The influence of item-level contextual history on lexical and semantic judgments by children and adults

Yaling Hsiao, Megan Bird, Helen Norris, Ascensión Pagán & Kate Nation

University of Oxford, UK

Correspondence to this article should be addressed to Yaling Hsiao or Kate Nation, both at the Department of Experimental Psychology, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK. E-mails: yaling.hsiao@psy.ox.ac.uk and kate.nation@psy.ox.ac.uk. Helen Norris is now at the Institute of Education, University of Reading and Ascensión Pagán is now at the Department of Neuroscience, Psychology and Behaviour, University of Leicester. This work was supported by grants from The Economic and Social Research Council (ES/M009998/1) and The Leverhulme Trust (RPG-2015-070) to Kate Nation. Yaling Hsiao is supported by a British Academy Post-Doctoral Fellowship (PF2\180013). We are grateful to the Children's Dictionaries Department at Oxford University Press for their collaboration and support. All the materials associated with the study are also available on the Open Science Framework website (https://osf.io/7hz5p/).

Abstract

Semantic diversity quantifies the similarity in the content of contexts a word has been experienced in. Four experiments investigated its effect on lexical and semantic judgments in 9-10 year-olds and adults. In Experiment 1, a cross-modal semantic judgment task, participants decided whether a visually presented word matched an audio definition. Both groups were slower to respond to words high in semantic diversity and this effect was modulated by task demands. Experiment 2 used the same items but in a lexical decision task. Children were faster to respond to words high in diversity but there was no effect in adults, failing to replicate previous work. Experiment 3 examined possible reasons for this while Experiment 4 tested the effect of semantic diversity on lexical decision via secondary analysis of two large megastudies. Overall, the facilitative effect of semantic diversity on lexical decision was robust. Our findings show that contextual experience influences subsequent lexical processing, consistent with context inducing semantic representations that reflect continuities and gradations in meaning. These gradations are captured by semantic diversity, and in turn, this interacts with task demands to influence behavioural performance.

Keywords: semantic diversity, semantic judgment, lexical decision, corpus analysis, word processing

The influence of item-level contextual history on lexical and semantic judgments by children and adults

The meaning of a word is closely related to the contexts in which it appears. Converging evidence from experiments and computational models indicates that this contextual history shapes lexical organisation (for review, Jones, Dye, & Johns, 2017). In this paper, we investigated how contextual experience influences lexical processing as children and adults made lexical and semantic judgments to words that varied in semantic diversity.

Semantic diversity is a metric that quantifies the similarity in the content of contexts a word has been experienced in. To calculate semantic diversity, Hoffman, Lambon Ralph and Rogers (2013) used latent semantic analysis of large text corpora to represent in multidimensional space each context a given word appears in. A word's semantic diversity value corresponds to the mean distance between all the contexts it appears in. Contexts that are more similar to each other are closer in space, reflecting the fact that words high in semantic diversity appear in overlapping and inter-related contexts. As an example, Hoffman and Woollams (2015) considered *spinach*, a low semantic diversity word that tends to occur in a restricted range of contexts related to food. In contrast, a high diversity word like *chance* appears in a range of different contexts. Put simply, *spinach* provides a reasonable clue as to the content of a context whereas *chance* does not. Not surprisingly, semantic diversity is moderately correlated with frequency (r= .49 according to Hoffman et al., 2013); this makes sense as words that occur more often are more likely to feature in more varied contexts.

In related work, Jones and colleagues (Jones, Dyes, & Johns, 2017; Jones, Johns, & Recchia, 2012 Johns, Dyes,& Jones, 2016) developed a similar construct of semantic diversity, using the Semantic Distinctiveness Model. In this distributional model of semantics, each new contextual encounter with a word compares that context with all of the contexts the word has appeared in previously, stored in the word's memory vector. If a word is encountered in a similar context to what the model predicts based on previous encounters, there is little pressure on the system (the memory vector) to update. If, however, the new context is less predicted by previous experience, this provides an opportunity for the word's representation to be updated in semantic memory. In this way, semantic representations are updated across experience in a way that is sensitive to contextual change. Over time, this results in words that have been experienced in more variable contexts becoming less associated with a particular context. One product of this contextual freedom is that as words become more semantically diverse they should become easier to identify.

Jones et al. (2012) used data from the English Lexicon Project (Balota et al., 2007) to test this hypothesis. They found that words high in semantic diversity enjoyed a processing advantage in both lexical decision and reading aloud. This effect could not be explained by frequency, and semantic diversity accounted for more unique variance than both frequency and document count. This suggests that the content of contextual experience shapes lexical development, not just the quantity of experience. Jones et al. (2012) related their findings to the principle of likely need, in line with rational models of memory (e.g. Adelman, Brown, & Quesada, 2006; Anderson & Milson, 1989). On this view, contextual variation during experience is an indicator that the word will appear in future unknown contexts, leading it "to be more accessible in the lexicon" (Jones et al., 2017, p. 242). While the principle of repetition describes why frequency influences the ease of word identification, the principle of likely need describes why semantic diversity exerts an influence beyond frequency. A different type of explanation starts with the suggestion that variations in semantic diversity reflect gradations in semantic representation with high diversity words having richer semantic representations (Hoffman et al., 2013; Hoffman & Woollams, 2015). High diversity words map to a range of multiple or nuanced meanings, based on the notion that variation in the meaning of a word is an emergent property of variation in the context in which it is used. This is akin to theoretical accounts of the polysemy processing advantage (Rodd, under review; Rodd, Gaskell, & Marslen-Wilson, 2004). Upon each encounter with the polysemous word, its semantic representation would be shaped to its particular sense in that particular context. Over time, this results in patterns of activation that have distinct but overlapping representations. In Rodd et al.’s model, these form a single large attractor basin providing strong input to word form representations. In turn, this input allows for better performance in lexical decision and reading aloud for high semantic diversity words.

So far, we have considered semantic diversity as a variable that shares a positive association with performance in tasks tapping visual word recognition (Hoffman & Woollams, 2015; Jones, Johns, & Recchia, 2012). If the task is changed to one that taps semantic relatedness, however, both of the theoretical accounts outlined above predict a different pattern of association between semantic diversity and behavioural performance. According to Hoffman and Woollams (2015), words that are experienced in varying contexts develop “intrinsically noisy” semantic representations. By their very nature, high diversity words are flexible in meaning and depend on local context for precision in comprehension. A consequence of this for the identification of words presented in isolation might be facilitation, but if the task requires participants to actively reflect on meaning, words high in semantic diversity should be harder to process than less diverse words. Likewise, within the Semantic Distinctiveness Model, more stable semantic representations emerge for words experienced in less variable contexts (Jones et al., 2017). Once again, this leads to the same prediction that less diverse words should be easier to process for meaning than more diverse words – words that have developed a less stable semantic representation as a direct consequence of contextual variation during experience.

In line with this prediction, Hoffman and Woollams (2015) reported opposing effects of semantic diversity on lexical decision vs. semantic judgment: adults were faster to make lexical decisions to words high in semantic diversity, but slower to make semantic decisions about the same words. Complementary evidence comes from Johns, Dye and Jones (2016) word learning experiment with adults. Words induced to be more semantically diverse (via contextual variation during training) were easier to recognize at test but harder to make meaning-based judgments about, relative to words trained in more uniform contexts.

Our discussion has focussed on item-level semantic diversity in connection with how easily adults process words in different tasks. Some previous studies have investigated whether psycholinguistic effects vary as a function of age. For example, Davies and colleagues (Davies, Arnell, Birchenough, Grimmond, & Houlson, 2017) collected lexical decision and naming data from 500 participants ranging in age from 8 to 83 years. They found that both frequency and age of acquisition exhibited a U-shape trajectory with age (and with reading ability) such that the size of effect was initially, then grew larger, and then decreased. This pattern has also been observed in computational models: Monaghan, Chang, Welbourne, and Brysbaert (2017) showed that some exposure was needed for any frequency effect to emerge but that the frequency effect gradually diminished with further input.

Importantly, however, semantic diversity is inherently developmental and none of these investigations have considered semantic diversity. At any point in time, a word's semantic diversity can be thought of as the product of an individual's contextual experiences with that word and the opportunities for learning that are afforded by those experiences, culminating in variations in lexical quality that in turn govern item-level variation in lexical processing (Hsiao & Nation, 2018; Nation, 2017). As semantic diversity is a variable that has its roots in learning and experience, developmental data from children are important – we need to know how semantic diversity changes over time, and how these changes in word knowledge relate to children's language experience. Ultimately, this requires longitudinal studies that follow developmental trajectories as words (and children) grow and semantic networks develop. In the meantime, however, there is some evidence that words acquired earlier in life have more semantic connections than those acquired later (Hills, Maouene, Riordan & Smith, 2010) but few studies have explored the effect of semantic diversity on children's lexical processing. Working with 8-9 year-olds, Rosa, Tapia and Perea (2017) attempted to induce variation in contextual history by placing novel words in passages that varied in theme (high diversity) or maintained the same topic (low diversity). After reading the passages, they assessed how well the children had learned the meaning of the new words using a picture-word matching tasks. Children performed at much higher levels on words learned across diverse encounters, leading Rosa et al. to conclude that contextual diversity supports the learning of word meaning. Joseph and Nation (2018) used a similar approach but found that words learned in semantically diverse passage were not advantaged in subsequent post-tests. These contrasting results mean that the effect of semantic diversity on new word learning by school-age children is not yet clear and more research is needed.

In this paper we take a different approach to capture the influence of semantic diversity on children's processing of language. Rather than induce variation in diversity by training new words in different environments, we extracted semantic diversity estimates for existing words from a large developmental corpus that provides a proxy for children's language experience. We then developed experimental tasks that were suitable for children, and used these to test whether performance was sensitive to item-level variation in contextual history, as indexed by semantic diversity. Relevant to our investigation are data reported by Hsiao and Nation (2018). They found that across three experiments (lexical decision and word naming) with 8-11 year-olds, words high in semantic diversity were easier to recognize than words in low in diversity. This effect could not be explained by frequency, but instead suggests that previous contextual experience with a word influences lexical development and that this is reflected in faster word recognition for words with a history of semantic diversity. If correct, this same contextual experience should result in words that have accrued high semantic diversity being harder to process in a task that demands reflection on meaning, rather than just word recognition. Following experiments with adults (Hoffmann & Woollams, 2015) then, there should be opposing effects of semantic diversity on children's ability to recognize words vs. reflect on their meanings. Our first goal in this paper was to investigate whether children's judgments about word meaning are influenced by semantic diversity and how this compares to adult processing by developing a paradigm suitable for both children and adults. Our second goal was to take a closer look at semantic diversity effects in adults in both semantic judgment and lexical decision.

