on the added parameters.

We started our exploration of the characteristics of a computational model that allows simulation of the effects of phonological neighbourhood in the human data, first using a set of "default" DRC-SEM parameters, and then systematically modifying those parameters that were most likely to influence activation of neighbours. We chose to investigate whether the resulting simulated "response times" were predicted by two phonological neighbourhood measures: PND and summed higher frequency PNF. We investigated effects of PND on different simulations (even if no effect of PND was observed in the present study) to investigate whether parameter settings that were able to generate facilitatory effects of PND (as in Chen & Mirman, 2012, and Dell & Gordon, 2003) were or were not able to produce the critical interaction we found between target log frequency and summed higher frequency PNF.

Methods

DRC-SEM's vocabulary is comprised of monosyllabic words. Hence, targets in this experiment were 171 words from the Australian dataset (Experiment 1) that were monosyllabic (ranging between two and five phonemes). Of these, 162 belonged to the British dataset that we analysed in Experiment 2.

DRC-SEM is considered to have produced a word when all of the phonemes of that word have reached a criterion level of activation. Response time is measured in terms of the number of processing cycles required to reach this criterion. A response is considered accurate when the output string of phonemes is the one corresponding to the target.

The number of cycles required to produce a response was used as the dependent variable in a series of multiple regressions. Trials which did not result in a correct response were excluded from the analysis. DRC-SEM is able to implement potential effects of frequency and of length, but not, for example, of visual complexity, age of acquisition or phonotactic probability. It is expected that frequency effects will be stable in the analyses of the simulations, as the omission of these variables removes the intercorrelations between frequency and other predictors that is likely the cause of unstable frequency effects in the human data (e.g., inhibitory effects of frequency in the Australian analyses when including age of acquisition and familiarity as covariates, see discussion above). Target log frequency and length in phonemes were therefore included as predictors in linear regressions run on the number of cycles, together with either, 1) PND, or 2) the PNF variable that had an effect on latencies in both Experiment 1 and 2 (summed higher frequency PNF), with an interaction with target log frequency. As DRC-SEM's phonological lexicon's units are weighted according to their CELEX spoken frequency, CELEX spoken (word-form) log frequency was used both for the target and for the calculation of summed higher frequency PNF (as opposed to combined spoken and written frequency in the human data analyses). All predictors were standardised. Analyses were run in R (R Core Team, 2014).

The majority of DRC-SEM's parameters are shared with DRC, and consequently for our simulations, we used the standard DRC default settings (see Coltheart et al., 2001), with the exception of feature-letter excitation and feature-letter inhibition that were both set to 0 so that no written unit would receive external activation.

DRC-SEM has 16 additional parameters:

1) a parameter that controls the strength of excitation of a specified semantics unit (SemanticsExternalExcitation),
2) an onset parameter specifying the starting cycle for semantic layer activity,
3) six parameters that pertain to the semantic layer itself (lateral inhibition, noise, decay, decay trigger, threshold, strength of the imageability scaling),

4) eight parameters that regulate the strength of connections between the semantic layer and the orthographic and phonological lexicons (in both directions, and both inhibitory and excitatory). Connection strength determines the speed in the rise of activation of the end point unit. For instance, weaker connections between the semantic layer and the phonological output lexicon (as modulated by the parameter named SemanticPhonlexExcitation) mean that the activation of units in the phonological output lexicon will rise more slowly.

By default, these specific parameters are all set to 0. DRC-SEM has never been used to simulate "simple" picture naming, although it has been used to simulate picture-word interference (Ceccherini, 2015), and in colour naming in a Stroop paradigm (Coltheart et al., 1999). To model simple picture naming we set SemanticsExternalExcitation to 0.5 (as in Ceccherini, 2015), and following consultation with Max Coltheart, we set semantics to phonological lexicon excitation (SemanticPhonlexExcitation) to 0.1. Given that the focus of this study was in lexical rather than semantic effects, and we did not have particular hypotheses regarding effects that would arise at the semantic level, all the other DRC-SEM-specific parameters remained at zero, as they had been in previous studies. These initial parameters settings were taken as "default" in the present study (see Supplementary materials for a full list of parameter values).

In a series of simulations, we then systematically adjusted those settings most likely to impact on the activation of neighbours in the phonological lexicon. The criterion that was chosen to deem a simulation successful was the presence, in the analyses of the latencies obtained with the respective simulations, of an interaction between target frequency and summed higher frequency PNF. In addition to this effect, however, we also documented the following criteria, with no particular hierarchy: overall accuracy, target frequency and length effects, and number and nature of units in the phonological lexicon that were activated. This was in order to assess whether the obtained simulations generated other plausible effects that would strengthen the value of these particular simulations, or unusual effects that might undermine this value.

Results

The output of each simulation and the main effects of phonological neighbours are summarised in Table 5. Full results of each analysis are reported in the supplementary materials.

<Insert Table 5 about here>

Simulation 1: Default parameters

With the Default parameters, all 171 target words were correctly 'named'. Multiple regression showed no main effect of PND on the number of cycles DRC-SEM required to produce a response, as well as no main effect of summed higher frequency PNF and no significant interaction between this variable and target log frequency. In addition, a significant facilitatory effect of target log frequency was observed, while length in phonemes did not significantly predict the number of cycles.

Examining the activation levels within the phonological lexicon, revealed that there was rarely activation of any other unit than the target in the phonological lexicon. Only the Phonlex units corresponding to homographs (for bow for instance) were activated, most likely because of bidirectional activation between the phonological and orthographic lexicon units. Our criterion for 'activation' of a unit in the phonological lexicon (Phonlex unit) was set at 0.02, but even when this value was set to 0.000001, no additional Phonlex units were considered activated.

The fact that no effect of phonological neighbourhood measures was found, and that almost no phonological neighbours were activated, is perhaps unsurprising, since the default parameters of DRC feature only limited feedback from the phoneme level to the phonological lexicon. Effects of phonological neighbours have previously been explained within models that feature interactivity between the phonological lexical level and the phonemes. Hence, in Simulation 2, we increased feedback from phonemes aiming to increase activation of phonological neighbours.

Simulation 2: Increasing feedback from the phoneme level to the phonological lexicon

The DRC parameter that regulates feedback from the phoneme level to the phonological lexicon is called PhonemePhonlexExcitation and is set to 0.04 by default. In Simulations 2a and 2b, this parameter value was increased to 0.15 and 0.2 respectively with all other parameter settings remaining identical to Simulation 1.

Simulation 2a: PhonemePhonlexExcitation = 0.15

Once again, all target words were correctly named. Increasing the feedback to the Phonlex units from phonemes did result in more Phonlex units being activated, and the number of activated Phonlex units was significantly correlated to the total number of phonological neighbours of each target word (17 on average; $r = .229$, $p = .003$) and the number of neighbours of higher frequency than the target (5 on average; $r = .209$, $p = .006$). However, the main effect of PND did not reach significance in the multiple regression, and no main effect of summed higher frequency PNF nor interaction of summed higher frequency PNF with target log frequency was observed. A significant facilitatory effect of target log frequency was observed, as well as (interestingly), a significant facilitatory effect of length (longer words were faster named, although the correlation between Simulation 2a number of cycles and length was not significant: $r = -.053$, $p = .492$).

Simulation 2b: PhonemePhonlexExcitation = 0.2

In Simulation 2b, feedback from the phoneme level to the phonological lexicon was further increased (0.2), resulting, once again, in error free naming. The number of activated Phonlex units was also higher with this additional increase in feedback (11.72 on average), these values were still below the actual number of phonological neighbours (17 on average), and remained only moderately correlated to PND ($r = .372$, $p < .001$) and were not correlated with summed higher frequency PNF ($r = .077$, $p = .314$).

There was a main facilitatory effect of PND, but the effect of summed higher frequency PNF did not reach significance, and there was no significant interaction between this variable and target log frequency.

In addition to significant facilitatory effects of target log frequency, effects of length were observed, once again in the unexpected direction: longer words were more likely to be named in fewer cycles. Note that there were no concerning levels of multicollinearity – all VIFs were below 1.73. Length and the number of cycles were negatively correlated with Simulation 2b number of cycles (longer words were associated with fewer cycles, $r = -.221$, $p = .004$).

Therefore, it appears that sufficient increase of feedback from phonemes to the phonological lexicon resulted in facilitatory main effects of phonological neighbourhood (PND), but these effects were accompanied by effects of length (in an unexpected direction) that were not found in the human data. In addition, the critical interaction between target log frequency and summed higher frequency PNF was not reproduced with the parameter settings used here in Simulation 2. Consequently, the next simulations explored the effects of changes of other parameters that could amplify effects of neighbours: Decreasing inhibition from the phoneme level to the phonological lexicon and lateral inhibition within the phonological lexicon.

Simulation 3: Decreasing inhibition of phonological lexical units

In contrast to DRC-SEM, the previous computational models used to simulate facilitatory main effects of PND in spoken word production (Chen & Mirman, 2012; Dell & Gordon, 2003) did not include inhibitory links between phonemes and the phonological lexicon. Additionally, Chen and Mirman (2012) included lateral inhibition within the phonological lexicon (or "word layer"), although Dell and Gordon did not. Furthermore, Coltheart et al. (2001) found that the initial DRC model was only able to model facilitatory effects of orthographic neighbourhood density on word reading aloud if both inhibition from letter to the orthographic lexicon was reduced, and lateral inhibition within the (orthographic and phonological) lexicons was set to zero (p. 224). Hence, in the next three simulations, we systematically investigated the effects of reducing lexical inhibition in DRC-SEM.

