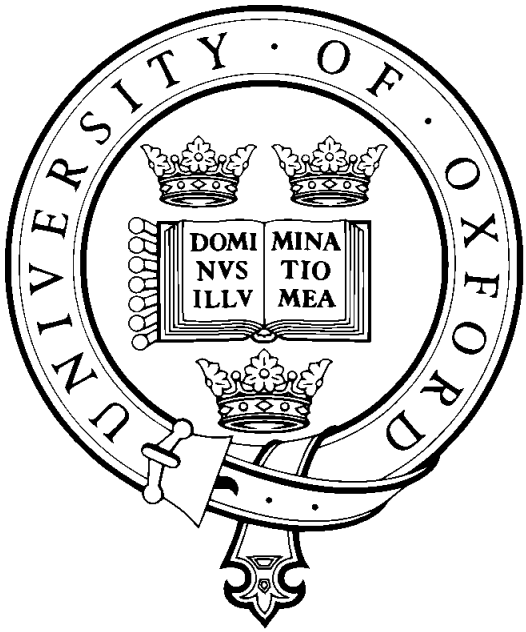
Tracking contextual histories: how long-term and recent lexical experience shapes our processing of words



Ellen Taylor

St John's College

A thesis submitted to the Department of Experimental Psychology, University of Oxford,

in partial fulfilment for the degree of

Doctor of Philosophy

Michaelmas 2022

Table of Contents

ACKNOWLEDGEMENTS **7**

ABSTRACT **8**

CHAPTER 1 LITERATURE REVIEW **11**

Introduction 11
Lexical variation 12
Investigating word frequency *14*
The role of context 17
Document count *18*
Contextual distinctiveness *19*
Semantic diversity *20*
Manipulating contextual variation *24*
Theoretical approaches to understanding semantic diversity 27
The Semantic Distinctiveness Model *27*
Semantic diversity as a measure of polysemy *29*
Relationships with other variables 34
Semantic diversity as a component of semantic richness *34*
Differences between semantic measures *36*
Summary 39
The aims of this thesis 41

CHAPTER 2 COMPARING MEASURES OF SEMANTIC DIVERSITY **44**

Introduction 44
Study 1: Exploring measures of semantic diversity 56
Measures *58*
Results *60*
Discussion *70*
Study 2: Predicting semantic diversity 74
Measures *77*
Analysis *77*
Discussion *82*
General Discussion 84

CHAPTER 3 ESTABLISHING VALUES FOR CONTEXT AVAILABILITY AND SENTENCE AVAILABILITY **92**

Introduction 92
Method 97
Participants *97*
Materials and Procedure *97*
Results 98
Discussion 107

CHAPTER 4 LONG-TERM AND RECENT EXPERIENCE ON PRODUCTION AND COMPREHENSION **111**

Introduction 111
Long-term lexical experience *112*
Recent lexical experience *114*
The current study *118*
Method 121
Participants *121*

Design ... *121*
Materials and Procedure ... *122*
RESULTS ... 126
DISCUSSION ... 139

CHAPTER 5 TESTING THE EFFECTS OF RECENT EXPERIENCE ON SUBSEQUENT WORD PROCESSING THROUGH LEXICAL AND SEMANTIC DECISION TASKS ... 145

INTRODUCTION ... 145
METHOD ... 150
Participants ... *150*
Materials ... *151*
Procedure ... *153*
RESULTS ... 154
The effects of semantic diversity and imageability on lexical and semantic decision at pre-test ... *155*
Summary ... *162*
The effect of pre-test condition on lexical decision performance at post-test ... *163*
Summary ... *175*
DISCUSSION ... 176

CHAPTER 6 CONCLUSIONS, CONTRIBUTION TO THE LITERATURE, AND REFLECTIONS FOR FUTURE RESEARCH ... 184

INTRODUCTION ... 184
UNDERSTANDING LEXICAL-SEMANTIC SPACE ... 184
TOOLS FOR FUTURE RESEARCH ... 188
THE ROLE OF CONTEXT IN LANGUAGE PROCESSING ... 191
RELATIONSHIP BETWEEN LONG-TERM AND RECENT EXPERIENCE ... 193
REFLECTIONS ON PREREGISTRATION ... 198
REFLECTIONS ON ANALYTICAL METHODS ... 201
REFLECTIONS ON ONLINE TESTING ... 203
FINAL THOUGHTS ... 206

BIBLIOGRAPHY ... 208

APPENDIX 1 ... 222

APPENDIX 2 ... 234

APPENDIX 3 ... 239

APPENDIX 4 ... 241

APPENDIX 5 ... 252

APPENDIX 6 ... 256

APPENDIX 7 ... 260

APPENDIX 8 ... 264

APPENDIX 9 ... 266

APPENDIX 10 ... 267

APPENDIX 11 ... 272

APPENDIX 12 ... 275

Table of Tables

.....Chapter 2
Table 2.1: Descriptive statistics for the measures of semantic diversity. 60
Table 2.2: Results from linear mixed effect models. 66
Table 2.3: Comparisons between the base model and semantic diversity model 70
Table 2.4: Principal components analysis reported by Hoffman et al. (2013) 75
Table 2.5: Regression statistics and covariances between factors and variables for Hoffman 80
Table 2.6: Regression statistics and covariances between factors and variables for Johns 80
.....Chapter 3
Table 3.1: Factor loadings for an exploratory factor analysis 103
Table 3.2: Fixed effects predicting lexical decision data for BLP 106
Table 3.3: Fixed effects predicting lexical decision data for ELP 107
.....Chapter 4
Table 4.1: Means and standard deviations for high and low availability wordlists. 123
Table 4.2: Example sentence pairs for AWE and YIELD 124
Table 4.3: Descriptive statstics of reading time and rating response time and score. 128
Table 4.4: Mean and standard deviations of RTs for target words, split by condition 133
Table 4.5: Fixed effects for predicting target word and spill-over word reading time 133
Table 4.6: Mean and standard deviations of response times for rating, split by condition 138
Table 4.7: Rating score, split by condition. 138
Table 4.8: Fixed effects for predicting rating response time (upper section) and Likert score (lower section) 138
.....Chapter 5
Table 5.1: Descriptive statistics for pre-test task performance, separated by condition. 157
Table 5.2: Fixed effects predicting lexical decision response time and accuracy. 160
Table 5.3: Fixed effects predicting semantic relatedness decision response time and accuracy 161
Table 5.4: Descriptive statistics for lexical decision performance, collapsed across conditions 165
Table 5.5: Descriptive statistics for post-test performance for participants in the LD condition 168
Table 5.6: Descriptive statistics for post-test performance for participants in the SD condition 169
Table 5.7: Fixed effected predicting lexical decision performance within participants.....157
Table 5.8: Fixed effects predicting post=test performance between participants.....158
Table 5.9: Fixed effects predicting post-test performance for participants in the LD condition 173
Table 5.10: Fixed effects predicting post-test performance for participants in the SD condition.....174