EXPERIMENT 1

This experiment tested the hypothesis that semantic diversity is negatively associated with the ease of making semantic judgments. Our experiment builds on the approach taken by Hoffman and Woollams so we begin by describing their methodology. They selected 240 words that varied orthogonally in semantic diversity and imageability. The four groups of 60 words (high-high, high-low, low-high and low-low) were matched listwise for frequency and a range of other psycholinguistic variables. These 240 words formed the second word of sequentially presented pair to which adults made a yes/no judgment as to whether the two words were related in meaning. A further 240 words, each selected to be semantically related to one of the original words, formed the first word in a pair; these mirrored the second word in terms of classification as high or low semantic diversity and imageability. Each participant

(N= 25 adults) saw all 240 second words, half preceded by their semantically-related partner to form a related 'yes' trial, and half preceded by an unrelated word to form a 'no' trial.

The task was quite difficult for participants, especially for related pairs where there was an effect of semantic diversity (error rates were above 10% for low semantic diversity pairs and around 20% for the high semantic diversity pairs, see Figure 2, Hoffman & Woollams, 2015). Unrelated pairs were easier to dismiss (error rates below 5%) and performance was not influenced by semantic diversity. Turning to RT, there was a main effect of semantic diversity; as predicted, people were slower (41 msec) to make decisions to high semantic diversity pairs.

Hoffman and Woollams considered the finding that semantic diversity influenced performance on both related and unrelated trials in the RT data as theoretically informative. They argued that a fresh encounter with a semantically diverse word results in a 'blended' or 'averaged' pattern of semantic activation, reflecting previous encounters in multiple and nuanced contexts. Two predictions follow from this. First, as high semantic diversity words elicit a noisy or underspecified pattern of semantic activation, it should be harder to make semantic decisions to words as semantic diversity increases; in contrast, words that have been experienced in more uniform contexts should be easier to process as they elicit more decisive patterns of semantic activation. Second, if these differences between words high vs. low in semantic diversity reflect differences in semantic activation, this should influence performance on both related and unrelated trials. A different type of theoretical account predicts that negative effects of semantic diversity will be seen on related trials only. This derives from the ambiguity literature where homonyms (words with distinct meanings such as *bank*) are slower to process in semantic relatedness tasks. This effect is seen more strongly on related trials, consistent with the idea that response conflict plays a role (e.g. Pexman, Hino, & Lupker, 2004). Hoffman and Woollams favoured a semantic activation type account and this certainly fits with their RT data which showed clear effects of semantic diversity across trial types. Their accuracy data, however, are less straightforward to interpret. Here, only the related trials showed an effect of semantic diversity, suggesting perhaps that response conflict cannot be ruled out.

In summary, although Hoffman and Woollams' findings are in line with the prediction that high semantic diversity is associated with greater difficulty making meaning-based judgments, there are three reasons why further investigation is warranted. First, error rates were high, reducing the number of trials available to enter the RT analyses. Second, although there was an effect of semantic diversity for both accuracy and RT, its nature was different across the two dependent variables. Finally, Hoffmann and Woollams used a categorical design and analysed their data using ANOVA. While semantic diversity effects were reliable across items and participants, treating semantic diversity as a continuous variable and analysing the data within a linear mixed environment has the advantage that it can account for random effects originating from individual participants and items and reduce the chance of Type-1 error.

To address these issues, we developed a cross-modal semantic judgment task, designed to produce high levels of accuracy. This was intended to increase power for RT analyses, and make the task suitable for children (with skilled readers making upwards of 20% errors to high diversity words in the related condition in Hoffman and Woollams' experiment, we anticipated that their task would be too difficult for children). In our task, participants heard a sentence definition of a target word, and at the same time the target word appeared on the computer screen. Their task was to decide as quickly as possible whether the word fitted with the definition or not. We predicted that high semantic diversity target words would be responded to more slowly than low diversity words. As our task contained both related and unrelated trials, we had the opportunity to test whether any effect of semantic diversity is restricted to one trial type or not. We also modelled the effect of word frequency on semantic judgment to test whether any effects of contextual history (i.e., semantic diversity) remained, once frequency had been accounted for. The interaction between semantic diversity and frequency was also of interest to test whether the effect of one variable varies at different levels of the other.

Method

Participants

Sixty-three adults participated in this experiment (42 male; M age = 27.17 SD = 4.45). They were recruited via Prolific (Prolific Academic Ltd.). All participants were native English speakers and were paid for their participation. Fifty-two children participated in this experiment, recruited from primary schools in Oxfordshire. Two children were excluded: one due to experimenter error and one because they failed to complete the task. The final sample comprised 50 children (27 female; M age = 9.8 years, SD = 0.95). Two children were bilingual but had been educated in English only; they had no difficulty with the task (scoring 94% and 96% correct; M for child sample = 93%) and so were included in the analyses. The experiment and all the others in the study were approved by Oxford University's Research Ethics Committee.

Materials

Target words. We selected 160 words that varied in semantic diversity and frequency (see Appendix 1 for all items). The words averaged 6.8 letters in length (range 5-8 letters, SD= 0.87) and were estimated to be familiar to 9-year-old children. Frequency and semantic diversity values are summarized in Table 1. The children's norms were extracted from the Oxford Children's Corpus, with semantic diversity values taken from Hsiao & Nation, 2018;

the adult norms are from the British National Corpus, with semantic diversity values taken from Hoffman et al. (2013).

Insert Table 1 around here

Definitions. A definition was created for each of the 160 words (see Appendix 1), based on the most salient meaning of each word. This was used in the matched trials. Each definition was randomly paired with a different target word, forming the non-matched trials.

Plausibility and predictability of the definitions was checked via two pre-tests, both run using the online survey platform Qualtrics (https://www.qualtrics.com). For plausibility, 86 adults rated how appropriate they thought each definition was for its target word, using a scale of 1 (extremely inappropriate) to 7 (extremely appropriate). Each person rated 60 definitions, 30 matched and 30 non-matched. Matched definitions were rated as highly appropriate (M=6.82, SD =.03; range= 5.5-7) whereas non-matched trials were considered highly inappropriate (M= 1.65, SD= 0.41; range = 1-3). To assess predictability, the 160 definitions were randomly assigned to one of eight lists and a separate group of adult participants (N=14 per list) were asked to supply (in writing) up to five words that first came to mind. These data were used to calculate a predictability score for each target (see Appendix 2 for a worked example) and this was included as a covariate in the analyses. We did not pre-test predictability and predictability for children, but had no reason to suppose it would be different. To preface our findings, the high levels of accuracy shown by both adults and children is consistent with the materials being valid for both age groups.

The written definitions were converted into mp3 audio files using MacOS text-to-speech software, using a female British accent voice (‘Kate’). The software produces realistic speech that is highly intelligible, as confirmed by native speakers of British English (and supported by the high levels of accuracy seen in the experiment). The definitions were distributed into four lists with semantic diversity, frequency, word length and word class matched across lists. The lists were also matched for target word prevalence, a measure based on a mega-scale study that documented the percentage of people who know each word (Brysbaert, Mandera, McCormick, & Keuleers, 2018).

Procedure

For adults, stimulus presentation was via the online platform Gorilla (www.gorilla.sc) and participants completed the experiment on their own computer. For children, the experiment was presented on a Dell Latitude E6400 laptop using E-prime (version 2.0; Schneider, Eschman, & Zuccolotto, 2012) and was completed individually in a quiet area adjacent to their classroom. Participants in both age groups were randomly assigned to one of the four lists. Each person completed 80 trials (40 matched and 40 non-matched) and nobody encountered the same definition twice.

Participants were told that they would see a written word on the screen and at the same time they would hear a brief definition. They were instructed to press the ‘j’ key if they thought the definition matched the word and the ‘f’ key if they thought the definition didn’t match the word (the keys were marked with stickers for the children). Participants were asked to respond as quickly as possible whilst trying not to make any mistakes and the instructions emphasized that they could make their response at any time, even while the definition was still playing. Children were provided with a set of headphones and adults were asked to put on headphones, or make sure they were in a quiet space. Before starting, participants placed a finger of each hand onto the response keys. The task began with four practice trials. The experiment was split into two blocks for the adults and four blocks for the children, each beginning with two dummy trials. The order of the blocks was counterbalanced between participants. At the start of each trial, a central fixation cross was displayed for 500ms before the word appeared and the definition played. The word remained on the screen until a response was made. Trials were presented in a random order and the experiment lasted approximately 15 minutes for both adults and children.

Results

Our results are presented in three parts. We first describe data pre-processing and general analytic approach before presenting the results for the adults and then the children. We analysed the data for adults and children separately because semantic diversity and frequency values differed for the two groups.

The data were checked for participant and item outliers. For adults, no items were excluded but four participants were excluded due to low accuracy (below 50% correct). The final sample comprised 59 adults (38 female; M age= 27.07, SD= 4.42). For children, four items generated low accuracy scores, suggesting they were too difficult for some children with accuracy (less than 60%). These items were removed. One child who performed at chance on the non-matched trials was also excluded. The final sample comprised 49 children (27 female; M age= 9.84 years, SD= 0.91). Individual trials were removed if its RT fell more than 2.5 (for adults) or 3 (for children) standard deviations away from a participant's overall mean RT. Most of the data was retained for both adults (92%) and children (94%).

We analysed the data using linear mixed effects (LME) models with maximal model structure (see Barr, Levy, Scheepers, & Tily, 2013), computed using the lme4 package (version 1.1-15; Bates, Mächler, Bolker, & Walker, 2015) in R (version 3.4.1; R Core Team, 2018). By-subject random intercept and slopes were included for each fixed main effect and interaction and by-item random intercept and slope of the matchedness effect were included.

Models which failed to converge with maximal structure were simplified by removing random interactions and effects explaining the least amount of variance. Random slopes included for each model were specified in the corresponding result sections. For models with a continuous outcome variable (RT), statistical significance was determined based on the criteria of t>2. For models with a binary outcome variable (accuracy), mixed effects logistic regression models were computed and the cut-off for significance was p<.05. The fixed factors included in the analyses were the main effects of semantic diversity, frequency (both continuous variables), matchedness (matched vs. non-matched trials), the two-way interactions of semantic diversity*frequency, semantic diversity*matchedness and frequency*matchedness. Matchedness was specified as a categorical variable using effects coding (0.5/-0.5 for matched vs. non-matched, respectively), such that the intercept corresponded to the grand mean and the fixed effects corresponded to the main effect of the fixed factors. We also included two covariates in the RT analyses, namely definition length and predictability score. We had planned to include predictability as a covariate in the analysis of accuracy data. As reported below, however, accuracy was high and the data not analysed further. All continuous variables were centred and scaled.