Simulation 3a: Decreasing inhibition from the phoneme level to the phonological lexicon

The PhonemePhonlexInhibition parameter has the effect of limiting the activation of other units in the phonological lexicon. In simulation 3a, this value was changed from its default of 0.16 to zero (with the rest of the parameter settings identical to Simulation 1). This set of parameter values was similar to Chen and Mirman's (2012) parameters: featuring both some feedback from phonemes and some competition at the level of lexical units, but no inhibition coming from phonemes, allowing us to assess whether a model similar to Chen and Mirman's model could either yield the facilitatory effects of PND originally found by these authors, or successfully simulate our behavioural findings. This parameter change had very little effect compared to Simulation 1. Only one item showed a difference in number of cycles for naming ("rose" was named in 77 instead of 78 cycles), and no more

Phonlex units were activated compared to Simulation 1. Consequently, as for Simulation 1, there was no effect of PND, no main effect of summed higher frequency PNF and no significant interaction between this variable and target log frequency. In addition, there was a significant facilitatory effect of target log frequency and no effect of length. Hence, a computational model similar to Chen and Mirman's (2012) did not allow us to either replicate their finding of a facilitatory effect of PND, nor did it allow us to successfully simulate the behavioural effects seen in the present study.

Simulation 3b: Decreasing inhibition within the phonological lexicon

Another source of inhibition that could potentially reduce activation of units in the phonological lexicon is the lateral inhibition parameter (PhonlexLateralInhibition). This parameter has a default value of 0.07. In Simulation 3b, this parameter was set to zero, with every other parameter identical to Simulation 1. This change resulted in the exact same output as Simulation 3a: there was the exact same number of cycles for each item, and exactly the same Phonlex units were activated. The results of the multiple regression were, therefore, identical.

Simulation 3c: Decreasing inhibition from the phoneme level to the phonological lexicon, and within the phonological lexicon

Coltheart et al. (2001) reduced both types of inhibition to observe orthographic neighbourhood effects on reading with the initial DRC model. Simulation 3c applied this principle to spoken word production from semantics in DRC-SEM by setting both PhonemePhonlexInhibition and PhonlexLateralInhibition to zero. This parameter setting then approaches Dell and Gordon's (2003) characteristics: it features some feedback from phonemes to the lexical units, but no inhibitory mechanisms either between lexical units or from phonemes to lexical units. Hence, this simulation allowed us to observe whether such a model could successfully reproduce facilitatory effects of PND, and/or account for our behavioural findings.

This simulation resulted in activation of additional units in the phonological lexicon, and, similar to Coltheart and colleagues' findings for orthographic neighbourhood in reading aloud with the initial DRC model, a facilitatory effect of PND. Interestingly, this finding was in the absence of any length effect (but with the expected significant facilitatory effect of target log frequency). This parameter set did not, however, result in a significant effect of summed higher frequency PNF nor any interaction between summed higher frequency PNF and target log frequency.

We can see here that using model settings that resemble those of Dell and Gordon (2003), we could not simulate the critical interaction found in the present behavioural data, while we could successfully simulate the facilitatory effects of PND that Dell and Gordon observed.

Overall, while some main facilitatory effects of both PND and summed higher frequency PNF were seen in Simulations 2 and 3, our human data was not successfully simulated: There was no interaction between summed higher frequency PNF and target log frequency in any of these simulations. Consequently, the next set of simulations aimed to investigate whether parameter settings that should amplify the effects of frequency for the target and for co-activated Phonlex units would result in the critical interaction.

Simulation 4: Increasing the frequency scaling of phonological units

The FrequencyScale parameter regulates how excitable entries in the (phonological) lexicon are. As noted earlier, frequency is implemented in DRC by the resting levels of activation of units in the lexicon (higher frequency targets have higher resting levels of activation) as determined by a constant (ranging from 0 to -1). Whenever the activation level of a unit is updated, this frequency constant is added to the activation as part of the updating. This addition results in activation levels rising more quickly for high frequency units than for units of low frequency (if all other factors are held constant). However, prior to being added, the frequency constant is multiplied by the FrequencyScale parameter (set by default to 0.05), modulating the size of the constant. For larger FrequencyScale values, the difference between the constants added for high and low frequency words is increased and therefore the effects of frequency are magnified.

In Simulation 4, we examined the effects of this manipulation: Default parameters were used, with the exception of an increase in the FrequencyScale parameter.

Simulation 4a: FrequencyScale = 0.07

When FrequencyScale was set to 0.07, response latencies were slowed, but no additional units were activated in the phonological lexicon compared to the default settings. Furthermore, there was no significant effect of PND, summed higher frequency PNF nor an interaction between summed higher frequency PNF and target log frequency. Increased significant facilitatory effects of target log frequency were observed, but no significant effect of length.

Simulation 4b: FrequencyScale = 0.095

When we further increased the frequency scaling parameter to 0.095, latencies were slowed further but no more Phonlex units showed activation. There were still no effects of PND or summed higher frequency PNF nor their interaction with target log frequency. As expected, significant facilitatory effects of target log frequency increased even further than in Simulation 4a. No effects of length were observed.

Simulations 2 to 4 have demonstrated that several parameter settings can lead to the observation of main facilitatory effects of PND, but none of these parameter settings result in the critical interaction between target log frequency and summed higher frequency PNF seen in the human response time. In the next simulation, we investigate whether reducing the speed at which units in the phonological lexicon are activated might allow for this interaction to develop. We achieved this by reducing the parameter that controls the strength of excitatory links from semantics to the phonological lexicon (SemanticsPhonlexExcitation). We had previously set the "default" value for this parameter to 0.1, however, this value was somewhat arbitrary given the paucity of previous simulations using DRC-SEM.

Simulation 5: Decreasing activation from semantics to the phonological lexicon

Simulation 5a: SemanticsPhonlexExcitation = 0.045

In simulation 5a, the default parameters were used, except that the SemanticsPhonlexExcitation parameter was decreased from 0.1 to 0.045.

This was the first simulation where naming errors occurred: 27 words were not named before the threshold (1,000 cycles), resulting in 84% accuracy (human accuracy was around 86%). These words were typically low in frequency. Few additional units were activated in the Phonological Lexicon.

In the multiple regression model including the 144 words that were named accurately, there was no significant main effect of PND, nor of summed higher frequency PNF. However, the critical interaction between target log frequency and summed higher frequency PNF that was present in the human data was observed. As in the human data, high summed higher frequency PNF tended to be inhibitory on low frequency targets, and facilitatory on high frequency targets, compared to low summed higher frequency PNF (see Figure 4). Strong significant effects of target log frequency were observed, but no significant effects of length.

<Insert Figure 4 about here>

This simulation successfully replicated the pattern observed in the human data, and was the first to reproduce the interaction between target log frequency and summed higher frequency PNF. However, there were no additional observable phonological lexicon units activated. Therefore, the next simulations aimed to build on this finding and explored whether combining the parameter changes that induced activation of additional phonological lexicon units, in combination with the reduced activation from semantics, would provide a closer fit to the human data.

Simulation 5b: SemanticsPhonlexExcitation = 0.045, no PhonLex lateral inhibition, no inhibition from Phonemes to Phonlex

As noted above, previous computational models (Chen & Mirman, 2012; Dell & Gordon, 2003) have not included inhibitory links from phonemes to lexical units when simulating (main) facilitatory effects of phonological neighbourhood. In addition, in Simulation 3c we found that removing within level inhibition and inhibitory feedback from the phoneme level resulted in activation of additional phonological lexicon units and a facilitatory main effect of PND. Therefore, in Simulation 5b, in addition to setting the SemanticsPhonlexExcitation parameter to 0.045, we set both the lateral inhibition within the phonological lexicon and the inhibition from phonemes to the phonological lexicon to zero.

This parameter set resulted in 84% accuracy. Overall, the number of cycles for accurate responses correlated very highly between Simulation 5a and Simulation 5b (r = .991, p <.001), showing that the presence or absence of inhibition in the system had little effect on response latencies.

No effect of PND was observed. However, the critical interaction between target log frequency and summed higher frequency PNF was found, which was in the same direction as in Simulation 5a. Additionally, strong significant effects of target log frequency were observed, but there was no significant effect of length.

This observation suggests that the critical interaction does not *require* lateral inhibition within the phonological lexicon, nor inhibition from phonemes. This parameter set led to activation of a large number of Phonlex units. The numbers of units activated for a target were not significantly correlated with the actual number of phonological neighbours (r = .077, p = .361) but were correlated with the number of phonological neighbours of higher frequency than the target (r = .211, p = .011). In addition, although co-activated Phonlex units were most often actual phonological neighbours of the target, they did not always correspond to phonological neighbours, but rather to words with some phonemes in common and with high frequency. For example, with the target *bridge* the Phonlex unit corresponding to the word *from* was activated: While they do have a phoneme in common, clearly, they are not phonological neighbours.

We note here that, although Simulations 5a and 5b do not differ greatly in response latencies, they are very different with regards to the number of Phonlex units they activate. This shows that within this DRC-SEM simulation, the number of activated Phonlex units (or at least, of Phonlex units that may only be distantly related to the target in form) does not seem to be strongly related to response latency.