..... Appendix 1
Table 1: Fixed effects of variables for predicting BLP lexical decision with Hoffman's measure. 223
Table 2: Fixed effects of variables for predicting ELP lexical decision with Hoffman's measure 224
Table 3: Fixed effects of variables for predicting BLP lexical decision with Cevoli's measure 225
Table 4: Fixed effects of variables for predicting ELP lexical decision with Cevoli's measure 225
Table 5: Fixed effects of variables for predicting BLP lexical decision with Johns' (2020) measure 227
Table 6: Fixed effects of variables for predicting ELP lexical decision with Johns' (2020) measure 227
Table 7: Fixed effects of variables for predicting BLP lexical decision with Johns' (2021) measure 229
Table 8: Fixed effects of variables for predicting ELP lexical decision with Johns' (2021) measure 229
Table 9: Fixed effects of variables for predicting BLP lexical decision with document count 231
Table 10: Fixed effects of variables for predicting ELP lexical decision with document count 231
Table 11: Proportion of the variance in lexical decision explained by the linear-mixed effect model 233
..... Appendix 3
Table 1: Range and mean values for Altarriba et al.'s CA norms and our CA norms. 240

Table 2: Descriptive statistics for words that occurred in both the current and Altarriba et al.'s item-set. 240
..... Appendix 4
Table 1: Factor loadings for an exploratory factor analysis using oblique rotation. 245
Table 2: Fixed effects predicting lexical decision in BLP data. 246
Table 3: Fixed effects predicting lexical decision in ELP data. 246
..... Appendix 5
Table 1: Main effects for predicting sentence availability 253
Table 2: Main effects for predicting sentence availability in an updated model 253
Table 3: Main effects for predicting context availability 254
Table 4: Main effects for predicting context availability in an updated model 254
..... Appendix 6
Table 1: Percentage of responses where words were correctly predicted 261
Table 2: Fixed effects of condition on predictability 263
..... Appendix 9
Table 1: Pearson correlations between participants' ratings and sentence availability norms. 267
Table 2: Fixed effects predicting lexical decision performance in BLP data 269
Table 3: Coefficients predicting lexical decision performance in ELP data. 270
..... Appendix 10
Table 1: Fixed effects predicting lexical decision performance 276
Table 2: Fixed effects predicting semantic decision performance. 277

Table of Figures

Chapter 1
Figure 1.1: Roadmap showing the headings and subheadings of this literature review. 12
Figure 1.2: A visual representation of latent semantic analysis 21
Chapter 2
Figure 2.1: Scatterplots showing the relationship between semantic diversity measures. 62
Figure 2.2: Heatmap showing correlation coefficients between the four diversity measures. 63
Figure 2.3: Heatmap showing correlations between semantic diversity and other measures. 66
Figure 2.4: Path diagram for the adjusted structural equation model predicting Hoffman's diversity 79
Figure 2.5: Path diagram for the adjusted structural equation model predicting Johns' diversity 82
Chapter 3
Figure 3.1: Predicted relationships between CA, SA, imageability and semantic diversity 96
Figure 3.2: Histograms showing the distribution of CA and SA 99
Figure 3.3: Scatterplot showing the relationship between CA and SA 100
Figure 3.4: Pearson correlations for CA, SA and other lexical variables 101
Figure 3.5: Scree plot showing the factors identified by parallel analysis 104
Chapter 4
Figure 4.1: Flowchart showing between subjects' priming conditions 122
Figure 4.2: Histograms for target word and spill-over RT, rating RT, and rating score. 130
Figure 4.3: Violin plot of target reading time, split by condition 136
Chapter 5
Figure 5.1: Factorial design of the target wordlists. 152
Figure 5.2: Histograms showing the distribution of lexical and semantic decision response times. 157
Figure 5.3: Bar plots showing the average performance for words 158
Figure 5.4: Histograms showing the distribution of lexical decision RTs 166
Figure 5.5: Box plots showing pre-test and post-test RT in the lexical decision condition. 167
Figure 5.6: Histograms showing the distribution of post-test lexical decision RTs 170
Figure 5.7: Box plots showing post-test response time, split by pre-test condition 172
Figure 5.8: Bar plots showing the average post-test performance, split by pre-test condition 175
Appendix 4
Figure 1: Histograms showing the distribution of CA and SA 242
Figure 2: Scatterplot showing the relationship between CA and SA 243
Figure 3: Pearson correlations for CA, SA and other lexical variables 244
Figure 4: Scatterplot showing the relationship between CA and SA in function words 249
Figure 5: Pearson correlations for CA, SA and other lexical variables in function words 250

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Kate Nation, for her incredible support and guidance throughout my DPhil. I would also like to thank Yaling Hsiao, who offered invaluable advice and expertise. Together, their encouragement and feedback helped to shape this thesis. I am immensely grateful for their patience and vision.

Thanks to the Medical Research Council, who funded this work and supported my academic development throughout my course.

I would like to thank the members of staff and other doctoral students from the Department of Experimental Psychology at the University of Oxford. Particularly, thanks to Maggie Snowling for her supervision; Dorothy Bishop for her inspiring enthusiasm for reproducible science; Paul Thompson for his guidance in structural equation modelling; and Mohen Zhang and Sean McCarron for ensuring I could remotely access my computer whilst away from Oxford.

Finally, I would like to thank my friends and family, and the postgraduate community at St John's College, Oxford for their support.