Note that the data for all four experiments is available in the supplementary materials, along with the analysis scripts. The materials are also available on the Open Science Framework website (https://osf.io/7hz5p/) (Hsiao, Bird, Norris, Pagán & Nation, 2019).

(i) Adults

Accuracy was at ceiling (M= 98%) and not analysed further. For RTs to correct trials (overall M= 2077 msec, SD=733), the converged model included a random slope of definition length for participants and matchedness for items. Both covariates were significant, with slower responses associated with longer definitions (b = 249.84, SE = 21.49, t = 11.63) and less predictable definitions (b = -82.70, SE = 16.27, t = -5.08). The main effect of semantic diversity was significant (b = 51.75, SE = 18.40, t = 2.81): RTs were slower for high diversity words. Semantic diversity interacted with matchedness (b = 129.50 SE = 30.30, t = 4.27), shown in Figure 1. Separate models were fitted that included matchedness with different reference levels confirmed that there was no effect of semantic diversity on the matched trials (t= -0.56) but a robust effect on the non-matched trials (b = 116.50, SE = 24.39, t = 4.78).

Neither frequency (t = -0.04) nor matchedness (t = -0.70) were significant main effects. There was however a significant frequency*matchedness interaction (b = -69.89, SE = 31.04, t = -2.25), shown in Figure 2. Higher frequency words tended to elicit faster RTs in matched trials but slower RTs in non-matched trials. However, two models with contrasting reference levels of matchedness confirmed an absence of a frequency effect on both the matched trials (t=-1.45) and the non-matched trials (t= 1.43). Finally, there was no interaction between semantic diversity*frequency (t = -0.20).

Insert Figure 1 and 2 around here

(ii) Children

Accuracy was high (M=93%, SD=.26) and a model with the full set of fixed factors failed to converge, consistent with performance being close to ceiling and therefore not warranting further analysis.

Turning to RT to correct trials (overall M= 2520 msec, SD= 996), the converged model included random slopes of definition length for participants and matchedness for items. Children were slower to respond to longer definitions (b = 306.06, SE = 25.34, t = 12.07), and those that were less predictable (b = -97.31, SE = 21.07, t = -4.62). The main effect of semantic diversity was significant (b = 53.71, SE = 25.74, t = 2.09), with slower

RTs to words higher in semantic diversity, as shown in Figure 3. There was also a main effect of frequency (b = -59.38, SE = 25.96, t = -2.29) with faster RTs to higher frequency words, shown in Figure 4. Children were also faster to respond to matched trials (M= 2397, SD= 983) than non-matched trials (M= 2640, SD= 994; b = -214.11, SE = 38.26, t = -5.60). Only one interaction was significant: semantic diversity * frequency (b = 76.26, SE = 23.06, t = 3.31). Holding frequency constant, increases in semantic diversity were associated with increase in RT. No other interactions were significant: semantic diversity*matchedness (t = -0.03), frequency*matchedness (t = 0.07).

Insert Figure 3 and 4 around here

Discussion

This experiment investigated whether semantic diversity and frequency influenced how easily children and adults judged the meaning of a word by asking them to decide whether it fitted with a definition. Its primary aim was to test the hypothesis that making decisions about meaning is harder for words that are high in semantic diversity. Overall, our findings are consistent with this hypothesis. Children were slower to respond as semantic diversity increased and this effect was seen for both matched and non-matched trials. Adults also showed a negative effect of semantic diversity, mainly driven by the non-matched trials.

Our findings from children were similar to the results of Hoffman and Woollams' relatedness judgment task performed by adults. There was an overall slowdown for high semantic diversity words across 'yes' and 'no' trials. In contrast, the adults in our experiment only showed an effect of semantic diversity for the non-matched trials. This might be attributable to differences in task demands. In the relatedness task used by Hoffman and

Woollams, participants needed to decide whether two words (with similar levels of semantic diversity) were related; presumably, this entails generating and comparing two sets of meanings. In our experiment, participants only needed to come up with one set of meanings for a single target word and to consider whether it matched the definition provided. This difference in experimental design might have given Hoffman and Woollams' greater power to detect an effect, even in the 'yes' trials. Consistent with this, another difference between the two experiments is that adults made few errors on our task (accuracy 98%) whereas the adults in Hoffmann and Woollams' experiment found that the task quite difficult, especially for the high semantic diversity items (accuracy approximately 80% on a forced choice task).

The children in our experiment were much slower overall than the adults (2520ms and 2077ms, respectively), and slower still on the non-matched trials. Arguably, this more effortful processing led to effects of semantic diversity emerging for both 'yes' and 'no' trials for children.

Children's RTs were also sensitive to word frequency, with faster responses to high frequency words across both matched and non-matched trials. Frequency interacted with semantic diversity for the children, such that the advantage of frequency was increasingly eliminated as semantic diversity increased. There was no main effect of frequency for adults, but it did interact with matchedness, with slower responding to high frequency words on matched trials yet faster responding to the same words on non-matched trials. The reasons for this are not clear, but we note that the frequency effect was not reliable at either level of matchedness in follow-up analyses on the interaction.

EXPERIMENT 2

Experiment 1 found that words high in semantic diversity are more difficult to process when the task requires active reflection on word meaning. This diversity disadvantage contrasts with the previously reported advantage in tasks tapping word identification, both for adults (e.g., Hoffman & Woollams, 2015) and children (Hsiao & Nation, 2018). It is however difficult to compare across experiments and across tasks, not least because different tasks use different sets of items. Hoffman and Woollams (2015) is the only study to directly compare word identification (lexical decision) and semantic judgment for the same items. While they concluded that there are opposing effects of semantic diversity on lexical decision and semantic judgment, their lexical decision data are not robust: there was a significant effect across participants, but the effect was not reliable across items. Clearly, this finding needs to be strengthened and replicated before strong conclusions can be made about the opposing effects of a semantic diversity across different types of task for the same items. In this spirit, Experiment 2 used the same words as Experiment 1 but in a lexical decision task. We predicted that this should cause the semantic diversity effect to flip direction: children and adults should be faster at responding ‘yes’ to words higher in semantic diversity.

Method

Participants

Two new groups of adults and children participated in this experiment. Sixty-five adults were recruited via Prolific (as per Experiment 1). Three quit before the experiment had started, leaving a final sample of 62 (32 male, 29 female, 1 other; M= 28.34 years, SD= 4.53). All participants were native English speakers and were paid for participation.

Forty-six children participated in this experiment. Five children were excluded: three who were too tired to do the task, one child who struggled to read and one due to response box error. This left 41 children in the final sample (M= 9.73 years, SD= 0.45). Eight children had some experience with a language other than English, but all had been educated in English only, and none had any difficulty with the task (M= 81%, SD= 0.07; rest of sample M= 85%, SD=0.34, range=63%-98%); they were therefore retained for analysis.

Materials and Procedure

The 160 target words from Experiment 1 served as word stimuli and 160 nonwords were generated using Wuggy (Keuleers & Brysbaert, 2010), matched to the words for length and number of syllables.

The adults completed all 320 trials (160 words and 160 nonword) on their own computer, using the online platform Gorilla (www.gorilla.sc). Written instructions explained that they would see a word on the screen and they had to decide as quickly as possible if it was ‘real’ or ‘made-up’, pressing the ‘p’ key for ‘yes’ and ‘q’ for ‘no’. Following 20 practice trials (with feedback), the 320 items were presented in a random order. Following a fixation cross of 500 ms, the word remained on the screen until a response was made.

The procedure for children was largely similar, although like Experiment 1, children were tested in a quiet area adjacent to their classroom. The experiment was presented on a Dell Latitude E6400 laptop using E-prime 2.0 software (Schneider et al., 2012). Children responded via a response box, pressing the right button for ‘yes’ and the left button for ‘no’. The items were split into two lists, each containing 80 words and 80 nonwords. Children were randomly assigned to one list. Items were presented in a random order (each following a 1000ms fixation cross) with a break halfway through.

Results

The data were checked for participant and item outliers. One adult was removed as their accuracy was very poor (40%), making the final sample N=61 (29 female; M= 28.39 years, SD = 4.54). One item was removed as it generated a low accuracy rate below 60%. All children were retained, but 14 items were removed from the children's data due to accuracy below 60%. Individual trials were removed if responses were exceptionally fast (<250ms) or slow (>2500 for adults and >5000 for children). Following this, a trial was removed if its RT was more than 2.5 (for adults) or 3 (for children) standard deviations away from a participant's overall mean RT. A high proportion of the RT data remained for the adults (94%) and the children (79%). Accuracy on word trials and RT for correct word trials served as dependent variables. We used the data modelling approach described in Experiment 1. The fixed factors were semantic diversity, frequency and the interaction between semantic diversity and frequency.

(i) Adults

Accuracy was high, at 95% (SD= 0.22). The model with frequency random slope for participants and random intercept for items converged. People responded more accurately to higher frequency words (b = 0.96, SE = 0.13, *p* < .001) but there was no effect of semantic diversity (*p* = .26), or frequency*semantic diversity interaction (*p* = .32). The same random structure was included in the RT model. Mean RT was 645ms (SD=193). Adults were faster to respond to higher frequency words (b = -31.36, SE = 3.86, t = -8.13) but neither the main effect of semantic diversity (t = 0.68) nor its interaction with frequency (t = 1.82) were significant.

(ii) Children

The accuracy model converged with only intercepts for participants and items. Overall accuracy was high, M= 89% (SD= 32) and was influenced by semantic diversity (b = 0.35, SE = 0.12, t = 2.80, *p* = .005) and frequency (b = 0.65, SE = 0.12, t = 5.34, *p* < .0001).

Children were more accurate to respond to words with higher semantic diversity and frequency. The interaction was not significant (*p* = .15). The RT model contained the random slope of frequency for participants and intercept for items. Mean RT was 870ms (SD= 379). Children were faster to respond to words high in semantic diversity (b = -28.12, SE = 10.21, t = -2.76) and high in frequency (b = -55.53, SE = 11.95, t = -4.65); there was no interaction between semantic diversity and frequency (t = -0.82).

Discussion

Using the same words as in Experiment 1, this experiment investigated whether semantic diversity and frequency influences lexical decision. Replicating previous work (Hsiao & Nation, 2018), children showed clear effects of both frequency and diversity: they responded more quickly to words as their frequency and semantic diversity increased. In adults however, the pattern of results was different: there was an effect of frequency, but no effect of semantic diversity. This contrasts not only with our own data from children, but also previous work showing facilitative effects of semantic diversity on lexical decision in adults (Hoffman & Woollams, 2015; and using a similar metric, Jones et al., 2012).