The fact that Phonlex units that were only distantly related to the target were activated can be explained by the absence of inhibition in a model where interactivity is very strong: Slow activation from semantics results in many more processing cycles prior to the phoneme production threshold being reached and hence, more opportunities for activation of (very frequent) supplementary Phonlex units when there is no inhibition to suppress this activation. Because of the lack of inhibition, one would not expect inhibitory effects of phonological neighbours on latencies. However, it is possible to envisage that there might be facilitatory effects of phonological neighbours, that we do not observe here.

Simulation 5c: SemanticsPhonlexExcitation = 0.045, increased feedback from phoneme level.

In Simulation 5c, we assessed whether the combination of weaker activation from semantics to the phonological lexicon (0.045) and increased feedback from the phoneme level to the phonological lexicon (set to 0.2, as in Simulation 2b) led to the critical interaction.

Accuracy remained at 84%. Additional phonological lexicon units were activated (these Phonlex units corresponded to actual phonological neighbours, unlike in Simulation 5b) and the number of activated Phonlex units was marginally significantly correlated with the actual number of neighbours (PND: r = .169, p = .077), and did not correlate significantly with the number of phonological neighbours of higher frequency than the target (r = .013, p = .895). There was no effect of PND, but the interaction between summed higher frequency PNF and target log frequency was significant, replicating the patterns found in the human data. Target log frequency had a strong significant facilitatory effect, and length did not significantly predict the number of cycles.

In sum, the simulations that showed the desired effects (Simulation 5a, b and c) resulted in some items never accurately named, with roughly the same 144 items being named correctly across the three simulations. The target items that were incorrectly named were significantly lower in log frequency than those items that were correctly named ($t(134)$ = 1.978, p <.001). One possibility is that if the analysis was performed on these 144 items (rather than the full set of 171), the interaction may be present, whatever the parameter settings. Consequently, we reanalysed the data from Simulation 1 (default parameters) but only including this subset of 144 words. This change did not result in the critical interaction, and no effects of summed higher frequency PNF were observed nor any effects of PND (all $p > .05$), showing that the effect was not due to this particular subset of items, and can be attributed to the simulation parameters settings instead. This same procedure was applied to Simulations 2 to 4 and once again, the interaction was not observed. The results of these additional analyses are reported in the supplementary materials.

Simulations 5a to c did successfully reproduce the critical interaction between target word frequency and summed higher frequency PNF. Aside from this finding, all three simulations also showed plausible effects of target frequency and length, but imperfect accuracy. Inspection of the adjusted R squared of each "summed higher frequency PNF model" run on the respective number of cycles for each of these three simulations showed very similar values ($R^2 = 0.679$ for Simulation 5c, and $R^2 = 0.675$ for both Simulation 5a and 5b). While none of the simulations showed activation of a set of phonologically related units corresponding to phonological neighbours of higher frequency than the target, Simulation 5b showed the activation of words that were not phonological neighbours of the target word. This makes it difficult to give preference to this model. Given that there seems to be no frank superiority of any of them, we are unable to adjudicate between these three simulations. Therefore, the only parameter change that we can confidently consider able to contribute to accounting for our findings is a slow rise in activation from semantics to the phonological lexicon that leads to increased interactivity.

Discussion

In order to explore the mechanisms underpinning effects of phonological neighbourhood on word production, Experiment 3 used an extended version of the DRC model of reading, capable of modelling spoken word production through the addition of semantic nodes that can directly activate the phonological output lexicon. The main findings relating to the different simulations, and their possible interpretations will be discussed here.

First, using this model, we have increased our understanding regarding the processing conditions under which effects of PND can be simulated. Indeed, we were able to obtain the facilitatory effects of PND that have previously been simulated using other computational models (Chen & Mirman, 2012; Dell & Gordon, 2003), by using similar parameter settings to these models.

The authors of the two computational models that have been used to model facilitatory effects of PND on spoken word production (Chen & Mirman, 2012; Dell & Gordon, 2003) attributed such effects to, at least in part, the presence of feedback from phoneme nodes to the phonological word nodes. Dell and Gordon (2003) and Chen and Mirman (2012) both used targets of equal (short) length (respectively six CVC words, and five two- phoneme "words"), making it impossible to assess any effects of length in their simulations. In the context of the DRC model with its much larger and more "realistic" lexicon including words of different lengths, and that simulates word frequency effects, we found that the default low levels of feedback, that successfully replicate patterns of adult skilled reading, were not enough to produce facilitatory effects of PND. Although increasing the amount of feedback from phonemes to the phonological lexicon (Simulation 2b) did reproduce an effect of PND on latencies, this was at the expense of disproportionate effects of length in the unexpected direction of facilitation. In contrast, lack of inhibition alone (with little feedback) produced facilitatory effects of PND, and no effect of length, suggesting that no or limited inhibition between phonemes and the phonological level (as is the case in both Dell & Gordon, 2003, and Chen & Mirman, 2012), in combination with limited feedback from phonemes, are critical for successful simulation of facilitatory effects of PND.

While our DRC simulations whose parameter settings resembled those of Chen and Mirman (2012), and Dell and Gordon (2003), did produce facilitatory effects of PND, they did not reproduce the critical interaction between summed higher frequency PNF and target log frequency that we observed in our behavioural findings. It needs to be acknowledged that Chen and Mirman (2012) did implement frequency of (orthographic) neighbours in a simulation of visual word recognition. This was achieved by increasing connection weights between the higher frequency “word” units and their corresponding letter units. Chen and Mirman’s (2012) parameter settings also included inhibitory links within the “word layer” (i.e., the connection strength between units in that layer). The resulting simulations allowed for inhibitory effects of higher frequency (orthographic) neighbours to appear. It is therefore possible, if connection weights between words and corresponding phoneme units were increased in higher frequency phonological neighbours, that this model could simulate inhibitory effects of higher frequency neighbours in spoken word production. However, it is unclear whether such a simulation would also have resulted in the critical interaction between target frequency and the summed frequency of these neighbours of higher frequency than the target.

With Simulation 5, we successfully found three sets of parameter settings capable of reproducing the critical interaction between target log frequency and summed higher frequency PNF that we found in the behavioural data. This set of simulations featured reduced activation from semantics to the phonological lexicon, with or without inhibition (within Phonlex units or between phonemes and the phonological lexicon) and whether or not increased feedback from phonemes was present. These simulations that allowed for the desired effects to emerge had in common, that they allowed for increased effects of frequency (target and neighbour frequency), with activation spreading relatively slowly through the system, allowing increased interactivity between the word form level and the level of phonemes before the word is successfully produced. To obtain this effect, no changes were required from the DRC default settings for reading aloud (that are also used for picture naming). The parameter regulating the excitation of Phonlex units from semantics had to be changed from the initial DRC-SEM value to successfully model the behavioural findings of this study. This initial "default" value of the parameter had been set somewhat arbitrarily and our results suggests its setting should be changed. Implementation of this change did, however, lead to less than perfect accuracy, because there was insufficient activation for very low frequency items to be accurately named. Perfect accuracy would be expected in our simulations that feature no noise and in which activation is always sent to the correct phonological unit from the correct semantic unit. What the less than perfect accuracy suggests, is that the activation of the very low frequency units in the phonological lexicon resulted in a very low level of activation sent from the phonological lexicon to phonemes. That amount of activation was probably insufficient to offset the phoneme decay that is featured in the system. As a consequence, target phonemes for those very low frequency words did not reach enough activation for successful "naming".

Despite the success of these simulations to reproduce behavioural findings with minimal changes in initial DRC-SEM parameter settings, some limitations need to be noted. First, a paradoxical finding in these simulations, is how either effects of PND or an interaction between target log frequency and summed higher frequency PNF could be found in the absence of observable activation of "phonological neighbours" (Phonlex units, such as in Simulations 4b and 5a). Second, it needs to be acknowledged that DRC-SEM, in its current form, has limited ability to simulate effects happening or arising at the level of semantics

(aside from the imageability scaling for some words). For instance, a small number of picture naming studies have shown that the number of semantic features of a given concept (or semantic richness) has a facilitatory effect on latencies and accuracy (Lampe et al., 2021; Rabovsky et al., 2016; Taylor et al., 2012). Consequently, a more comprehensive version of DRC-SEM should incorporate such effects. More generally, a comprehensive model of spoken word production should also be able to simulate the activation of multiple semantically related candidates. This model would need to incorporate an ability to simulate effects related to this co-activation that arise in any task involving spoken word production (e.g., cumulative semantic interference in semantic blocking: e.g., Howard et al., 2006, semantic interference in picture-word interference: e.g., Schriefers et al., 1990, etc.), while being able to account for the present effects of phonological neighbours, and would also allow exploration of whether these semantic effects would be predicted to interact with any other effect, including the phonological effects that are observed here.
In addition, future developments should go beyond DRC-SEM's ability to simulate response time and also improve the extent to which it can simulate accuracy and error types.
Finally, in addition to documenting neighbourhood density and frequency effects in spoken word production, this study suggests a change in the default parameter set for DRC-SEM (decreasing semantics to phonological lexicon excitation) that should be used for further investigation of picture naming.