ABSTRACT

This thesis investigated the role of contextual information on lexical processing. Using secondary data analysis and online experiments, we tested how written language processing is shaped by recent experience and by the lexical histories words have accrued across the lifetime. Chapter 1 is a literature review. It reviewed the different ways of conceptualising semantic information, and discussed the importance of 'semantic diversity' (Jones et al., 2012; Hoffman et al., 2013), a metric which captures the variation in words' contextual histories. Semantic diversity has been conceptualised and operationalised in different ways. Chapter 2 compared four different measures of semantic diversity by exploring their relationships with other item-level variables and their effects on lexical decision data. Small differences between the different measures demonstrated the importance of the corpus size and sample from which diversity is calculated. In Chapter 3, 3,000 English words were normed for context availability and sentence availability, two variables that reflect the ease with which words can be associated with a context or a sentence. These norms allowed for the exploration of lexical-semantic space on a larger scale. Context availability and sentence availability tapped different facets of semantic knowledge, and each explained unique variance in lexical decision performance. Context availability appeared to capture information relating to our interactive and imageable experience with words, while sentence availability captured linguistic variation in a similar way to semantic diversity. As sentence availability is a new metric and warrants further investigation, we employed this in Chapter 4 as a marker of long-term linguistic experience. Chapter 4 investigated the role of recent experience, by testing the effects of repeated exposure to words across sentence reading and sentence availability rating tasks. By presenting the same words in each task, we tested how longer-term lexical knowledge (indexed by sentence availability) interacts with the short-term influence of recent encounters (induced by priming). Words high in sentence availability were associated with more efficient performance in

both tasks, therefore validating the measure. There were no main effects of priming but the interaction between priming and sentence availability was significant during sentence reading: words with high sentence availability were read faster during task 1 (effectively unprimed) compared to task 2 (primed). We ventured that this might be due to a difference between the context participants themselves produced in the rating task and the context they read in the subsequent task, but this result was difficult to interpret. Chapter 5 also tested long-term and recent experience using a more robust experimental design. Long-term experience of words was indexed by their semantic diversity and imageability values, which were manipulated orthogonally to compare high vs low imageability and high vs low semantic diversity. Participants first encountered words via either lexical decision or semantic decision (task 1; serving as unprimed performance), then all participants responded to the same words in a subsequent lexical decision task (task 2; serving as primed performance). All participants therefore encountered the same words, but half took part in the same task pre-test and post-test, while for the other half of participants, the task differed. Analysing the results from task 1, we found that semantic diversity was associated with more efficient lexical decision and less efficient semantic decision performance, while imageability was associated with more efficient performance in both tasks. These observations replicate those in the literature. Lexical decision performance was better in task 2 than 1, suggesting that repetition facilitated performance. Semantic diversity was no longer associated with lexical decision performance when it was administered as task 2, suggesting that recent experience (via task 1, regardless of whether that was lexical decision or semantic decision) might have temporarily masked the effects of a word's long-term contextual history, as estimated by its semantic diversity.

Together, the results of this thesis demonstrate the importance of prior lexical experience on current linguistic behaviour. Through studying item-level variables, it shows the importance of both theoretical conceptualisations and computation methods in calculating semantic measures. It supports the idea of semantic richness as a complex and multifaceted construct (Pexman et al., 2002) and identifies potential distinct facets of meaning-based knowledge. It also gives a basis for

understanding the relationship between long-term experience and recent word encounters through priming studies. The thesis closes with a general discussion that offers reflections and suggestions for future research.

CHAPTER 1
LITERATURE REVIEW

INTRODUCTION

Some words are easier to process than others. For example, the word *cat* is likely to be much less cognitively taxing to read, comprehend and produce than the word *incumbent*. But why is this? *Cat* is noticeably shorter than *incumbent* and is monosyllabic. It is typically learnt much earlier in life and is more frequently encountered in both written and spoken language. The word *cat* is concrete; it's easy to imagine a cat, and you can physically interact with one, while *incumbent* is more difficult to visualise or to place within a context. These factors are all known to affect lexical processing, even when words are presented in isolation in the context of a simple experimental task such as lexical decision or word naming (e.g., Balota et al., 2007; Kueleers et al., 2012).

Factors that are thought to affect processing have been quantified to give values known as lexical variables. Identifying and studying lexical variables gives insight to how language is organised in the mental lexicon. By determining which variables predict word processing in various tasks, we can understand how words are encoded, stored and retrieved from memory. This has important implications both for understanding typical processing, and for understanding processing in individuals with language deficits (Luzzatti et al., 2002).

The extent to which variables are distinct from each other, and their unique influence on lexical processing, is unclear not least because lexical factors tend to correlate with each other. For example, words that are higher in frequency are also likely to be acquired early in life, be short in length, and occur in a broader range of linguistic contexts (Baayen, 2010). As such, it can be difficult to tease apart different lexical measures in order to determine what is driving linguistic behaviour.

This chapter reviews research involving some key lexical variables and how they relate to models of lexical organisation. A chapter roadmap is shown in Figure 1.1. First, I give a summary of the frequency effect, a well-researched effect in the literature whereby words that occur more often are processed more efficiently. I then discuss the role of context in building semantic representations of words. Next, I review different theoretical approaches to semantic diversity. Finally, I describe the relationship between diversity and other variables relating to word meaning, exploring Pexman et al.'s (2007) idea of a multifaceted semantic richness. The aims of this thesis are then presented considering this evidence.

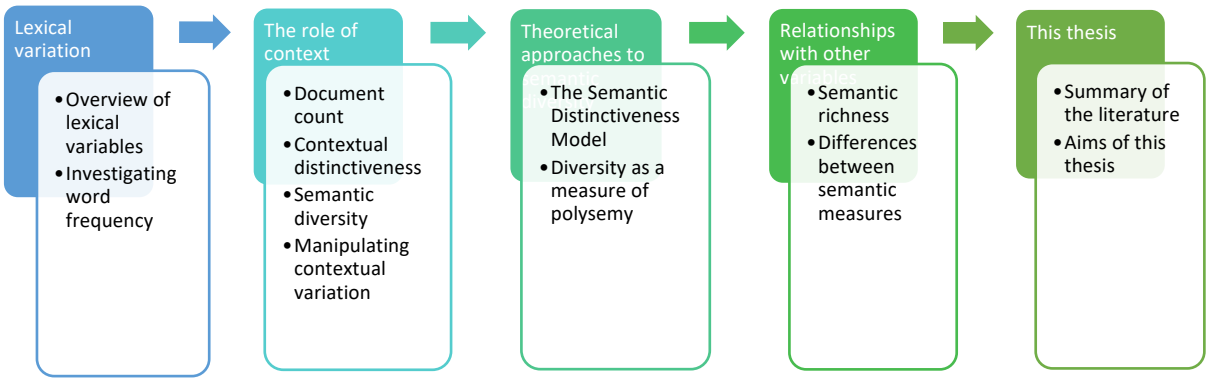


Figure 1.1: Roadmap showing the headings and subheadings of this literature review.