There are a number of methodological differences between our experiment and that of Hoffman and Woollams. First, our experiment was remote, with adult participants completing it via an online platform whereas Hoffman and Woollams collected their data in the lab. Second, we investigated semantic diversity and frequency whereas Hoffman and Woollams focussed on semantic diversity and imageability. Disentangling the effects of semantic diversity and frequency is difficult given the natural correlation between the two variables; what might follow is that a larger (or more varied) set of words is needed to detect an effect in lexical decision when frequency is also included in the model. The effect size for semantic diversity in Experiment 2 (adult data) was small at 0.05; by comparison, the effect size for the adults in semantic judgment in Experiment 1 was 0.1. For Experiment 2 to have a power of 0.8 with such a small effect size, and with 160 words, we would need 146 participants, or an infinite number of items if the number of participants remained at its current level of N=61 (Brysbaert & Stevens, 2018; Westfall et al., 2014). In contrast, for Experiment 1, a power analysis showed that with our current item set of 160 items, a sample size of 50 participants is required to detect an effect size of 0.8; 63 people participated in Experiment 1 and we could have reached power of 0.8 with 136 items. The results of these power analyses suggest that Experiment 1 was sufficiently powered to detect an effect of semantic diversity, Experiment 2 was underpowered. A final difference between our experiment and that of Hoffman and Woollams is that our experiment took a continuous approach to design and analysis, and we used linear mixed effects taking both participant and item random effects into account. They used an orthogonal design, manipulating semantic diversity and imageability across four lists with frequency matched across lists and then testing for effects using analysis of variance. It is also possible that the range of semantic diversity values varied across the two item sets. To establish whether there is an effect of semantic diversity in lexical decision in adults, Experiments 3 and 4 systematically addressed these methodological differences.

EXPERIMENT 3

This experiment used the same items as Hoffman and Woollams (2015, Experiment 1) but presented them in an online experiment, following our procedure in Experiment 2. We also analysed the data in two ways: analysis of variance, repeating the categorical approach of Hoffman and Woollams, and continuously in a linear mixed effect environment.

Method

Participants

Twenty-three people participated in Hoffman and Woollams' experiment. We recruited 38 adults via Prolific (https://prolific.ac). Six quit before the start of the experiment leaving a final sample of N=32 (16 male; M= 27.84 years, SD = 4.58). All participants were native English speakers and were paid for participation.

Materials and Procedure

As noted earlier, Hoffman and Woollams selected four lists of 60 words that varied orthogonally in semantic diversity and imageability and frequency was matched list-wise. We used the same 240 words and 240 nonwords but our experiment was run online using Gorilla, as per our previous experiment. Following 20 practice trials, adults responded to all 480 items, presented in a random order.

Results

The data were cleaned in the same way as Experiment 2. One item had a low accuracy rate (55%) and was removed. One person appeared to have mixed up the response keys (accuracy = 9%); their data were excluded, leaving a final sample of N=31. Trials that were very fast (<250ms) or very slow (>2500ms) were removed, as were trials that were more than 2.5 SDs away from an individual's mean RT. This resulted in 93% of data retained for analysis. Note that the pattern of results was identical when we followed the same data cleaning procedures reported by Hoffman and Woollams.

Our data are summarised in Figure 5. Mean accuracy was 93% (SD= 0.26). Accuracy was not specified by Hoffman and Woollams but visual inspection of Figure 1 in their paper indicates a similarly high level of performance.

Insert Figure 5 around here

We first analysed our data using a 2 (high vs. low semantic diversity) x 2 (high vs. low imageability) repeated measures analysis of variance on both accuracy and RT; our findings are summarised alongside those reported by Hoffman and Woollams in Table 2. For accuracy, both experiments produced main effects of semantic diversity and imageability, with fewer errors to high imageability and high semantic diversity words. In addition, we saw a reliable semantic diversity*imageability interaction, but only in the by-participants analysis. Turning to RT, we replicated the semantic diversity effect seen by Hoffman and Woollams with faster responses to more diverse words. Our analysis also revealed an effect of imageability (by-participants only).

To check the robustness of these findings, and to compare them with those reported in Experiment 2, we fitted linear mixed effects models to the data from Experiment 3, treating both semantic diversity and imageability as continuous variables (with variables centred and scaled). Both the accuracy and RT models converged with the random structure with intercepts. Once again, the main effect of semantic diversity was significant, both for accuracy (b=.18, SE=.07, z=2.47, p=.01) and RT (b=-5.35, SE=2.58, t=-2.08). Imageability was also significant for both accuracy (b=.19, SE=.08, z=2.45, p=.01) and RT (b=-10.08, SE=2.57, t=-3.93). There was no interaction between the two variables (accuracy: p=.51; RT: t=.05).

Insert Table 2 around here

Discussion

Our aim in Experiment 3 was to ask again whether semantic diversity influences lexical decision in adults. Experiment 2 found a faciliatory effect of semantic diversity in children's lexical decision, replicating previous findings (Hsiao & Nation, 2018). In adults however, the effect was not reliable, failing to replicate findings reported by Hoffman and Woollams. It is also worth noting that semantic diversity was only significant by-participants in their experiment, not by items. The results of Experiment 3 are clear in showing that people are faster to respond to words high in semantic diversity. This main effect emerged in the analysis of variance that treated semantic diversity and imageability as categorical variables; it was also present when we analysed the data continuously using linear mixed models.

The semantic diversity effect does seem reliable, at least in the stimulus set used here. Why then was there no effect on Experiment 2? We can rule out the likelihood of it being due to its online nature, given Experiment 3 was also conducted online. One possibility is that the effect is limited to item sets that have particular characteristics, or that the range of semantic diversity values in the two experiments differed: Experiment 3 did contain some lower diversity words relative to Experiment 2 (Experiment 2 range 1.42-2.16, SD= 0.28; Experiment 3 range 0.61-2.17, SD= 0.25). It might be that frequency plays an important role too. In Experiment 2, semantic diversity was assessed directly alongside frequency. In Experiment 3 however, and following Hoffman and Woollams, semantic diversity was compared against imageability and although frequency was matched across stimulus lists, the effect of frequency or its interaction with semantic diversity was not tested. This is an important point as frequency and semantic diversity are naturally associated. The correlation between Hoffman et al.'s (2013) metric of semantic diversity and frequency was .49 and Johns, Gruenenfelder, Pisoni, & Jones (2012) report correlations ranging from .46 to .95, depending on the corpus used to derive their semantic diversity values. It remains plausible, therefore, that some of the variance associated with lexical decision speed in Experiment 3 is associated with frequency rather than imageability and that this is the reason why the results vary between Experiments 2 and 3. This possibility is investigated in Experiment 4.

EXPERIMENT 4

In this final experiment, we used data from two megastudies to test at scale the relationship between item-level semantic diversity and lexical decision performance. This avoided the stimulus selection problems inherent in Experiments 2 and 3, and allowed us to control both frequency and imageability in the same analyses. As discussed above, frequency and semantic diversity are naturally correlated. Semantic diversity also correlates with imageability, r= -.48 (Hoffman et al., 2013). Note this is a negative relationship: words high in semantic diversity tend to be low in imageability.

The English Lexicon Project (ELP; Balota et al., 2007) contains data from 815 participants on nearly 40,000 words. Of these, 2679 had both semantic diversity values (Hoffman et al., 2013) and imageability ratings (Cortese & Fugett, 2004). The British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012) contains data from 78 participants on around 14,300 words. Semantic diversity and imageability values were available for 2705 of these words. The total number of observations was over 90,000 and 100,000 for the ELP and BLP datasets respectively.

For both the ELP and the BLP data, we fitted linear mixed effects models that controlled for both item-level (only random intercepts as there were no participant-level variables) and participant-level random effects (see Table 3 for the random slopes of the final models that converged). We included zipf frequency (van Heuven, Mandera, Keuleers, & Brysbaert, 2014) based on the British National Corpus, along with semantic diversity

(Hoffman et al., 2013) and imageability (Cortese & Fugett, 2004), and the interactions between semantic diversity and the other two variables: semantic diversity*frequency, and semantic diversity*imageability. All variables were centred and scaled. The results are summarized in Table 3.

We observed large main effects for semantic diversity, frequency and imageability across both datasets, for accuracy as well as RT. Words were easier to process when they were more semantically diverse, more frequent and more imageable. The semantic diversity*frequency interaction was robust, such that among low diversity words, words that were also low in frequency were the most difficult to process. The interaction between semantic diversity and imageability was significant in the ELP data but marginal in the BLP: among lower diversity words, words lower in imageability were harder. Taken together, these analyses show clear facilitative effects of semantic diversity on lexical decision, independent of both frequency and imageability. This suggests that the varying effects seen across earlier experiments are likely due to restricted sets of items, and small sample size.

Insert Table 3 around here

General Discussion

Our goal in this paper was to investigate how a word's contextual history, as indexed by its semantic diversity, influences lexical processing in children and adults. Our findings are clear in finding an effect of semantic diversity, consistent with the idea that the contextual nature of a person's previous experience with a word influences how that word is represented and processed. However, the behavioural manifestation of semantic diversity varied according to the nature and demands of the task.

Hoffman and Woollams (2015) first reported opposing effects of semantic diversity across lexical decision and semantic judgment in adults. High diversity was associated with faster lexical decision but slower semantic judgment. Importantly, however, the effect in lexical decision was not statistically reliable across items, casting some doubt on the reliability and generalisability of their findings. While their semantic judgment data showed similar effects across items and participants, it patterned differently across RT and accuracy. For RT, semantic diversity was negatively associated with performance in both related and non-related trials, whereas for accuracy, its negative effect was restricted to related trials. We thus sought to test the robustness of Hoffman and Woollams' findings, and extend them to children.

Experiment 1 established the utility of the cross-modal definition task. Error rates were low, even for children, and RTs were clearly sensitive to item-level properties. The task is suitable for a broad range of words, unlike tasks like animacy or size judgment that are restricted to nouns. Our findings from Experiment 1 are consistent with the hypothesis that meaning-related decisions are harder to make for words that are high in semantic diversity. Children showed a clear effect of semantic diversity, with slower responses to high diversity words across both matched and non-matched trials. For adults, although there was a main effect of semantic diversity, this interacted with trial type such that its effect was only evident on non-matched trials.