General Discussion

We have reported two picture naming experiments in English, one with a population of Australian English monolingual speakers, the other using a published dataset of picture naming latencies from British English monolingual speakers. Given the inconsistencies in the previous literature, our aim was to examine the effects of several phonological neighbourhood measures, and in particular, focus on neighbours of higher frequency than the target. Motivated by the literature on word recognition, our study was the first to examine whether there was an interaction between target frequency and phonological neighbourhood in picture naming.

The human data in Experiments 1 and 2 found that the most consistent effects were related to interactions between target frequency and phonological neighbourhood, in particular, summed frequency of phonological neighbours that are of higher frequency than the target. These interactions demonstrated that effects of phonological neighbours can be both facilitatory and inhibitory, and that these effects are intimately connected to frequency effects.

While it has previously been noted that phonological neighbours can generate forces that are either facilitatory or inhibitory, this claim was made in the context of different patterns dependent on task modality (production as opposed to reception) (Chen & Mirman, 2012; Dell & Gordon, 2003). Here we suggest that opposite forces can be generated within spoken word production depending on the frequency of the target and depend critically on those neighbours of higher frequency rather than all neighbours. This possibility gives a potential insight regarding why the literature shows inconsistent results – depending on the frequency of items in a given experimental set and the relative frequency of these items' neighbours, a net facilitatory, null, or even inhibitory main effect of (higher frequency) phonological neighbourhood frequency may emerge.

In Experiment 3, we undertook a systematic series of simulations using the Dual Route Computational model of reading, with an implemented semantic component (DRC-SEM). Our simulations involved a much larger and more realistic model vocabulary compared to the two other computational models of spoken word production that had previously investigated effects of PND (Chen & Mirman, 2012; Dell & Gordon, 2003). DRC-SEM had not been designed with these simulations in mind and hence there was no guarantee that there would be a parameter set in this model that could simulate the human data. Nevertheless, the interaction between frequency and summed higher frequency PNF was reproduced using parameters that resulted in a relatively slow rise of activation in the phonological output lexicon from semantic input.

While Chen and Mirman (2012) and Dell and Gordon (2003) did not investigate the interaction between target frequency and neighbourhood effects, we set the parameters of some simulations so that they would resemble those of Chen and Mirman (2012) and Dell and Gordon (2003), respectively. These parameter configurations were not successful at simulating the crucial interaction. However, it is worth noting that Chen and Mirman did perform a simulation of visual word recognition taking into account the frequency of words: to simulate frequency, connection weights were increased between the higher frequency word units and their corresponding letter units. Chen and Mirman found that words that had higher frequency neighbours were recognised more slowly because those high frequency neighbours became more highly active, more quickly, and took longer to inhibit compared to equal frequency neighbours, and therefore delayed the recognition of the target word. It would be interesting to see if, within this model, increasing the connection weights between the higher frequency word units and their corresponding phoneme units would lead to words being inhibited by neighbours of higher frequency, and whether the effects would also interact with the target's frequency.

A finding of the present study that would merit further investigations is the significant interaction that was observed in the British dataset between target log frequency and higher frequency PND. Although we judged that the evidence was not strong enough in the present study to justify modelling this interaction in Experiment 3, further behavioural research is needed to replicate this finding. If it was discovered that this is actually a robust effect, then this interaction should be modelled in simulations as well, and the potential changes in parameter settings discussed.

An important limitation of the present study resides in the limited consideration of semantic factors, in particular of the influence of semantically related representations in spoken word production. Picture naming is a semantically driven task, but as we noted earlier, previous attempts at demonstrating the effect of the number of semantically related alternatives in picture naming have failed to show significant effects, although no study to date has investigated possible interactions between this and other factors (e.g. frequency). Further research is needed to address this important issue.

Finally, we reported a strong correlation between frequency and both age of acquisition and familiarity. One possibility is that the effects that are attributed to frequency in the current analyses are actually driven by either of these other variables. For instance, some authors have claimed that effects that were attributed to frequency in picture naming were actually related to age of acquisition, based on the fact that when age of acquisition was included as a covariate, effects of frequency disappeared (e.g., Morrison et al., 1992, cited in Bonin et al., 2002). Future research could address this issue by testing the interaction between the age of acquisition and/or familiarity of target words and measures pertaining to the age of acquisition and/or familiarity of phonological neighbours.

Conclusion

In this investigation of effects of phonological neighbourhood on spoken picture naming, we identified a critical interaction between the summed frequency of phonological neighbours of higher frequency than the target and the frequency of the target word. This phonological neighbourhood measure exerted inhibitory effects on low log frequency targets, but facilitatory effects on high log frequency targets. We argue that this observation may underpin the inconsistent findings in previous research, as it shows that neighbours can exert both facilitatory and inhibitory forces, but that the direction of these forces depends on both the frequency of the target and the frequency of its neighbours. Computational simulations allowed us to determine that, in a model featuring interactivity between words and phonemes, increased frequency effects generated by a slow rise in activation from semantics to the phonological lexicon units were needed for this critical interaction to occur.

Our research suggests that in the study of phonological neighbourhoods in spoken word production, the number of neighbours does not matter as much as frequency differences between the target and its neighbours that determine the activation dynamics in the lexicon. As we have demonstrated this research constrains theoretical models and has important implications for comparisons across studies.

Acknowledgements

We thank Max Coltheart and Steven Pritchard for their contribution to computational modelling training of the first author. Both provided valuable discussions during the preparation of this paper. We also thank three reviewers including Steve Lupker for their valuable comments and suggestions that have led to improvements of this manuscript.

References

Abdel Rahman, R., & Melinger, A. (2009). Semantic context effects in language production: A swinging lexical network proposal and a review. *Language and Cognitive Processes*, *24*(5), 713–734. https://doi.org/10.1080/01690960802597250

Alario, F.-X., & Ferrand, L. (1999). A set of 400 pictures standardized for French: Norms for name agreement, image agreement, familiarity, visual complexity, image variability, and age of acquisition. *Behavior Research Methods, Instruments, & Computers*, *31*(3), 531–552. https://doi.org/10.3758/BF03200732

Alario, F.-X., Ferrand, L., Laganaro, M., New, B., Frauenfelder, U. H., & Segui, J. (2004). Predictors of picture naming speed. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 140–155. https://doi.org/10.3758/BF03195559

Allison, P. (2012). *When Can You Safely Ignore Multicollinearity? | Statistical Horizons*. Statistical Horizons. https://statisticalhorizons.com/multicollinearity

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. https://doi.org/10.3758/BF03214334

Arnold, H. S., Conture, E. G., & Ohde, R. N. (2005). Phonological neighborhood density in the picture naming of young children who stutter: Preliminary study. *Journal of Fluency Disorders*, *30*(2), 125–148. https://doi.org/10.1016/J.JFLUDIS.2005.01.001

Baayen, H. R., & Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychological Research*, *3*(2), 12–28. http://www.redalyc.org/html/2990/299023506003/

Baayen, R. H., Piepenbrock, R., & Rijn, van H. (1993). *The {CELEX} lexical data base on {CD-ROM}*.

Barry, C., Morrison, C. M., & Ellis, A. W. (1997). Naming the Snodgrass and Vanderwart

Pictures: Effects of Age of Acquisition, Frequency, and Name Agreement. *The Quarterly Journal of Experimental Psychology Section A*, *50*(3), 560–585. https://doi.org/10.1080/783663595

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models Douglas Bates. *ArXiv E-Print, under Revision*, 1–27. https://doi.org/arXiv:1506.04967

Baus, C., Costa, A., & Carreiras, M. (2008). Neighbourhood density and frequency effects in speech production: A case for interactivity. *Language and Cognitive Processes*, *23*(6), 866–888. https://doi.org/10.1080/01690960801962372

Bernstein Ratner, N., Newman, R., & Strekas, A. (2009). Effects of word frequency and phonological neighborhood characteristics on confrontation naming in children who stutter and normally fluent peers. *Journal of Fluency Disorders*, *34*(4), 225–241. https://doi.org/10.1016/j.jfludis.2009.09.005

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2002). The determinants of spoken and written picture naming latencies. *British Journal of Psychology*, *93*(1), 89–114. https://doi.org/10.1348/000712602162463

Boukadi, M., Zouaidi, C., & Wilson, M. A. (2016). Norms for name agreement, familiarity, subjective frequency, and imageability for 348 object names in Tunisian Arabic. *Behavior Research Methods*, *48*(2), 585–599. https://doi.org/10.3758/s13428-015-0602-3

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*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Ceccherini, L. (2015). *The effects of a concomitant distractor on word reading aloud and picture naming tasks* [Doctoral dissertation, Università delle studi di Padova and

Macquarie University].
http://paduaresearch.cab.unipd.it/7504/1/Ceccherini_Lisa_Thesis.pdf

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive Science*, *34*(4), 685–697. https://doi.org/10.1111/j.1551-6709.2010.01100.x

Chedid, G., Wilson, M. A., Bedetti, C., Rey, A. E., Vallet, G. T., & Brambati, S. M. (2019). Norms of conceptual familiarity for 3,596 French nouns and their contribution in lexical decision. *Behavior Research Methods*, *51*(5), 2238–2247. https://doi.org/10.3758/s13428-018-1106-8

Chen, Q., & Mirman, D. (2012). Competition and cooperation among similar representations: toward a unified account of facilitative and inhibitory effects of lexical neighbors. *Psychological Review*, *119*(2), 417–430. https://doi.org/10.1037/a0027175

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, *30*(2), 441–469. https://doi.org/10.1017/S0305000903005579