LEXICAL VARIATION

OVERVIEW OF LEXICAL VARIABLES

Broadly, lexical variables are thought to capture our long-term experience with words. Corpus-derived statistics (like *frequency* and *document count*) use large collections of language to calculate metrics that reflect a word's usage within that language sample. Participant-based ratings (like *concreteness*) are conceptually driven measures which reflect a person's experience with a

word, built across their lifetime. Both corpus- and participant-derived measures provide a proxy for the knowledge we gather about words across a lifetime of experience. Research demonstrates that numerous lexical qualities affect the efficiency of word processing, and there are many ways to measure lexical properties of words. Variables can be grouped by similarities in their theoretical nature and the effects they exert. For example, some measures capture the complexity of orthographic form such as counts of letters or syllables. We might also want to know how regular the word is by calculating its orthographic similarity to other words (*bigram frequency* or *orthographic neighbourhood size*). Or, we might want to know how often words appear as part of a derived word or a compound word (*morphological family size*). Other variables focus on how often we encounter words. The most notable measure here is *word frequency*, which is typically a count of how often a word occurs within a corpus. However, there are other ways of conceptualising occurrence. We might want to consider how many people know a word (*word prevalence*) or how many different documents a word occurs in (*document count*).

Turning to word meaning, there are multiple ways we can reflect semantic knowledge associated with a word. Broadly, semantic richness has been proposed as a multifaceted construct relating to how rich the meaning associated with a word is (Pexman et al., 2008). This has been captured in different ways by different measures. For example, *concreteness* refers to how concrete or abstract the word's referent is perceived to be; while *number of senses* refers to the number of nuanced meanings a word has. These two metrics both relate to the amount of semantic information a word is associated with but appear to capture different components of semantic knowledge (Pexman et al., 2004). The long-term experience captured by lexical variables allows individuals to build knowledge of this word in the lexicon. Word knowledge might be linguistic, containing information about how the word is used in language; or they might be experiential, containing information about how we interact with the word concept (Vigliocco et al., 2009). In order to process language efficiently, people rely on both components of word-knowledge; we

combine statistical information we have gathered about how words are used, along with our conceptual understanding of what those words mean. This thesis is primarily concerned with understanding the relationships between semantic variables. My goal is to tease apart complex relationships within lexical-semantic space and to aid our understanding of lexical organisation. Before addressing the wealth of research into semantic variables, I first review a lexical variable that has been widely investigated: word frequency.

INVESTIGATING WORD FREQUENCY

Frequency refers to how often a word is encountered. This is typically measured by counting the number of times a word occurs within a large language sample. Examples include the summed count of the occurrences of written and spoken words in the British National Corpus (Leech & Rayson, 2014), and the standardised Zipf frequency measure taken from the SUBTLEX corpus of British television subtitles (Van Heuven et al., 2014).

Words that occur more often in language are generally easier to process than those that occur less often (Monsell et al., 1989). This is known as the frequency effect. The frequency effect is well established in the literature and has been observed in multiple tasks (for review, see Brysbaert et al., 2018). To list just a few examples, in an eye-tracking study of text reading, Rayner and Raney (1996) found that gaze durations were longer for low than high frequency words. Savin (1963) found fewer errors in the perception of low-decibel speech when word frequency was high. As well as frequency effects emerging in a wide range of lexical processing tasks (see Monsell, 2012), they are consistent across languages (Bates et al., 2001; Bonin et al., 2001), and are seen when other key linguistic variables are controlled for (Brysbaert, 1996).

The frequency effect can be accommodated by a number of different theoretical accounts. Theories from memory research can explain why some words might be easier to retrieve and process than others. Human memory has been theorised to use structures of statistical optimisation to prioritise information most likely to be needed in the future (Anderson & Milson, 1989). Words in our lexicon are therefore proposed to be more accessible if they have been retrieved before. The Production-Distribution-Comprehension hypothesis (MacDonald, 2013) builds upon this idea. This account suggests that our use and experience of language shapes our subsequent use and perception. Internal biases exist to reduce the burdens of lexical access and retreival from memory. Words that we recently read, hear or produce are made more accessible. This in turn increases our likelihood of producing these words, and makes it easier for us to comprehend them when encountered in our environmnet. Words high in frequency will be more likely to have been recently comprehended or produced, therefore making it easier for them to be retreived again.

Given an extensive enough sample of language, word frequency should reflect our experience with words, at least as a proxy. However, word frequency measures have been criticised. Notably, they are limited by the size and sampling techniques of the corpus used. Differences in corpus size can result in different frequency values, which can be difficult to compare. To address this problem, frequency can be standardised by calculating rate of occurrence per set number of words (Brezina & Timperley, 2017), giving metrics such as frequency per million words. Alternatively, measures such as Zipf frequency perform logarithmic transformation on standardised frequency to give an evenly distributed scale (Brysbaert et al., 2013). This is calculated from the word's frequency count, the total number of tokens in the corpus, and the number of unique tokens in the corpus.

The corpus from which frequency is calculated affects how representative the measure is. Corpora of written texts have been criticised as they are unlikely to reflect everyday language use (Johns, 2021). In line with this idea, word frequency as estimated via a corpus of television subtitles

accounts for more variance in lexical decision than text-based frequency measures (Brysbaert et al., 2011), as do calculations from social media posts (Gimenes & New, 2016). Limitations appear particularly salient for low frequency words. Examining a corpus of over 1,000,000 printed words, Lovelace (1988) noted that relatively common words in day-to-day usage like *cucumber* and *toaster* received a frequency count of 0, while more obscure words like *conjugate* and *polynomials* had a count of 10. This highlights the importance of a representative corpus. Moreover, using computational simulations, Kuperman and Van Dyke (2013) demonstrated that corpus-based frequency measures overestimate the counts of low frequency words. This makes measures of frequency a less reliable fit to behavioural data, particularly for readers with less experience. Kuperman and Van Dyke therefore asked participants to rate how frequently they use words, measured on a Likert scale of 1 (never) to 7 (several times a day). The researchers found that this 'subjective frequency' explained more variance in participants' eye-movement latencies during sentence reading, and lexical decision latencies. Another way of considering word occurrence is to count what proportion of the population know a word. Following this method, 'word prevalence' gives a count similar to frequency, but across participant vocabularies rather than across a corpus. Prevalence ratings also provide a better fit to word processing data than word frequency (Brysbaert et al., 2019).