Experiment 1 established a negative association between semantic diversity and meaning judgments. Experiment 2 tested whether the same items would show an opposing effect of semantic diversity on lexical decision performance. For children, there was a clear facilitative effect with faster responses to words higher in semantic diversity, replicating previous findings (Hsiao & Nation, 2018). For adults, however, there was no effect of semantic diversity. This was not as predicted, so to investigate further, Experiment 3 used the same item set as Hoffman and Woollams. Here, we did observe a positive effect of semantic diversity on lexical decision performance, replicating their findings. The most obvious explanation for the different pattern of results for adults across Experiments 2 and 3 is variation within small sets of items and associated lack of statistical power. In line with this suggestion, Experiment 4 found clear facilitative effects of semantic diversity on lexical decision in both the ELP and the BLP mega datasets, even when frequency and imageability were both controlled. Taken together, we are confident in concluding that variations in semantic diversity are positively associated with ease of word identification, as tapped by lexical decision.

Having confirmed the opposing effects of semantic diversity in lexical decision and semantic judgment, we now turn to consider what semantic diversity is, and why it influences lexical processing. It is clear that its effects are separable from those of frequency, meaning that a theoretical account based on the principle of repetition is not adequate. Experiments 3 and 4 (and Hoffman & Woollams, 2015) demonstrate that its effects are also separable from imageability, a variable considered to tap semantic richness. Indeed, high diversity words tend to be lower in imageability (r= - .48; Hoffman et al., 2013), yet both semantic diversity and imageability are positively associated with lexical decision suggesting that semantic diversity does not reflect a construct such as ‘semantic richness’ in the same way as traditional semantic variables such as imageability and concreteness (see also Pexman, Heard, Lloyd, & Yap, 2017 who reported that semantic diversity facilitated semantic decisions for abstract words, but showed the reverse effect for concrete words). Clearly, semantic diversity is not the same as frequency, or semantic richness. Instead, our findings suggest that something about the contextual nature of previous experience with a word, or something correlated with this, drives item-level differences in lexical processing. The next challenge is to capture item-level developmental trajectories for words as they emerge through contextual experience. This highlights the need for more large-scale lifespan data across tasks that will allow a range of psycholinguistic variables (derived from developmentally-informed corpora) to be modelled in the same datasets.

Ultimately, a different type of experimental design is needed to inform more precisely what semantic diversity is – a design that directly manipulates and induces variation in semantic diversity during language learning and measures the consequence of this on subsequent lexical processing as words and children develop. In the meantime, however, and as described in the Introduction, two theoretical accounts might be relevant to understanding the influence of semantic diversity, both of which view word meaning as graded and varying as a function of context (Rodd et al., 2004). According to Hoffman and Woollams (2015), when a word is experienced, the semantic activation it generates reflects a blend state or composite of previous semantic associations, representing its contextual history. For words with a rich and varied contextual history, including words high in semantic diversity, this initial blend state will be activated quickly and while sufficient to drive a lexical decision response, a word's blend state needs to settle and resolve before a semantic decision response can be made. The behavioural observations of faster lexical decisions but slower semantic decisions for words high in semantic diversity sit comfortably with this account. The data also fit within the Semantic Distinctiveness Model framework, as discussed by Jones, Johns and colleagues. This sees semantic representations developing as an emergent property of each episodic encounter with a word; those words experienced in more varied contexts have greater opportunity to update than words experienced in similar episodes, leading to processing differences over time. The two theories differ in how they represent the semantic space of a word. Hoffman and colleagues describe an instance of aggregation at the time a word is processed that reflects its previous contextual experience (i.e. the blend state that is activated), whereas Jones, Johns and colleagues describe a word's representation being updated each time the word is experienced, with opportunities for updating being greater as new contexts are experienced.

Common to both theoretical accounts is the notion of continuous and graded semantic representations characterising all words, not just words that are categorised as ambiguous or polysemous. These gradations of meaning are captured by semantic diversity, and in turn, this interacts with task demands to influence behavioural performance. While these models measure semantic variability associated with contextual change, future work should investigate how contexts associated with a given word cluster differentially to form distinctive meanings homonym such as *bark* (Jamieson, Avery, Johns, & Jones, 2018, Rodd et al., 2004). Our data cannot distinguish between the two types of theoretical account, but they nevertheless add to the evidence base in three important ways. Our findings show that a word's contextual history influences lexical processing, even when a word is experienced in isolation and out of context. They demonstrate opposing effects of the same variable on tasks that tap word identification and those that require reflection on meaning, and extend these observations to children. Finally, they add to the work of others (e.g. Keuleers & Balota, 2015; and see Schröter and Schroeder, 2017, for lexical decision developmental megadata from children and adults reading German) in demonstrating the utility of using corpus-based statistics in combination with secondary analysis of mega-datasets to investigate item-level effects, complementing traditional categorical designs that control and manipulate particular lexical properties.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity not frequency determines word naming and lexical decision times. *Psychological Science*, *17*(9), 814–823.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719. https://doi.org/10.1037/0033-295X.96.4.703

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1077-9

Chang, Y. N., Monaghan, P., & Welbourne, S. (2019). A computational model of reading across development: Effects of literacy onset on language processing. *Journal of Memory and Language*, *108*(September 2018), 104025. https://doi.org/10.1016/j.jml.2019.05.003

Chateau, D., & Jared, D. (2000). Exposure to print and word recognition processes. *Memory & Cognition*, *28*(1), 143–153. https://doi.org/10.3758/BF03211582

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words.

Behavior Research Methods, Instruments, & Computers, 36(3), 384–387. https://doi.org/10.3758/BF03195585

Davies, R., Arnell, R., Birchenough, J. M. H., Grimmond, D., & Houlson, S. (2017). Reading through the lifespan: Individual differences in psycholinguistic effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1298–1338. https://doi.org/10.1037/xlm0000366

Hersch, J., & Andrews, S. (2012). Lexical Quality and Reading Skill: Bottom-Up and Top-Down Contributions to Sentence Processing. *Scientific Studies of Reading*, *16*(3), 240–262. https://doi.org/10.1080/10888438.2011.564244

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

Hoffman, P., & Woollams, A. M. (2015). Opposing effects of semantic diversity in lexical and semantic relatedness decisions. *Journal of Experimental Psychology. Human Perception and Performance*, *41*(2), 385–402. https://doi.org/10.1037/a0038995

Hsiao, Y., Bird, M., Norris, H., Pagán, A., & Nation, K. (2019, October 11). The influence of item-level contextual history on lexical and semantic judgements by children and adults. Retrieved from osf.io/7hz5p.

Hsiao, Y., & Nation, K. (2018). Semantic diversity, frequency and the development of lexical quality in children’s word reading. *Journal of Memory and Language*, *103*(February), 114–126. https://doi.org/10.1016/j.jml.2018.08.005

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An Instance Theory of Semantic Memory. *Computational Brain & Behavior*, *1*(2), 119–136. https://doi.org/10.1007/s42113-018-0008-2

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Johns, B. T., Gruenenfelder, T. M., Pisoni, D. B., & Jones, M. N. (2012). Effects of word frequency, contextual diversity, and semantic distinctiveness on spoken word recognition. *The Journal of the Acoustical Society of America*, *132*(2), EL74–EL80. https://doi.org/10.1121/1.4731641

Jones, M. N., Dye, M., & Johns, B. T. (2017). Context as an Organizing Principle of the Lexicon. In *Psychology of Learning and Motivation* (Vol. 67, pp. 239–283). https://doi.org/10.1016/bs.plm.2017.03.008

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology*, *66*(2 SPL.ISSUE), 115–124. https://doi.org/10.1037/a0026727

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *The Quarterly Journal of Experimental Psychology*, *68*(8), 1457–1468. https://doi.org/10.1080/17470218.2015.1051065

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods*, *44*(1), 287–304. https://doi.org/10.3758/s13428-011-0118-4

Monaghan, P., Chang, Y. N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21. https://doi.org/10.1016/j.jml.2016.08.003

Nation, K. (2017). Nurturing a lexical legacy: reading experience is critical for the development of word reading skill. *Npj Science of Learning*, *2*(1), 3. https://doi.org/10.1038/s41539-017-0004-7

Pexman, P. M., Heard, A., Lloyd, E., & Yap, M. J. (2017). The Calgary semantic decision project: concrete/abstract decision data for 10,000 English words. *Behavior Research Methods*, *49*(2), 407–417. https://doi.org/10.3758/s13428-016-0720-6

Pexman, P. M., Hino, Y., & Lupker, S. J. (2004). Semantic Ambiguity and the Process of Generating Meaning From Print. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1252–1270. https://doi.org/10.1037/0278-7393.30.6.1252

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria. Retrieved from https://www.r-project.org/

Rodd, J. (n.d.). Settling into Semantic Space : An Ambiguity-Focused Account of Word-Meaning Access Jennifer Rodd Department of Experimental Psychology , University College London Word Meaning Access : The Challenge of Lexical Ambiguity.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. https://doi.org/10.1016/j.cogsci.2003.08.002

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). E-Prime User’s Guide. Pittsburgh: Psychology Software Tools, Inc.

Schröter, P., & Schroeder, S. (2017). The Developmental Lexicon Project: A behavioral database to investigate visual word recognition across the lifespan. *Behavior Research Methods*, *49*(6), 2183–2203. https://doi.org/10.3758/s13428-016-0851-9

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: a new and improved word frequency database for British English. *Quarterly Journal of*

Experimental Psychology (2006), *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Table 1. Frequency and semantic diversity values for the 160 target words used in the Experiment 1

	Semantic Diversity			Frequency (zipf)[1]		
	M	SD	Range	M	SD	Range
Adults	1.61	.28	1.42-2.16	4.06	.77	2.78-5.59
Children	2.02	.17	1.28-3.64	4.60	.64	3.64-5.85

[1] Zipf frequency was calculated using the following formula, as recommended by van Heuven et al. (2014): log10(fpmw) + 3.

Table 2. Summary of ANOVA results from Hoffman & Woollams and Experiment 3

	Hoffman & Woollams[1]		Experiment 3[2]	
	F	*p*	F	*p*
Accuracy				
Semantic Diversity				
F1	5.93	.023*	16.17	.0004*
F2	1.28	.26	6.73	.01*
Imageability				
F1	13.5	.001*	9.37	.005*
F2	4.74	.03*	1.91	.17
Sem Diversity x Image				
F1	2.74	.11	10.98	.002*
F2	.84	.36	2.45	.11
RT				
Semantic Diversity				
F1	9.99	.005*	8.86	.006*
F2	2.80	.096	7.66	.006*
Imageability				
F1	n.s. (stats not reported)		16.91	.0003*
F2	n.s. (stats not reported)		7.81	.006*
Sem Diversity x Image				
F1	n.s. (stats not reported)		1.74	.20
F2	n.s. (stats not reported)		1.06	.30

Notes: [1]For Hoffman & Woollams, F1: df= 1, 22; F2: df= 1, 235; [2]For Experiment 3, F1: df= 1,30; F2: df= 1,235. *p< .05.