Coltheart, M. (1981). The MRC Psycholinguistic Database. *The Quarterly Journal of Experimental Psychology Section A*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

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

Coltheart, M., Woollams, A., Kinoshita, S., & Perry, C. (1999). A position-sensitive stroop effect: Further evidence for a left-to-right component in print-to-speech conversion. *Psychonomic Bulletin and Review*, *6*(3), 456–463. https://doi.org/10.3758/BF03210835

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65–70. https://doi.org/10.3758/BF03206399

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142. https://doi.org/10.1016/0749-596X(88)90070-8

Dell, G. S., & Gordon, J. K. (2003). Neighbors in the lexicon! Friends or foes? In N. O. Schiller & A. S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production* (pp. 9–47). Mouton de Gruyter.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical Access in Aphasic and Nonaphasic Speakers. *Psychological Review*, *104*(4), 801–838. https://doi.org/10.1037/0033-295X.104.4.801

Ellis, A. W., Lum, C., & Lambon Ralph, M. A. (1996). On the use of regression techniques for the analysis of single case aphasic data. *Journal of Neurolinguistics*, *9*(3), 165–174. https://doi.org/10.1016/0911-6044(96)00008-5

Ellis, A. W., & Morrison, C. M. (1998). Real age-of-acquisition effects in lexical retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(2), 515–523. http://psycnet.apa.org/buy/1998-00017-015

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503

Frank, A. F., Tanenhaus, M. K., Aslin, R. N., & Salverda, A. P. (2007). Frequency, neighborhood density, and phonological similarity effects in picture naming: An artificial lexicon study. *Proceedings of the 29th Annual Cognitive Science Society*, 995–1000. http://csjarchive.cogsci.rpi.edu/proceedings/2007/docs/p995.pdf

Ginestet, C. (2011). Ggplot2 : elegant graphics for data analysis. *Journal of the Royal Statistical Society, Series A*, *174*, 245–246.

Gordon, J. K., & Kurczek, J. C. (2013). The ageing neighbourhood: Phonological density in naming. *Language, Cognition and Neuroscience*, *29*(3), 326–344. https://doi.org/10.1080/01690965.2013.837495

Grainger, J., O'Regan, J. K., Jacobs, A. M., & Segui, J. (1989). On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics*, *45*(3), 189–195. https://doi.org/10.3758/BF03210696

Hall, K. C., Allen, B., Fry, M., Mackie, S., & McAuliffe, M. (2016). *Phonological CorpusTools, Version 1.2.* https://phonologicalcorpustools.github.io/CorpusTools/

Hameau, S., Nickels, L., & Biedermann, B. (2019). Effects of Semantic Neighbourhood Density on Spoken Word Production. *Quarterly Journal of Experimental Psychology*, *72*(12), 2752–2775. https://doi.org/10.1177/1747021819859850

Howard, D., Nickels, L., Coltheart, M., & Cole-Virtue, J. (2006). Cumulative semantic inhibition in picture naming: experimental and computational studies. *Cognition*, *100*(3), 464–482. https://doi.org/10.1016/J.COGNITION.2005.02.006

Hutcheson, G., & Sofroniou, N. (1999). Ordinary Least-Squares Regression. In *The Multivariate Social Scientist* (pp. 56–114). Sage Publications. https://doi.org/10.4135/9780857028075

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word Frequency Effects in Speech Production: Retrieval of Syntactic Information and of Phonological Form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 824–843. https://doi.org/10.1037/0278-7393.20.4.824

Johnston, R. A., Dent, K., Humphreys, G. W., & Barry, C. (2010). British-English norms and naming times for a set of 539 pictures: The role of age of acquisition. *Behavior Research Methods*, *42*(2), 461–469. https://doi.org/10.3758/BRM.42.2.461

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lampe, L., Hameau, S., & Nickels, L. (2021). Semantic variables both help and hinder word production: behavioural evidence from picture naming. *Manuscript Submitted for Publication*, *Department of Cognitive Science*, *Macquarie University*.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1999), 1–37.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1–16. https://doi.org/10.3389/fpsyg.2015.01171

Luce, P. A. (1987). Neighborhoods of Words in the Mental Lexicon. *Dissertation Abstracts International, B: Sciences and Engineering*, *47*(12). https://eric.ed.gov/?id=ED353610

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. http://www.ncbi.nlm.nih.gov/pubmed/9504270

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *ACL–MIT Press series in natural language processing. Cognitive models of speech processing: Psycholinguistic and computational perspectives* (The MIT Pr, pp. 122–147). http://psycnet.apa.org/record/1991-97555-005

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLoS ONE*, *7*(8), e43230. https://doi.org/10.1371/journal.pone.0043230

Mirman, D., & Graziano, K. M. (2013). The Neural Basis of Inhibitory Effects of Semantic and Phonological Neighbors in Spoken Word Production. *Journal of Cognitive Neuroscience*, *25*(9), 1504–1516. https://doi.org/10.1162/jocn_a_00408

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition*, *20*(6), 705–714. https://doi.org/10.3758/BF03202720

Newman, R. S., & Bernstein Ratner, N. (2007). The Role of Selected Lexical Factors on Confrontation Naming Accuracy, Speed, and Fluency in Adults Who Do and Do Not Stutter. *Journal of Speech, Language, and Hearing Research*, *50*, 196–213. https://doi.org/10.1044/1092-4388(2007/016

Newman, R. S., & German, D. J. (2002). Effects of Lexical Factors on Lexical Access among Typical Language-Learning Children and Children with Word-Finding Difficulties. *Language and Speech*, *45*(3), 285–317. https://doi.org/10.1177/00238309020450030401

Newman, R. S., & German, D. J. (2005). Life Span Effects of Lexical Factors on Oral Naming. *Language and Speech*, *48*(2), 123–156. http://journals.sagepub.com/doi/pdf/10.1177/00238309050480020101

Pérez, M. A. (2007). Age of acquisition persists as the main factor in picture naming when cumulative word frequency and frequency trajectory are controlled. *Quarterly Journal of Experimental Psychology*, *60*(1), 32–42. https://doi.org/10.1080/17470210600577423

Perret, C., & Bonin, P. (2018). Which variables should be controlled for to investigate picture naming in adults? A Bayesian meta-analysis. *Behavior Research Methods*, 1–13.

https://doi.org/10.3758/s13428-018-1100-1

Protopapas, A. (2007). Check Vocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*(4), 859–862. https://doi.org/10.3758/BF03192979

R Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. http://www.r-project.org/.

Rabovsky, M., Schad, D. J., & Abdel Rahman, R. (2016). Language production is facilitated by semantic richness but inhibited by semantic density: Evidence from picture naming. *Cognition*, *146*, 240–244. https://doi.org/10.1016/j.cognition.2015.09.016

Sadat, J., Martin, C. D., Costa, A., & Alario, F.-X. (2014). Reconciling phonological neighborhood effects in speech production through single trial analysis. *Cognitive Psychology*, *68*, 33–58. https://doi.org/10.1016/J.COGPSYCH.2013.10.001

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*(1), 86–102. https://doi.org/10.1016/0749-596X(90)90011-N

Severens, E., Lommel, S. Van, Ratinckx, E., & Hartsuiker, R. J. (2005). Timed picture naming norms for 590 pictures in Dutch. *Acta Psychologica*, *119*(2), 159–187. http://www.ncbi.nlm.nih.gov/pubmed/15877979

Stemberger, J. P. (2004). Neighbourhood effects on error rates in speech production. *Brain and Language*, *90*(1–3), 413–422. https://doi.org/10.1016/S0093-934X(03)00452-8

Suárez, L., Tan, S. H., Yap, M. J., & Goh, W. D. (2011). Observing neighborhood effects without neighbors. *Psychonomic Bulletin & Review*, *18*(3), 605–611. https://doi.org/10.3758/s13423-011-0078-9

Székely, A., & Bates, E. (2000). Objective visual complexity as a variable in studies of picture naming. *UCSD Centre for Research in Language Newsletter*, *12*(2), 3–33.

Székely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., Lu, C. C., Pechmann, T., Pléh, C., Wicha, N., Federmeier, K., Gerdjikova, I., Gutierrez, G., Hung, D., Hsu, J., Iyer, G., Kohnert, K., Mehotcheva, T., Orozco-Figueroa, A., … Bates, E. (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language*, *51*(2), 247–250. https://doi.org/10.1016/j.jml.2004.03.002

Tabak, W., Schreuder, R., & Baayen, R. H. (2010). Producing inflected verbs: A picture naming study. *The Mental Lexicon*, *5*(1), 22–46. https://doi.org/10.1075/ml.5.1.02tab

Taylor, K. I., Devereux, B. J., Acres, K., Randall, B., & Tyler, L. K. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*, *122*(3), 363–374. https://doi.org/10.1016/j.cognition.2011.11.001

Vitevitch, M. S. (1997). The Neighborhood Characteristics of Malapropisms. *Language and Speech*, *40*(3), 211–228.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(4), 735–747. https://doi.org/10.1037//0278-7393.28.4.735

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and Lexical Representations in Speech Production: Effects of Phonotactic Probability and Onset Density. *Journal of Experimental Psychology: Learning Memory and Cognition*, *30*(2), 514–529. https://doi.org/10.1037/0278-7393.30.2.514

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of Processing in Perception of Spoken Words. *Psychological Science*, *9*(4), 325–329. https://doi.org/10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487. https://doi.org/10.3758/BF03195594

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory and Cognition*, *31*(4), 491–504. https://doi.org/10.3758/BF03196091

Vitevitch, M. S., & Stamer, M. K. (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*(6), 760–770. https://doi.org/10.1080/01690960500287196

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Wurm, L. H., & Fisicaro, S. A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37–48. https://doi.org/10.1016/J.JML.2013.12.003

Zipf, G. K. (1935). *The psychology of language*. Hougton-Mifflin.