There has also been debate over what exactly frequency reflects. Words which are more frequent are also likely to be shorter, acquired earlier in life, be less morphologically complex, and occur in a diverse range of contexts (Baayan, 2010). Because of the correlations between different lexical properties, it is difficult to establish whether the frequency effect occurs due to word repetition (i.e., high frequency words are easier to process because we encounter them again and again) or due to overlapping variance that could be attributed to other lexical measures (i.e., high frequency words are easier to process because they are also shorter, acquired earlier in life, etc.). Baayen (2010) attempted to distinguish between these two possibilities. Using regression, he found

that 91% of variance in word frequency could be explained by other lexical variables, including measures of morphology and contextual variation. Using computational modelling, Baayen then demonstrated that lexical processing was best explained by the Naive Discriminative Reader model (Baayen et al., 2011). In this model, words are learnt from semantically rich input, suggesting that the context in which a word appeared is important for word learning. Baayen's analysis therefore demonstrates that, though frequency is an important predictor of word processing, it is not necessarily the repetition of words that drives the frequency effect. This illustrates the complexity of lexical variables more broadly. While multiple lexical measures have been found to affect linguistic processing (see below), the close relationships between these variables make it difficult to interpret such effects.

THE ROLE OF CONTEXT

Bayaan's work points towards the importance of a word's contextual history. This idea has been built upon in the lexical legacy hypothesis (Nation, 2017). Nation's theory describes the importance of diverse and meaningful contexts in reading development. According to this perspective, experience accrued across the lifetime provides a lexical history that is then reflected in behavioural reading measures. This hypothesis explains the frequency effect by proposing that repetition alone will not update a word's lexical history, nor encountering a word in an identical context to prior experience. Rather, it is the diverse occurrences that provide rich and nuanced semantic information that then benefits reading behaviours. In order to test whether there are differences between words that appear in uniform vs diverse contexts, we need to be able to quantify a word's lexical history. Various attempts have been made to capture contextual variation through lexical statistics. This section reviews different approaches to conceptualising contextual variation.

DOCUMENT COUNT

Document count is also referred to as 'contextual diversity' in the literature. This corpus-based measure counts the number of documents that a word occurs in, therefore accounting for contextual information that word frequency does not measure. Words that appear in many documents are assumed to be more contextually diverse than those appearing in only a few documents. Adelman et al. (2006) found that document count was a better predictor of word naming and lexical decision performance than word frequency. Perea et al. (2013) replicated these results with 9- and 10-year-old children, demonstrating that considering context is important across the lifetime.

While evidence suggests that contextual diversity is an important predictor of lexical processing, document count might not be the best metric to capture a word's contextual usage. Hoffman et al. (2013) reported that document count correlates highly with word frequency ($r = .95$). From this, they argued that this measure is more akin to occurrence and does not capture the variability in a word's contextual use. A word may have a high document count if it appears in several documents; but this count does not consider the semantic content of those documents. To give their example, *tax* might appear in several documents that are contextually similar (e.g., all discussing finances), but this word is not particularly contextually flexible. Therefore, document count is perhaps more similar to word frequency than a discrete measure of contextual diversity.

In a review of contextual diversity, Caldwell-Harris (2021) discussed research using event-related potentials to determine differences in the effects of contextual diversity and frequency. Recounting the work of Vergara-Martinez et al. (2017), Caldwell-Harris described that differences in scalp distribution and timing of event-related potentials suggest that the effects of contextual

diversity and word frequency arise from two different sources in linguistic processing, and relate to different word properties. Therefore, the context that words appear in does seem to be important; but a different way of capturing this information is needed in order to avoid confounding contextual variation with word frequency.

CONTEXTUAL DISTINCTIVENESS

Another attempt to quantify contextual variability comes from McDonald and Shillcock (2001). They stated the importance of including contextual information in corpus-derived statistics, arguing that the mental representation of a word is difficult to separate from the context in which it has occurred. They considered lexical co-occurrence to give a 'contextual distinctiveness' value for each word. McDonald and Shillcock calculated this by measuring the relative entropy between the distribution of context words surrounding a target word, and the distribution of those same context words in a corpus without considering the target word. Put more simply, this calculation looks at the similarity between the occurrence of words in a contextual window (i.e., the five words before and after a target word) vs the frequency of words in a large corpus. Having calculated contextual distinctiveness values for 144 words using data from the British National Corpus, McDonald and Shillcock demonstrated that this measure was a better predictor of lexical decision performance than word frequency. Although their findings indicate that a word's contextual history beyond frequency of occurrence is important, as with document count, this measure also has its limitations. Contextual distinctiveness, as Hoffman et al. (2013) noted, was calculated using a narrow window of five words either side of the target word. This window size was based on McDonald's (2000) experiments, which demonstrated that a window of five words provided the best fit to lexical decision data, compared to window sizes of one to ten words. However, Hoffman et al. (2013) argued that this metric might be skewed due to particular phrases that commonly occur together,

and therefore might not capture the broader variation in a word's context. In order to capture contextual variation at a wider level, *semantic diversity* has been proposed as a lexical measure.

SEMANTIC DIVERSITY

Semantic diversity is a corpus-derived metric that reflects variability in the contexts in which a word appears. This captures the contextual usage that a word has accrued across a corpus and serves as an approximation of the lexical history words develop through an individual person's experience. To give the example taken from Hoffman et al. (2013), *perjury* and *predicament* are two words with similar lexical properties: both have one core meaning, yet the variability of contexts they are likely to appear in is different. *Perjury* mostly appears in legal contexts, whereas *predicament* appears in contexts that describe a wide range of situations. To capture this type of variation, Hoffman et al. calculated semantic diversity.

Diversity values were calculated using latent semantic analysis, with data from the British National Corpus. Latent semantic analysis is a technique used to extract meaningful lexical information from a corpus. Following this method, first a document-term matrix is created. This is a matrix in which the occurrence of particular words within documents is counted (see Figure 1.2). This then gives a vector representation for each document (based on how words occur within this sentence) and for each word (based on the documents it occurs in). These are highlighted for the example in Figure 1.2. Here, 'young' would have a vector [0 0 1 0] based on the documents it occurs in. 'The wise young woman' would have a vector [0 1 1 0 0 0 0] based on the words that occur in it. These vectors provide information on where both words and documents sit within lexical semantic space. Singular value decomposition is then performed to find the latent features underlying these vectors. Here, data is reduced to show the higher-order relationships between words and contexts. This gives a new, smaller set of vectors which can be used to calculate semantic similarity.