Table 3. Results of linear mixed effect models on data from ELP and BLP megastudies, Experiment 4

	ELP (N= 2679 words)				BLP (N= 2705 words)			
	b	SE	t	*p*	b	SE	t	*p*
Accuracy								
Semantic diversity	0.21	0.03	6.28	***	0.20	0.04	5.15	***
Frequency	0.84	0.04	21.84	***	0.92	0.05	19.09	***
Imageability	0.84	0.03	27.07	***	0.77	0.04	21.50	***
SD*Frequency	-0.12	0.02	-5.19	***	-0.21	0.03	-7.41	***
SD*Imageability	-0.06	0.03	-2.41	*	0.02	0.03	0.87	0.38
RT								
Semantic diversity	-7.01	1.57	-4.44	*	-4.92	1.13	-4.34	*
Frequency	-46.53	1.81	-25.70	*	-37.08	1.67	-22.21	*
Imageability	-28.13	1.28	-21.94	*	-20.96	1.21	-17.28	*
SD*Frequency	9.62	1.12	8.63	*	7.59	0.76	9.95	*
SD*Imageability	3.43	1.21	2.84	*	1.43	0.82	1.75	-

Notes. For accuracy, **p< .01 and **p< .001. For RT, all results with t > 2 are considered significant and marked with *. No separate levels of significance are distinguished. When t < 2, no p-value is provided.

The random slopes included for participants were as follows. For accuracy: ELP: semantic diversity, frequency, imageability. BLP: semantic diversity, frequency, imageability, semantic diversity*frequency. For RT: ELP: frequency; BLP: semantic diversity, frequency, imageability

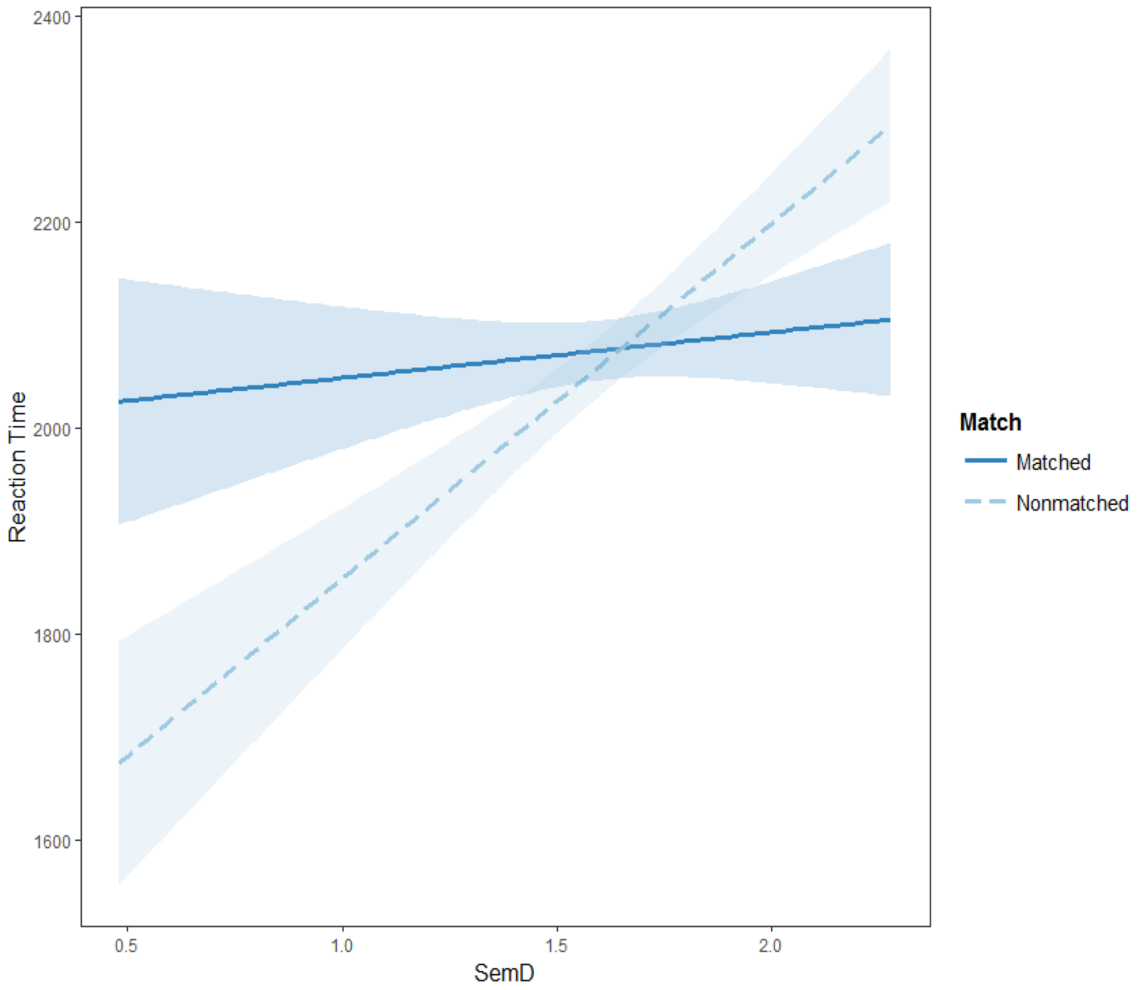

Figure 1. The interaction between semantic diversity and matchedness in adult RT data (ms) from Experiment 1.

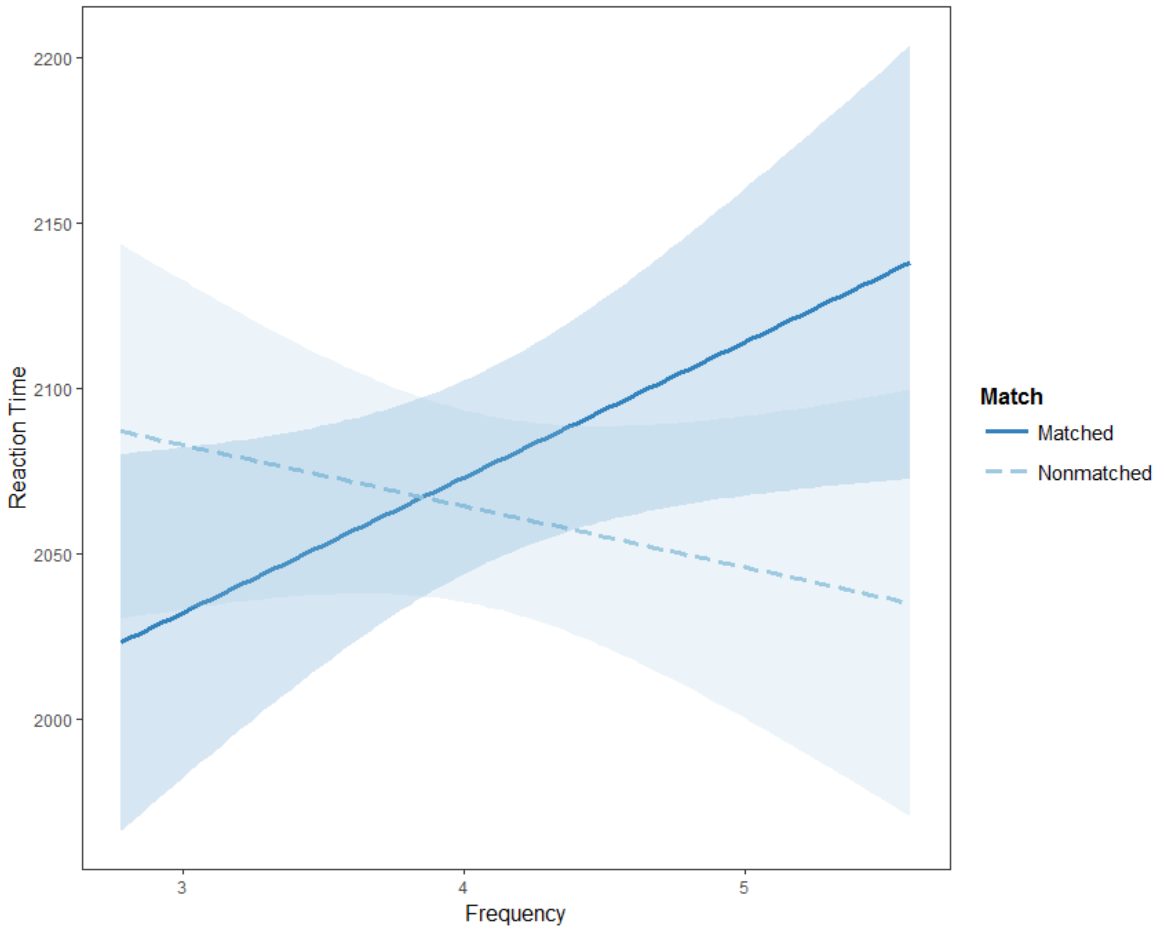

Figure 2. The interaction between zipf frequency and matchedness in adult RT data (ms) from Experiment 1.