Data availability

Raw data for this manuscript can be downloaded from https://data.mendeley.com/datasets/68f3fg56ff/draft?a=10f2aa14-fa49-4834-927c-400d0d336117

Table 1. Summary of the main findings of studies investigating the influence of PND/PNF on picture naming in unimpaired English speakers, sorted by measure of interest and target population.

Paper	Measure of interest	Population	Total Items	Controlled for / manipulation of	design	effect on latency	effect on accuracy
Gordon & Kurczek (2013)	(residual) PND*	31 young adults	200	Frequency, length, phonotactic probability	continuous	ø	ø
Newman & Bernstein Ratner (2007)	PND	24 young adults	44	Frequency, initial phoneme, PNF.	factorial	↗ (marginal)	↗
Vitevitch (2002): Expt 3	PND	34 young adults	48	Familiarity, frequency, length, PNF.	factorial	↗	ø
Vitevitch (2002): Expt 4	PND	25 young adults	48	Familiarity, frequency, length, PNF, phonotactic probability.	factorial	↗	ø

Vitevitch (2002): Expt 5	PND	25 young adults	49	Familiarity, frequency, length, PNF, phonotactic probability.	factorial	↗	ø
Vitevitch et al. (2004): Expt 3	PND	24 young adults	44	Familiarity, frequency, initial phoneme, length, PNF, phonotactic probability, visual complexity.	factorial	ø	ø
Newman & German (2005)	PND	690 adolescents & 530 adults	44**	Age of acquisition***, familiarity, frequency.	factorial	NA	↘
Gordon & Kurczek (2013)	(residual) PND*	42 older adults	200	Frequency, length, phonotactic probability.	continuous	↘	ø
Bernstein Ratner et al. (2009)	PND	15 children	44	Frequency, initial phoneme, PNF.	factorial	ø	↗
Arnold et al. (2005)	PND	14 children	8	Age of acquisition, familiarity, frequency,	factorial	↘	↘

				length, phonotactic probability.			
Newman & German (2002)	PND	≈ 270 children	72	Age of acquisition***, frequency, PNF.	factorial	NA	↘
Newman & German (2002)	Frequency-weighted PND	≈ 270 children	64	Age of acquisition***, familiarity, frequency.	factorial	NA	↘
Newman & German (2002)	higher frequency PND	≈ 270 children	26	Age of acquisition***, frequency, frequency-weighted PND, PND, PNF.	factorial	NA	↘
Newman & Bernstein Ratner (2007)	(average) PNF	24 young adults	44	Frequency, initial phoneme, PND.	factorial	↗	↗
Vitevitch & Sommers (2003): Expt 3	(average) PNF	21 older adults	54	Familiarity, frequency, length, PND, phonotactic probability.	factorial	↗	↗

Bernstein Ratner et al. (2009)	(average) PNF	15 children	44	Frequency, initial phoneme, length, PND.	factorial	↗	↗
Newman & German (2002)	(average) PNF	≈ 270 children	60	Age of acquisition***, familiarity, frequency, PND.	factorial	NA	↗
The present study	PND, higher frequency PND, PNF, summed higher frequency PNF	40 young adults	359	Age of acquisition, familiarity, target log frequency, imageability, length, name agreement, phonotactic probability, visual complexity.	continuous	Interaction between target log frequency and summed higher frequency PNF.	

↗: facilitation, ↘: inhibition, ø: non-significant effect, NA: not investigated in that study. *residuals of PND regressed on length. ** the dependent variable is a combination of tasks (picture naming, open-end sentences & category sorting). *** some words did not have age of acquisition ratings.

Table 2. Pairwise Pearson correlation coefficients between the control predictors and the PND/PNF predictors (n=359).

	VisComp	IMG	AoA	Fam	NameAg	LengthP	Phonotact	LogFreq	PND	Higher F PND	PNF
IMG	-0.001										
AoA	0.062	-.129*									
Fam	-.117*	0.042	-.649***								
NameAg	0.038	.139*	-.406***	.253***							
LengthP	0.082	.132*	.448***	-.282***	-.117*						
Phonotact	-0.054	-0.008	0.061	-0.039	0.02	.175***					
LogFreq	-0.04	-.121*	-.512***	.498***	.227***	-.500***	0.029				
PND	-0.089	-.136*	-.314***	.225***	.118*	-.708***	-0.092	.483***			
Higher F PND	-0.083	-0.082	-0.07	-0.017	0.047	-.535***	-.131*	0.068	.712***		
PNF	-0.095	-.125*	-.317***	.232***	.125*	-.692***	-0.088	.463***	.973***	.749***	
Summed higher F PNF	-0.034	-0.101	-0.101	-0.026	0.315	-.234***	-0.049	-0.026	.253***	.360***	.275***

*** Correlation is significant at p<.001 (2-tailed); *Correlation is significant at p<.05 (2-tailed)

Abbreviations: IMG = imageability (available for 265 items), VisComp = visual complexity, AoA = age of acquisition, Fam = familiarity, NameAg = name agreement, LogFreq = target log frequency, LengthP = length in phonemes, Phonotact = phonotactic probability, PND = phonological neighbourhood density, Higher F PND = number of phonological neighbours of higher frequency than the target, PNF = phonological neighbourhood (summed) frequency, Summed higher F PNF = summed frequency of the neighbours of higher frequency than the target.

Table 3. Australian response time and accuracy: summary of the effects of each PND/PNF predictor and their interaction with frequency when added to the base model: analyses on the whole set of 359 items (265 for accuracy analyses), and on the reduced set of 183 items (169 for accuracy analyses)).

		Response time						Accuracy					
		AIC	Estimate	Std. Error	t value	p value	vif	AIC	Odds ratio	Std. Error	z value	p value	vif
359 items	Model 1: PND	164833						6478					
	PND		-3.64	7.14	-0.51	.610	1.18		0.96	0.15	-0.25	.804	2.59
	PND: target F		-5.48	6.47	-0.85	.397	1.05		1.09	0.12	0.76	.448	1.55
	Model 2: Higher F PND	164832						6475					
	Higher F PND		-2.20	6.48	-0.34	.734	1.02		1.16	0.11	1.28	.201	1.69
	Higher F PND: target F		NA	NA	NA	NA	NA		1.25	0.14	1.59	.113	1.06
	Model 3: PNF	164832						6478					
	PNF		-5.79	6.32	-0.92	.360	1.09		1.05	0.14	0.34	.735	2.39
	PNF: target F		-9.04	7.43	-1.22	.224	1.09		1.07	0.12	0.57	.571	1.45
	Model 4: Summed higher F PNF	164830						6475					
	Summed higher F PNF		-7.28	5.39	-1.35	.177	1.01		1.13	0.11	1.08	.279	1.64

	Summed higher F PNF: target F		-20.23	9.64	-2.10	**.036**	1.16		1.25	0.14	1.66	.097	1.02
183 monosyllabic items	Model 1: PND	85469						3940					
	PND		16.56	10.10	1.64	.101	1.16		0.97	0.16	-0.23	.820	1.74
	PND: target F		-6.97	7.21	-0.97	.333	1.27		1.06	0.11	0.50	.617	1.06
	Model 2: Higher F PND	85469						3990					
	Higher F PND		3.57	14.25	0.25	.802	1.20		1.38	0.18	1.75	.079	2.56
	Higher F PND: target F		-14.89	8.35	-1.78	.075	1.29		1.19	0.13	1.40	.163	1.49
	Model 3: PNF	85470						3940					
	PNF		10.96	9.21	1.19	.234	1.10		1.08	0.15	0.53	.600	1.69
	PNF: target F		-10.49	7.38	-1.42	.156	1.25		1.04	0.12	0.33	.743	1.02
	Model 4: Summed higher F PNF	85468						3937					
	Summed higher F PNF		-2.35	9.94	-0.24	.813	1.28		1.35	0.17	1.78	.076	2.17
	Summed higher F PNF: target F		-19.98	9.08	-2.20	**.028**	1.49		1.20	0.13	1.46	.145	2.29

F = frequency, Std.Error = standard error, vif = variance inflation factor. Each PND/PNF predictor was added to the base model including, for the response time models: Name agreement, familiarity, log frequency and age of acquisition, and for the accuracy models, age of acquisition and log frequency, with random intercepts for participant and item.

Table 4. British English response time: Summary of the effects of each PND/PNF predictor and their interaction with frequency.