Turning back to Hoffman et al.'s (2013) semantic diversity, they calculated a document-term matrix in which *documents* were 1,000-word passages of text, and *terms* were words which occurred in at least 40 documents. After calculating vectors from this matrix, singular value decomposition was performed to reduce the dimensions and give two sets of vectors: vectors for contexts and vectors for words. Next, for every word, the authors calculated the cosine of the context-vectors in which that word appeared to determine similarity between contexts. From these cosines, the mean was taken, log transformed and reversed in order to give a semantic diversity value. These values ranged from 0.08 (low diversity) to 2.41 (high diversity).

	old	young	wise	tired	unhappy	man	woman	
"the old wise man"	1	0	1	0	0	1	1	CV
"the tired old man"	1	0	0	1	0	1	1	CV
"the wise young woman"	0	1	1	0	0	0	0	CV
"the tired unhappy woman"	0	0	0	1	1	0	0	CV
	WV	WV	WV	WV	WV	WV	WV	

Figure 2.2: A visual representation of a document-term matrix used in latent semantic analysis. Document vectors are represented as rows, and word vectors as columns.

Note: WV refers to word vector; CV refers to context vector. These labels correspond to the column above (for WV) or the row beside (CV) the label. Vectors for "young" and "the wise young woman" are highlighted and referred to in the text.

Hoffman et al.'s measure correlated positively with word frequency, document count and number of senses; and negatively with concreteness, imageability, context availability and contextual distinctiveness. High diversity has been associated with faster performance in lexical decision tasks, but slower semantic judgement performance (Hoffman & Woollams, 2015). Latent

semantic analysis has also been used on the Oxford Children's Corpus to create a semantic diversity measure relevant to children's linguistic experience. Having calculated diversity values using a method similar to Hoffman et al.'s, research has examined the influence of semantic diversity on children's word reading (e.g., Hsiao & Nation, 2018; Hsiao et al., 2020; Pagán et al., 2020). Semantic diversity facilitates children's word naming and lexical decision (Hsiao & Nation, 2018), and facilitates children's word processing during sentence reading (Pagán et al., 2020).

Studying clinical populations, Hoffman et al. (2011) demonstrated that semantic diversity relates to lexical performance in a way distinct from word frequency. They used semantic diversity values to explain the lack of frequency effects observed in individuals with semantic aphasia following a stroke. Patients with semantic aphasia typically have damage to the right inferior frontal gyrus, which has been associated with semantic and memory functions (Winhuisen et al., 2005). In Hoffman et al.'s study, these patients showed no difference in the processing of high and low frequency words. In fact, in some tasks, the frequency effect was reversed, with patients performing better for low frequency words. Semantic diversity, however, did explain patients' performance. This can be understood by considering how information is selected for retrieval. In healthy individuals, we would expect information that is more frequently used (i.e., items higher in word frequency) to be more readily retrieved than those that we rarely use; and this is supported by the robust literature on the frequency effect. But, as evidenced by the growing research into contextual diversity, another important factor is the number of contexts activated by a word. In patients with semantic aphasia, words high in diversity are more difficult to process than low diversity words. This could be because diverse words are more likely to activate multiple semantic representations and lexical associates. The activation could be difficult to 'filter through' for individuals with semantic aphasia. The resulting effect is difficulty processing high diversity words. Indeed, this is what Hoffman et al. observed. Semantic diversity accounted for the lack of frequency effects, and, while patients did not show a frequency effect across all items, expected frequency effects were reported

in low diversity words. This demonstrates that, unlike document count which arguably is conflated by frequency-type information (Hoffman et al., 2013), semantic diversity captures distinct linguistic information.

Jones et al. (2012) have also coined the term semantic diversity to refer to contextual variation across language experience, as estimated from corpus-derived statistics. The researchers calculated linguistic dissimilarity between documents in which a word occurs. They then used this value to create a document count measure that is weighted by the semantic uniqueness of contexts in which a word appears. Like Hoffman et al.'s calculation of semantic diversity based on latent semantic analysis, Jones et al.'s measure explains more variance in lexical decision performance than both word frequency and document count (see Johns et al., 2012, Johns et al., 2016a, and Johns et al. 2016b for replication studies using language corpora).

While using corpora allows us to gather largescale information on how words are used, this can only act as a proxy for individual's language experience. This means that, by their nature, corpus-derived statistics will be limited by the sample of text on which they are based. To use the British National Corpus as an example, this comprises written and spoken texts, including books, legal documents, and essays. While this gives a large sample, it is unlikely that this is reflective of any individual's linguistic experience. The reasons for this are twofold. First, there are different registers of language, and the scripted, pre-prepared content of a corpus likely differs from the spontaneous speech from which most of our language encounters are based (Roland et al., 2007). Only 10% of the BNC is made up of spoken language; and the academic texts within the corpus might not reflect typical language use. Second, the BNC does not contain information about who produced the language, and therefore does not allow us to track the language use of individual people. To further explore the role of contextual variation at the individual level, research has begun to derive measurements that consider the producer of language. For example, Johns et al. (2020) created an

'author prevalence' semantic diversity metric, using a corpus of fiction and non-fiction books. This considered books as individual discourse contexts, but also counted how frequently a word was used by a particular author across texts. This measure was a more successful fit to lexical decision and naming data than other diversity measures. Similarly, Johns (2021) calculated semantic diversity values from posts on the internet forum Reddit. He counted the number of discourses (or *subreddits*) a word occurred in, as well as the number of users who had produced a word. He then transformed these values to reflect semantic diversity, and found that the user-based semantic diversity value explained more variance in behavioural data, providing a better fit than semantic diversity as calculated by Jones et al. (2012). These studies add to the literature in two important ways. Firstly, they demonstrate the value of using corpora that accurately reflect language experience. Johns et al.'s (2020) corpus was based on books which, though reflect a scripted register, might be a better reflection of typical written-word encounters than the academic and legal documents included in the BNC. Johns' (2021) corpus reflecting language usage on internet forums likely reflects everyday speech. Secondly, Johns et al.'s (2020) and Johns' (2021) experiments found that metrics which considered the producer of language provided the best fit to behavioural data. This demonstrates the importance of considering item-level data within the context of an individual's language use: differences in vocabulary and syntax between language producers accounts for differences in lexical processing. These measures are a step towards capturing individual language experience, and in turn they could allow more adequate proxies of language experience to be more accurate.