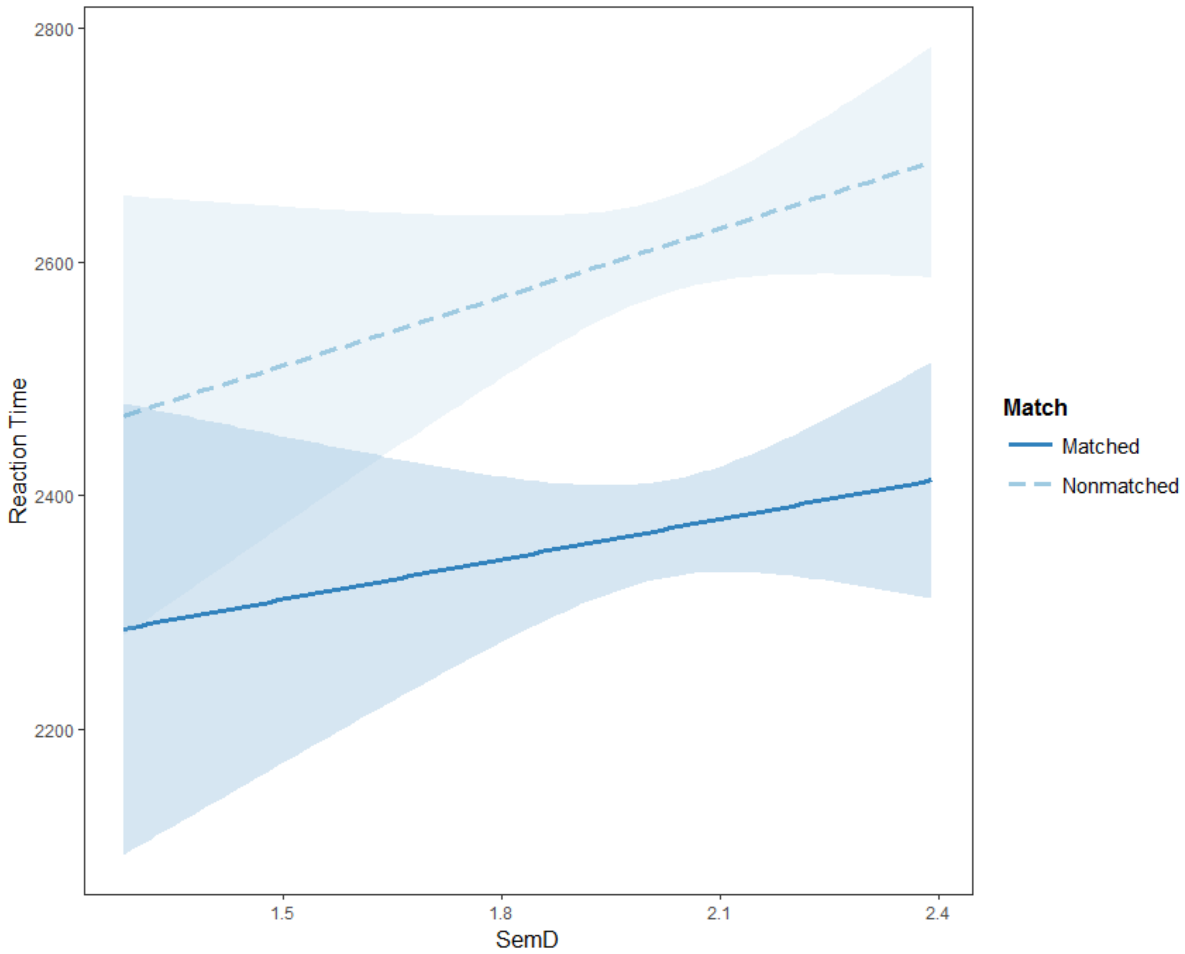

Figure 3. The effects of semantic diversity and matchedness in children's RT data (ms), Experiment 1.

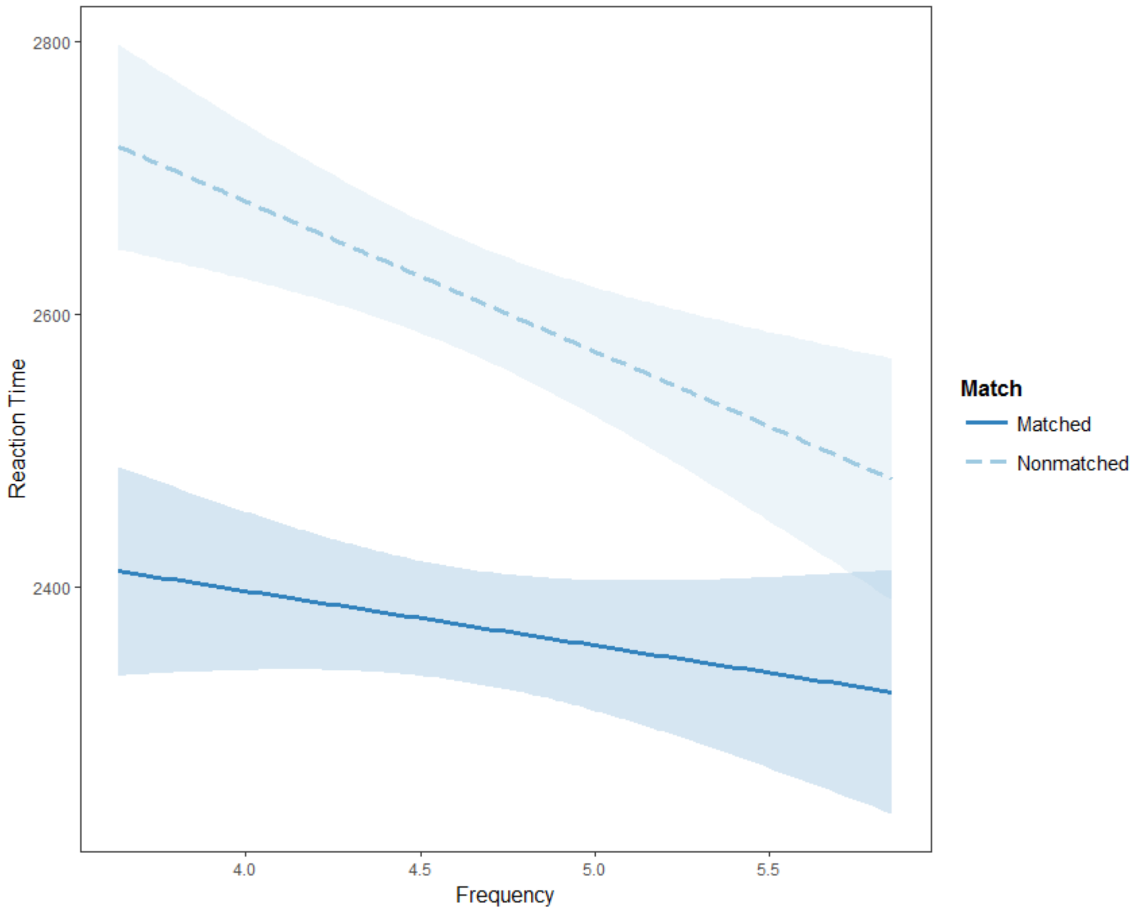

Figure 4. The effects of zipf frequency and matchedness in children’s RT data (ms), Experiment 1.

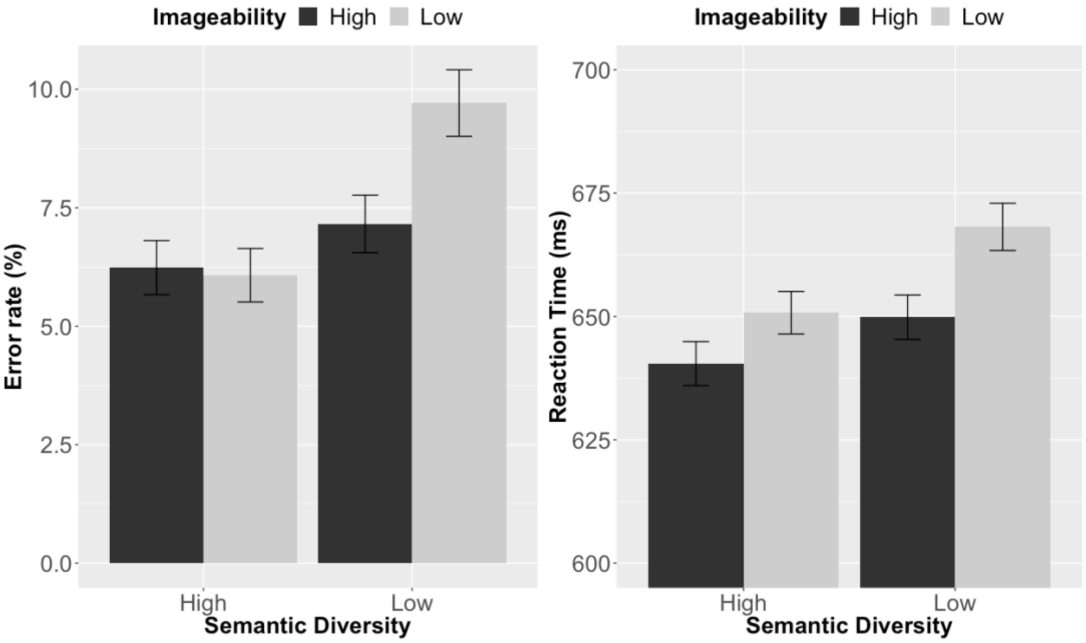

Figure 5. Mean error rate and RT by semantic diversity and imageability condition in Experiment 3. Error bars indicate standard error of mean.

Appendix 1. Experimental stimuli used in Experiment 1 and 2

Word	OCC SemD	OCC Freq	BNC SemD	BNC Freq	Definition
abruptly	1.99	4.81	1.43	4.12	When something stops, all of a sudden
advert	2.03	4.28	1.66	3.52	An announcement on television to persuade you to buy something
amazing	2.31	5.05	1.9	4.25	Something which causes a lot of surprise or wonder
amend	1.92	3.83	1.74	3.61	When you change something to make it better
ancient	2.26	5.31	1.7	4.74	When an object is extremely old
antidote	1.82	3.99	1.84	3.32	Something that cures poison
apple	2.07	5.24	1.54	4.62	A round and crunchy fruit, that is red or green
backyard	1.95	3.72	1.87	3.15	The area behind a house
baking	2.13	3.73	1.1	3.79	When you make a cake in the oven
banana	2.22	4.61	1.64	3.71	A long yellow fruit
barber	2.05	4.29	1.6	4.35	Someone who cuts hair and beards
beastly	1.66	4.24	1.4	3.11	When someone or something is really nasty and unpleasant
bottle	2.08	5.37	1.54	4.62	A container with a lid, that you can drink out of
branch	2.03	5.51	1.82	4.76	Something that is part of a tree
bravery	2.02	4.24	1.67	3.49	When you are not afraid to do something
bread	1.95	5.32	1.56	4.55	Something you use, to make a sandwich
butter	2.03	4.91	1.3	4.34	Something made from milk that you spread on toast
caramel	1.88	3.90	1.07	2.90	A light brown and sticky dessert
cardigan	1.97	3.88	1.42	3.48	A thin, knitted jacket, with buttons down the front
carpet	2.07	5.08	1.49	4.36	A fabric which covers the floor
central	2.28	5.00	2.13	5.33	When something is in the middle
charity	2.14	4.61	1.67	4.59	A group that helps people in need
clean	2.09	5.49	1.91	4.84	When something is fresh and not stained or dirty
cleanly	2.14	3.75	1.78	3.18	When someone does something without making any mess
cleanse	1.93	3.90	1.26	3.04	When you remove the dirt from something
clothes	2.06	5.62	1.59	4.87	Things that you wear
cloudy	2.12	4.14	1.63	3.40	When the sun is not out
cobra	1.65	4.05	1.61	3.36	A poisonous snake
cocoon	1.97	4.16	1.68	3.04	Something a caterpillar makes, to wrap itself in
coffee	2.08	5.20	1.53	4.81	A hot drink that is bitter
colour	2.20	5.65	1.78	5.09	Green, orange, red, and blue, are examples