		Response time					
		AIC	Estimate	Std. Error	t value	p value	vif
412 items	Model 1: PND	119007					
	PND		2.70	7.69	0.35	.726	1.25
	PND : target F		-6.11	6.76	-0.90	.366	1.07
	Model 2: Higher F PND	119003					
	Higher F PND		-2.27	6.92	-0.33	.743	1.02
	Higher F PND : target F		-25.23	8.55	-2.95	**.003**	1.12
	Model 3: PNF	119007					
	PNF		1.87	8.59	0.22	.828	1.50
	PNF : target F		-8.04	8.29	-0.97	.332	1.13
	Model 4: Summed higher F PNF	119004					
	Summed higher F PNF		1.51	6.67	0.23	.821	1.04
	Summed higher F PNF : target F		-23.16	9.29	-2.49	**.013**	1.11
210 mono	Model 1: PND	60944					

PND		-0.92	10.74	-0.09	.932	1.25
PND : target F		-3.22	10.42	-0.31	.757	1.71
Model 2: Higher F PND	60940					
Higher F PND		-2.89	8.71	-0.33	.740	1.09
Higher F PND : target F		-23.71	11.16	-2.12	**.034**	1.03
Model 3: PNF	60943					
PNF		-1.28	10.49	-0.12	.903	1.21
PNF : target F		-7.30	10.20	-0.72	.474	1.47
Model 4: Summed higher F PNF	60940					
Summed higher F PNF		-1.76	8.34	-0.21	.833	1.05
Summed higher F PNF : target F		-22.38	10.52	-2.13	**.033**	1.02

F=frequency, Std.Error=standard error, vif=variance inflation factor. Each PND/PNF predictor and interaction was added to the base model including picture name agreement, familiarity, log frequency, and age of acquisition, with random intercepts for participant and item.

Table 5. Summary of simulation output characteristics and effects of PND and higher frequency PNF on number of cycles, with a summary of behavioural results from Experiments 1 and 2.

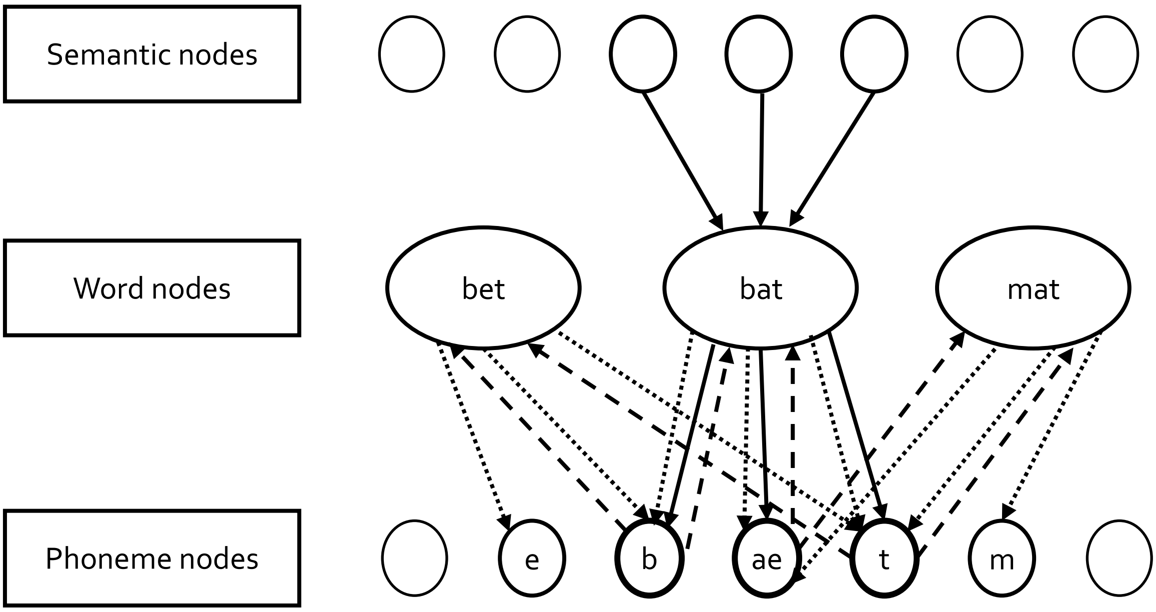
	% Accuracy	Mean Cycles per item	Mean Phonlex units activated (other than target)	Significant effects on response latency: t-values*		
				PND	Summed higher frequency PNF	Interaction Summed higher frequency PNF * target log frequency
HUMAN PICTURE NAMING		Mean RT (ms)				
Experiment 1: Australian latencies (full set)	86.72	917	NA	-0.51	-1.35	**-2.10**
Experiment 2: British latencies (full set)	87.17	923	NA	0.35	-0.10	**-2.49**
SIMULATION 1: Default settings	100	75.26	0.02	-0.63	0.28	0.16
SIMULATION 2: Increased feedback from phonemes to the phonological output lexicon						
2a. Phoneme-Phonlex activation 0.14	100	70.78	4.65	-1.67	-0.39	0.19

2b. Phoneme-Phonlex activation 0.2	100	67.75	11.73	**-3.69**	-1.82	0.02
SIMULATION 3: Reduced Inhibition						
3a. No inhibition from phonemes to Phonlex	100	75.25	0.02	-0.62	0.28	0.15
3b. No Phonlex lateral inhibition	100	75.25	0.02	-0.62	0.28	0.15
3c. No inhibition from Phonemes and no lateral inhibition	100	74.35	15.84	**-2.40**	-1.21	0.32
SIMULATION 4: Increased frequency scaling						
4a. PhonLex frequency scaling 0.07	100	106.12	0.02	-1.00	-0.02	-0.27
4b. Phonlex frequency scaling 0.095	100	114.05	0.01	-1.04	-0.68	-0.58
SIMULATION 5: Weakened Semantics to Phonlex excitation						
5a. Semantics to Phonlex 0.045	84	146.67	0.01	-0.09	-0.61	**-2.77**
5b. Semantics to Phonlex 0.045 AND no inhibition	84	146.31	25.53	-0.49	-0.86	**-2.73**
5c. Semantics to Phonlex 0.045 AND increased Phoneme-Phonlex feedback 0.2						
	84	114.07	12.02	-0.82	-0.45	**-2.52**

* Significant effects are in bold.

Note: For simulations, all parameters have default values except for those noted in the table (for default values see Supplementary materials)

Figure 1. Graphic representation of the activation of phonological neighbours within an interactive activation model without within or between level inhibition or competition (e.g., Dell et al., 1997).



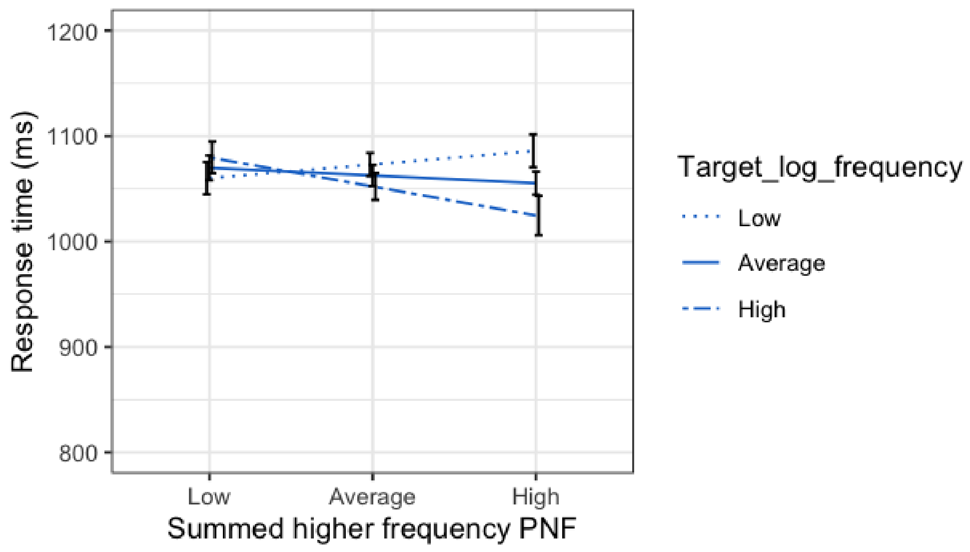
Legend:

Step 1 (plain arrows): Semantic nodes activate the target word node which activates the corresponding phoneme nodes.

Step 2 (dashed arrows): Target phonemes send activation back to the target word node AND its phonological neighbours.

Step 3 (dotted arrows): Target and neighbours send activation back to the phoneme nodes, which are then more strongly activated.

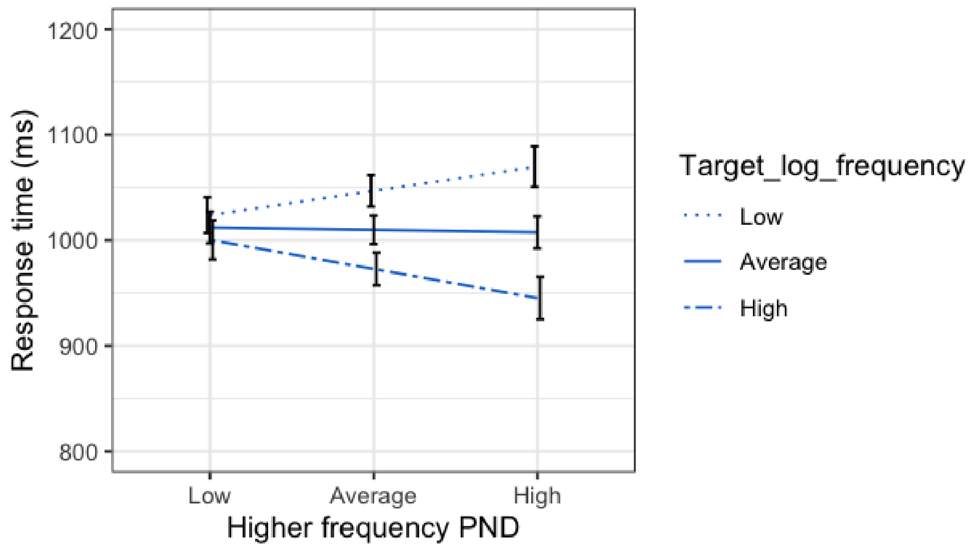
Figure 2. Effect of Summed higher frequency PNF on response time, moderated by target log frequency (Australian dataset). Plot created with ggplot2 (Ginestet, 2011).