MANIPULATING CONTEXTUAL VARIATION

As well as exploring the influence of lexical variables on behavioural tasks, word learning experiments have manipulated the contexts in which words appear. In these paradigms, participants

typically encounter novel words in linguistic contexts through incidental reading. These contexts will either be thematically similar or diverse. Participants' subsequent comprehension and processing efficiency of those non-words can then be tested via a variety of behavioural techniques. This allows us to examine the influence of contextual experience, while controlling for extraneous variables that correlate with semantic diversity (e.g., word frequency or age of acquisition).

Using a word learning paradigm, Bolger et al. (2008) found that words are learnt more effectively when presented in diverse rather than uniform contexts. This supports the idea that repetition alone is not the only driver of linguistic acquisition. Similarly, when non-words are presented in text with either consistent themes or differing themes, a benefit in speeded recognition is observed for differing-themed non-words (Johns et al., 2016a). The same study saw same-themed words were responded to faster during semantic judgement tasks, demonstrating a task-dependant effect of contextual uniformity. Results have also been reported in classroom settings. Rosa et al. (2017) found that Grade 3 children were better at remembering words learnt in diverse contexts than those learnt in redundant contexts. Here, contextual diversity was also associated with better performance in a pictogram matching task and in a multiple-choice sentence completion task. This demonstrates that contextual information is not only important for acquiring information on word form, but also in learning the meanings of words.

Joseph and Nation (2018) investigated the role of context on children's word learning using eye-tracking. This technique allows for the 'online' processing of words, which is thought to more accurately reflect live reading processes than 'offline' measures like comprehension questions (for further discussion, see Joseph et al., 2013). Novel words were presented in semantically diverse or uniform contexts and eye movements were monitored. There was no main effect of diversity condition on eye movements or offline learning measures. However, diversity did interact with children's reading comprehension ability. Children with poorer comprehension spent longer reading

words in both the diverse and non-diverse condition; but children with better comprehension spent less time reading in the non-diverse condition. The researchers concluded that diversity indirectly affects children's reading, with good comprehenders showing more effortful processing for words presented in diverse contexts.

Mak et al. (2020) further investigated the role of diverse contexts in novel word acquisition. They found that words are better learnt if 'anchored' in memory, a finding congruous with research demonstrating the benefits words receive when they connect to existing knowledge (Cervetti et al., 2016). Mak et al. initially reported a processing benefit for words that appeared in uniform contexts. However, the researchers then demonstrated that, once words had an opportunity to 'anchor' themselves to semantic knowledge (i.e., once words were secured via attachment to existing topics in long-term memory), high diversity words were processed more efficiently.

The ability of contextual history to explain word acquisition has also been evidenced by simulations. Landauer and Dumais (1997) demonstrated that latent semantic analysis was able to successfully simulate word learning at a similar rate to school-aged children. Importantly, this model relied on no prior linguistic similarity knowledge. This highlights the crucial role of context in language representation and supports the idea of a statistical learning mechanism in language acquisition. These effects are similar to those of other types of contextual variability within episodic memory research. Lohnas et al. (2011) describe 'contextual-variability theory', in which experiences encoded at different times are associated with contexts; each context provides an additional path to retrieving the item from memory.

Overall then, word learning experiments can help us to understand how contextual variation can influence the learning of new words. Alongside studies exploring the effects of lexical variables, word learning experiments can help us to understand how different contexts inform the linguistic knowledge we acquire. In these studies, variation appears to be beneficial to learning, though these

patterns are complex and might depend on the stage of learning and how learning is measured. This might connect with the effects of semantic diversity on tasks like lexical decision and word naming. However, different theoretical explanations have been proposed to explain semantic diversity's effects. The next section will explore these in more detail.

THEORETICAL APPROACHES TO UNDERSTANDING SEMANTIC DIVERSITY

From the growing literature, semantic diversity appears important to lexical processing in that variation in semantic diversity is associated with variance in task performance, even when other key variables are controlled (Hoffman et al., 2013; Johns et al., 2016a). But as with document count, we might question what these values reflect. Hoffman et al. (2013) proposed semantic diversity to be a continuous measure of polysemy. On this view, semantic diversity captures nuanced differences in word meaning, based on the contexts it has appeared in. Conversely, Jones et al. (2012) proposed that diversity reflects the history of a word's representation across experience, and that it is the updating of representations which leads to advantaged lexical processing. This distinction is important to understanding lexical processing more broadly. How should we interpret semantic diversity's effects? Are they a type of polysemy effect, or is the variation of contexts itself key to lexical organisation? Here, I give a summary of these two theoretical accounts.

THE SEMANTIC DISTINCTIVENESS MODEL

This model proposes that semantic diversity reflects the lexical history a word has acquired across a lifetime of experience. This fits with a broader theory that variability is important to learning and generalisation (for review, see Raviv et al., 2022). According to the Semantic Distinctiveness Model (Jones et al., 2012), words have a 'vector representation' that is continuously updated. When

a word is encountered for the first time, this vector reflects the linguistic context in which the word appears. The next time the word is encountered, the current context will be compared to the existing vector representation, and the vector will be adjusted. Greater updating is required when new contexts are very different to the current vector representation. Over time, after multiple encounters, the lexical representation will come to reflect a word's usage in language in a similar way to semantic diversity values. According to this model, the dynamic variation of contexts is key to how words are encoded into memory and from this the ease with which they are retrieved.