					of this
consider	2.03	5.40	2.12	5.10	When you think carefully, about something
consume	2.03	4.40	1.56	3.74	When you eat something
cottage	1.87	5.16	1.54	4.51	A small house in the countryside
creek	1.88	4.76	1.68	3.48	A small stream branched off from a river
custard	2.04	4.00	1.21	3.32	A yellow sauce, you have with dessert
customer	1.98	4.60	1.46	4.68	A person who buys something in a shop
dampen	2.03	3.71	1.83	3.15	When you make something wet
darling	1.86	4.92	1.31	4.31	What you call someone you love
debate	2.06	4.39	1.81	4.88	When people have different views, and talk about them
design	2.28	5.20	1.69	5.16	When you make a drawing before you make something
device	2.23	4.75	1.81	4.51	A small handheld machine, like a calculator
diagram	2.11	4.42	1.51	4.08	A drawing that describes how something works
digger	2.17	3.82	1.66	3.04	A machine that makes holes in the ground
display	2.23	5.02	1.84	4.85	When you put something up on a wall for people to see
divine	1.88	4.44	1.24	4.19	Something pleasing and wonderful
doctor	2.06	5.67	1.3	4.98	Someone who treats sick people
dress	1.98	5.78	1.46	4.73	What some girls wear to a party
drivers	2.03	5.15	1.43	4.52	People who steer a car
drowsy	1.77	3.98	1.43	3.20	When you're a bit sleepy
duchess	1.79	4.32	1.21	3.94	A noble woman, married to a duke
dynamite	2.02	4.21	1.73	3.15	A type of explosive
elderly	2.04	4.46	1.47	4.73	People who are very old. Like a grandma
enlarge	2.23	3.65	2.12	3.41	When you make something bigger
example	2.32	5.30	2.16	5.59	One thing picked out from a group, that shows what the group is like
expertly	1.93	3.81	1.68	3.32	When someone does something with a lot of skill
factory	2.20	5.09	1.73	4.68	A place where things are made
feebly	1.88	4.14	1.52	3.15	When you lift something, without any strength
fireman	2.06	4.14	1.35	3.34	Someone who stops burning buildings
footage	1.91	3.91	1.28	3.41	A short video clip, from a film
freedom	2.14	4.83	1.92	4.82	When you can do whatever you like
frozen	2.08	4.81	1.75	4.32	When ice covers everything
fudge	1.91	3.99	1.84	3.15	A soft brown sweet
funny	2.13	5.34	1.53	4.48	Something that makes you laugh
gadget	2.06	4.17	1.6	2.95	A small, and handy machine

giant	2.22	5.51	1.86	4.48	A creature who is super human in size
gilded	1.92	4.11	1.48	3.51	When something is covered in gold paint
goblet	1.59	4.17	0.9	3.15	A fancy wine glass
goblin	1.79	4.47	0.48	3.66	A short, ugly fairy tale creature
grain	2.09	4.86	1.62	4.32	Small seeds used to make flour
grumpy	2.10	4.15	1.62	3.00	When you're in a bad mood
hateful	1.86	4.03	1.41	3.08	Being very mean, to someone
healthy	2.14	4.78	1.85	4.59	When you look, and feel well
holiday	2.10	5.20	1.61	4.89	When school stops in the summer
hoover	1.93	3.80	1.7	3.36	A vacuum cleaner
island	2.15	5.71	1.56	4.85	A small piece of land, surrounded by water
jacket	2.03	5.14	1.43	4.48	Something with long sleeves, which you wear outside
jealous	2.05	4.64	1.54	4.00	What you feel when someone is better at something than you
kitchen	2.03	5.62	1.43	4.90	A room in a house, where you cook things
laptop	1.93	3.96	1.08	2.90	A computer you can bring with you
layout	2.12	3.82	1.6	4.12	The way a room is organised
letter	2.03	5.73	1.64	5.15	Something you find in the alphabet
library	2.39	5.26	1.35	4.94	A building full of books
lipstick	1.96	3.87	1.24	3.60	Makeup that goes on your mouth
lively	2.02	4.68	1.83	4.20	When someone is full of energy and life
lovable	1.88	3.64	1.66	3.20	When you're really nice and everyone likes you
loving	1.92	4.66	1.54	4.26	When you are caring towards someone
machine	2.29	5.50	1.76	4.99	A big device, that does something
manual	2.12	4.02	1.68	4.45	A long set of instructions to help you make something work
marker	2.10	3.96	1.64	3.86	What you use to write on a white board
massive	2.23	4.91	2.09	4.66	When something is very big and heavy
meeting	2.10	5.19	1.95	5.32	When people get together to discuss something
modesty	1.79	3.85	1.79	3.49	When you are not being proud of your achievements
modify	2.11	3.93	1.94	3.90	When you change something a little bit
mouldy	2.14	3.94	1.56	3.00	When fruit has gone off
mythical	1.28	4.14	1.66	3.43	Something from a legend, that's imaginary
narrow	2.06	5.41	1.95	4.74	When something is very thin
nurse	1.96	5.24	1.23	4.53	A person who looks after you, when you are sick in hospital
office	2.12	5.32	2	5.43	A place where people work at desks
officer	2.05	5.25	1.72	4.96	A type of policeman

ostrich	2.10	4.02	1.67	3.15	A big bird with a long neck
outhouse	1.76	3.74	1.42	2.78	A shed outside with a toilet in it
padlock	1.94	3.91	1.38	3.04	Something that stops people from opening a door
painting	2.12	5.01	1.26	4.67	A type of picture made with a brush
paper	2.13	5.67	1.83	5.23	Something you write, and draw on
pasta	1.97	3.84	1.03	3.78	Spaghetti and macaroni, are examples of this
pearly	2.01	3.77	1.46	3.04	When your teeth are very white
phone	1.97	5.43	1.53	4.83	Something you make calls, or texts with
photo	2.19	5.04	1.65	4.19	A picture that you take, with a camera
physics	2.11	4.19	1.07	4.28	A type of science about nature and energy
playful	2.02	3.98	1.67	3.36	When you like fun, and games
pound	2.10	5.31	1.45	4.59	A type of money, you can buy something with
prayer	1.83	5.03	1.26	4.35	When you thank or ask for something from a god
present	2.02	5.60	2.28	5.44	Something you get as a gift
princess	1.75	5.53	1.18	4.48	The daughter of the king and queen
produce	2.35	5.41	2.14	5.09	When you make something, like a factory makes cars
quest	1.98	4.60	1.88	3.99	A long and difficult journey, in search of something
radiant	1.79	4.19	1.6	3.43	When something is very bright and shiny
ranger	2.10	4.05	1.75	3.32	Someone who looks after a forest, or a large park
rapidly	2.12	5.01	2.13	4.70	When something happens very fast
receive	2.02	5.43	2.06	4.91	When someone gives you something
record	2.30	5.25	1.9	5.21	An account that is written about the past
remains	2.22	4.78	2.15	5.00	What is left over after a meal
reopen	2.08	3.64	1.8	3.32	When you unlock a door again, after it has been closed
report	2.18	5.25	1.98	5.38	Something teachers write about students
repress	1.86	3.99	1.49	3.00	When you stop yourself from saying, or doing something
retrace	1.87	4.09	1.37	3.00	When you go back over your footsteps to find something
rewrite	1.97	3.83	1.75	3.28	When you make a neat version of your school work
riddle	1.92	4.28	1.4	3.51	A confusing and fun word puzzle
river	2.12	5.85	1.54	5.00	Lots of flowing water, which you can row a boat down
security	2.20	4.90	1.9	5.18	When you are safe, from danger
single	2.20	5.45	1.99	5.34	When there is only one of something, not double
skilful	2.02	4.12	1.93	3.68	When you're really good at something

smoking	2.02	4.71	1.34	4.50	The habit of using cigarettes
softly	1.95	5.23	1.24	4.42	When you tiptoe without making a sound
spanner	2.20	3.90	1.9	3.20	A tool you use to turn a bolt
special	2.37	5.49	2.24	5.37	Something unique, and important
stadium	1.87	4.34	1.36	4.02	Where football fans go, to watch matches
stream	2.01	5.41	1.8	4.46	A small amount of flowing water, like a small river
superman	1.92	3.72	1.66	3.08	A make-believe hero who can fly
swamp	2.09	4.75	1.77	3.54	An area that is wet, and full of mud and plants
sweetie	1.93	4.02	1.25	3.00	A yummy, sugary snack
tartan	2.08	3.76	1.47	3.48	A checked wool pattern from Scotland
teacup	1.87	3.95	1.51	2.85	Something made of china, that you drink out of
thankful	1.88	4.54	1.68	3.59	When you are very happy for someone's help
trousers	2.03	4.91	1.38	4.32	A type of clothing that covers your legs
trusty	2.05	3.99	1.57	3.15	When something is always reliable
unkindly	1.86	3.68	1.58	3.00	When someone does something in a nasty way to someone else
unwanted	2.15	4.02	1.99	3.95	Something you would like to get rid of
unwell	1.91	3.70	1.59	3.38	When you are feeling sick
update	1.73	4.36	1.63	4.12	When new information is added to something
usually	2.38	5.45	2.11	5.29	When something happens most of the time
venison	1.55	4.00	1.11	3.18	The type of meat you can eat, from a deer
virtual	2.09	4.17	1.9	4.00	When something isn't real
wafer	1.93	3.83	1.7	3.08	A small, thin, crisp biscuit
waiters	1.78	4.39	1.35	3.48	People who work in a restaurant
wallet	1.91	4.24	1.44	3.78	A purse, where you keep your money
warming	1.91	4.24	1.33	4.11	When something heats you up
wedding	1.84	5.03	1.43	4.52	When people get married
workers	2.21	4.96	1.68	5.22	People who have a job

Appendix 2. Predictability rating calculation

Five points were awarded if the target word was produced in the first space (i.e., the target word was the first word that came into a rater's mind, indicating high predictability for that definition); 4 points were awarded if the target was produced in the second space, and so on. For example, the definition for the target word *cloudy* was "when the sun is not out". If a participant produced these five words: *dark, cloudy, overcast, gloomy* and *grey*, it received a score of 4 as *cloudy* was produced in the second space. We calculated a predictability score for each of the 160 definitions using the following procedure:

1. Total points received for the target word was calculated. For example, if 15 people produced *cloudy* in the first position (5 points) and one person produced it in fourth position (2 points), its score was (14 x 5) + (1 x 2).

2. The maximum points the target word could receive was calculated, e.g. the score if all 15 raters produced the word in position one (15 x 5).

3. The score for a target calculated in step 1 was divided by the maximum possible score derived in step 2, [(14x5) + (1x2)]/(15x5) = 0.96

The resulting predictability score represents the predictability of the target word, given its definition. The score varied between 0 and 1: the higher the score the more predictable the definition for identifying its intended target word.