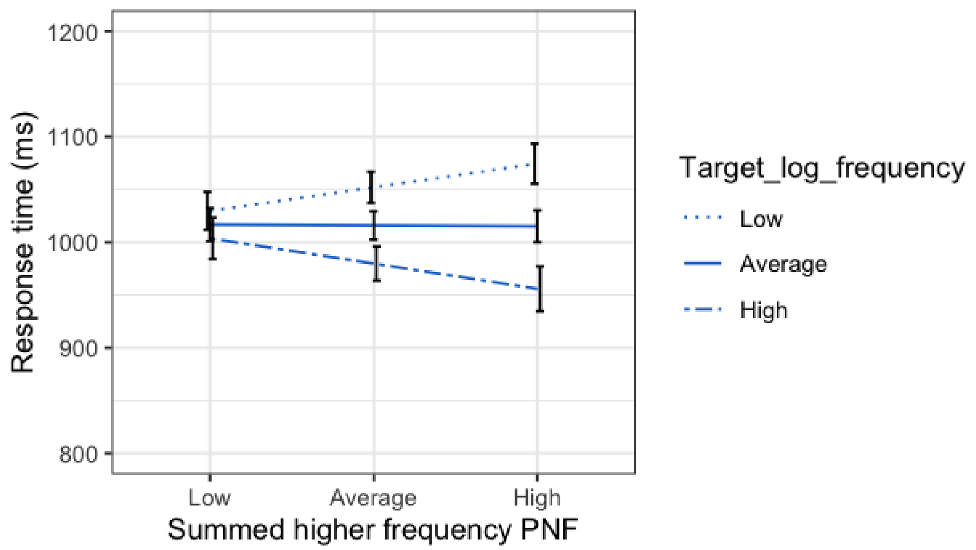
Note: “low” and “high” values correspond, respectively, to -1 SD and +1 SD

Figure 3. Effect of, 1) Higher frequency PND; and 2) Summed higher frequency PNF on response time, moderated by target log frequency (British dataset). Plot created with ggplot2 (Ginestet, 2011).

1)

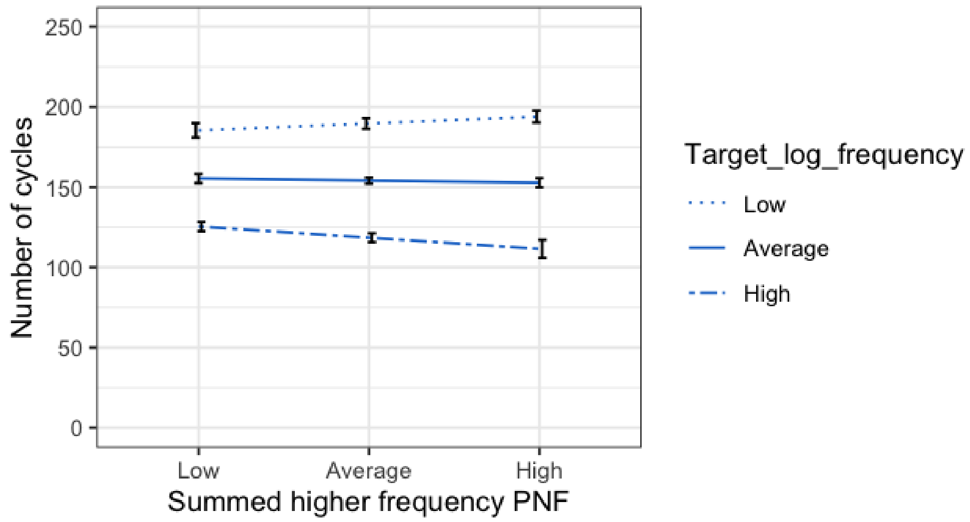


2)



Note: “low” and “high” values correspond, respectively, to -1 SD and +1 SD

Figure 4. Effect of Summed higher frequency PNF on number of cycles, moderated by target log frequency (DRC-SEM Simulation 5a). Plot created with ggplot2 (Ginestet, 2011).



Note: “low” and “high” values correspond, respectively, to -1 SD and +1 SD

APPENDIX A

Experiments 1 and 2: Predictors

A1. Description of measures:

Trial number: Baayen and Milin (2010) note that there are temporal dependencies between successive trials in many experiments. They argue that the inclusion of trial number in the model helps improving the fit and clarifying the role of the predictors of interest. Trial number reflects the order at which a given item was presented to the participant.

Name agreement refers to the degree to which participants agree on the name of the picture. Name agreement measures can either be obtained offline on a separate group of participants similar to the experimental group (e.g., Alario & Ferrand, 1999; Barry et al., 1997; Ellis & Morrison, 1998) or calculated on the basis of the accuracy of the experimental participants (Sadat et al., 2014; Severens et al., 2005). The stimuli used here were selected on the basis of high name agreement values from the IPNP, but these were from American English participants. As name agreement can differ across English varieties (as shown for example by the relatively low correlation ($r < .5$) between British and American English name agreement norms in Barry et al. (1997), in our analysis we used the mean accuracy of our participants on each of the 359 items as a measure of "speeded" Australian name agreement (the same way Sadat et al. 2014 and Severens et al., 2005, did).
Objective **visual complexity** values were retrieved from the IPNP (Székely et al., 2004) and were available for all items: this measure consists of the size of the digitized stimuli picture files in kilobytes. This measure has been suggested to be preferable to subjective ratings of

visual complexity which have been shown to be often confounded with familiarity (Székely & Bates, 2000).

Familiarity and **Age of Acquisition** ratings were drawn from a British English norming study (Johnston et al., 2010). Values were available for all the 359 final items. Familiarity values are ratings of how usual or unusual a concept/object is in the rater's realm of experience, on a seven-point scale ranging from very unfamiliar to highly familiar. Age of Acquisition is a subjective estimate of the age at which the name of the object was learned, choosing between seven age bands.

Values of log summed spoken and written word form **frequency** were obtained from the CELEX database (British English: Baayen et al., 1993). Additionally, frequency measures were retrieved from a more recent frequency database (SUBTLEX-UK: Brysbaert & New, 2009).

Ratings of **imageability** (the ease with which a word gives rise to a sensory mental image) were obtained from the MRC database (Coltheart, 1981), and were available for 265 of the final 359 experimental items.

Word length was the number of phonemes in each target word.

Phonotactic probability was calculated using Vitevitch & Luce's (2004) algorithm: average unigram or bigram positional probabilities across a word. The measure was computed using an online program (Phonological Corpus Tools: Hall et al., 2016).

Phonological neighbourhood density (PND) was calculated using the online program CLEARPOND (Marian et al., 2012), which uses the one-phoneme difference rule (words were neighbours if they shared all but one phoneme, either substituted, added or deleted)[10].

Higher frequency PND: In addition to total PND, we also considered more specifically the number of neighbours that were (numerically) higher in frequency than the target word.

Phonological neighbourhood frequency (PNF): Summed log frequency of phonological neighbours (frequencies were taken from the CELEX database, Baayen et al., 1993).

Summed higher frequency PNF: Summed log frequency of the phonological neighbours that are (numerically) higher in frequency than the target word.

[10] The CLEARPOND online program is designed such that possible neighbours are limited to words belonging to an educated monolingual adult's lexicon (a frequency threshold was used, so that the number of words in the corpus – 27,751 for English – is a reasonable vocabulary size estimate). This is to ensure that there are no rare words that are unlikely to be known by most adult monolinguals.

A2. Means and standard deviations of all item-related predictors used in analyses, before standardisation.

	359 Australian items mean *(SD)*	412 UK items mean *(SD)*
Visual complexity in Kbytes (IPNP)	16,665 *(8911)*	
Visual complexity rating /7 (Johnston et al., 2010)		3.65 *(1.00)*
Name agreement (per item percent accuracy)	86 *(34)*	87*(14)*
Familiarity /7 (Johnston et al., 2010)	4.89 *(1.17)*	4.90 *(1.17)*
Age of acquisition /7 (Johnston et al., 2010)	2.99 *(0.81)*	3.00 *(0.83)*
Celex Log combined word-form frequency	2.26 *(0.81)*	2.31 *(0.69)*
Imageability (range 100-700) (MRC)	(265 items) 592 *(34)*	(294 items) 591 *(37)*
Word Length (in phonemes)	4.37 *(1.64)*	4.49 *(1.57)*
Phonotactic probability (Vitevitch & Luce, 2004)	.005 *(.003)*	.005 *(.004)*
PND measures		
PND	15.15(15.41)	13.98 *(14.39)*
Higher frequency PND	3.38 (4.61)	3.96 *(5.37)*

PNF measures		
PNF	27.10 *(29.45)*	26.80 *(29.29)*
Summed higher frequency PNF	12.61 *(36.66)*	12.54 *(17.02)*