Reviewing the literature, Raviv et al. (2022) discussed the role of variability in learning across domains, including visual perception, motor learning and machine learning. Generally, stimuli presented under conditions of high levels of variability are initially learnt slower than low variability stimuli; but high variability resulted in better generalisation. Raviv et al. argued that different components of variability appear important to learning: numerosity, heterogeneity, situational or contextual, and training schedule. Applied to language processing, we can see how these four components might be captured by semantic diversity. A word with high diversity likely has high numerosity, as it likely has been learnt from a broader number of distinct encounters. These encounters will likely be more heterogeneous, with larger differences between the encounters. By the nature of this variable, high semantic diversity words will have greater situational variability, as the linguistic contexts in which they appear differ. Semantic diversity might also be related to training schedule, in that encounters with high diversity words are likely more randomly spaced over time, whereas low diversity words might be encountered in blocks. To illustrate this using Hoffman et al.'s example of *perjury* and *predicament*, *perjury* (lower diversity) which is constrained to legal contexts might appear many times when an individual is reading a legal text or watching a courtroom drama, but these instances may be few and far between. By comparison, *predicament* (higher diversity) appears in a wider variety of contexts, and this might make encounters with *predicament* more likely to be spaced across time. It may be that some of these four factors

described by Raviv et al. are more relevant to understanding the effects of semantic diversity than others, and it might depend on precisely how the construct is operationalised. The Semantic Distinctiveness Model focuses largely on situational variability, but through this general variability framework, we can begin to understand why semantic diversity has an influence on language learning and lexical organisation.

SEMANTIC DIVERSITY AS A MEASURE OF POLYSEMY

Another way of conceptualising semantic diversity is that it captures a continuous measure of polysemy. Most words in English have more than one meaning. Some words (e.g., *bark*) have distinct meanings which are not semantically related (the sheath of a tree trunk vs the cry of a dog). Other words (e.g., *glass*) have nuanced but related meanings, or 'senses' (the container we drink from vs the material it is made of). We can measure lexical ambiguity in different ways, ranging from homonymous words with distinct meanings (*bark*) to polysemous words that have nuanced, related meanings (*glass*). Hoffman et al. (2013) proposed semantic diversity as a continuous measure of polysemy. In this view, high diversity values reflect large amounts of polysemy; this makes sense as polysemous words need to be more flexible in their contextual use. In order to assess this characterisation of semantic diversity, it is useful to review the similarities between semantic diversity and other measures of ambiguity.

Different forms of ambiguity appear to have differing effects on lexical processing in the literature. Pexman and Lupker (1999) found that distinguishing between homonyms (words which have distinct meanings) and polysemous words (words with nuanced, related meanings) was key to understanding the effects on lexical processing. They found that while polysemous words were

associated with faster response times in lexical decision, homonyms were associated with slower response times.

First, let us review research investigating homonymy; that is, words with multiple distinct, unrelated meanings. Words with multiple meanings have been associated with slower reading times compared to words with one meaning (Piercey & Joordens, 2000; Rodd et al., 2002). Pexman et al. (2004) reported that processing is more difficult for words with multiple meanings when participants have to discriminate between multiple meanings in order to use the appropriate semantic information. This meaning discrimination can be informed by prior linguistic experience. Simpson (1981) found that lexical decision was facilitated for associated words after homographs were primed to their dominant meaning, but not when the same homographs were primed to their subordinate meaning. This suggests that participants relied on semantic knowledge learned through previous encounters with words. As well as this long-term contextual information, studies demonstrate that the current context plays an important role in determining which meaning should be processed. In ambiguity priming studies, Rodd et al. (2013) found that encountering a homophone in a disambiguating context can bias subsequent interpretation of the word. These studies demonstrate the importance of both current contextual information and long-term contextual histories to lexical processing. The effects of ambiguity can be explained by a competition-type account. For homonyms, each meaning is represented as an individual node in a network. During processing, these meanings compete for activation (Kellas et al., 1988). This can explain why words with multiple distinct meanings are read more slowly.

However, ambiguity effects appear different when we consider *polysemy*: the nuanced, related meanings of words. Polysemy can be measured by number of senses. Unlike for homonyms, words with more senses have been associated with faster lexical decision and word naming (Hino & Lupker, 1996). Rodd et al. (2002) has argued that the rich semantic knowledge associated with

polysemous words aids word recognition. The effects of polysemy can be explained by connectionist accounts, in which networks of semantic representations are activated simultaneously. For example, Kawamoto (1993) proposed that more dominant senses will be accessed first, with the difference between access-likelihood for senses increasing with relative frequency. Rodd et al. (2004) proposed that ambiguous words are activated in a blended state representing multiple possible meanings, and that activations are modified by connections between semantic units, which effectively force a shift to access one stable semantic representation. This can further be conceptualised by imagining a network of lexical-semantic knowledge. In a review of semantic space and ambiguity, Rodd (2020) described lexical-semantic knowledge as a complex and high-dimensional space. As words have multiple meanings which are context-dependant, linking a word form to a particular meaning can be highly variable. Rodd described that each meaning will correspond to a stable state within lexical-semantic space. For homonyms, the meanings will correspond to distant points within the space. For polysemous words, these points will be closer in proximity. In order to successfully retrieve a word's meaning, the network must settle on a stable point within lexical-semantic space. The differences in network proximity are proposed to explain the processing differences observed for homonymous and polysemous words.

Armstrong and Plaut (2008) further explored the effects of different types of ambiguity on processing by manipulating lexical decision task difficulty. They did this by manipulating the non-words in each condition: in the 'easy' condition, non-words had few orthographic neighbours and low bigram frequency, but in the 'hard' condition, non-words had large orthographic neighbourhood sizes and high bigram frequency. An advantage was found for both homonymous and polysemous words when the task was low in difficulty. For harder tasks, they found a polysemy advantage and homonymy disadvantage. The researchers argued that the amount of semantic precision required in order to make decisions is key. For example, precise semantic information is needed to decide whether a word belongs to a particular semantic category, while less precise semantic information

can be used to determine whether a letter-string is a word. This results in an ambiguity disadvantage for semantic categorisation tasks, and an advantage for lexical decision.

If semantic diversity is a continuous measure of polysemy, we would expect the effects of semantic diversity on lexical processing to be similar to other polysemy measures. That is, we would expect semantic diversity to be associated with more efficient reading, word naming and lexical decision, but less efficient semantic judgements. This is indeed the pattern of results reported in the literature. Hoffman and Woollams (2015) found that words high in semantic diversity were associated with faster responses in lexical decision and slower responses in semantic relatedness judgements. Semantic diversity effects have also been linked to task difficulty. In a meaning-matching task, both adults and children found it harder to match meanings for high- than low-semantic diversity words (Hsiao et al., 2020). This was particularly true for unmatched trials, which are thought to be harder. Using the same items, children made faster lexical decision times. Based on these results, semantic diversity does appear to have similar effects on lexical processing to other polysemy measures, supporting Hoffman et al.'s proposal of diversity as a continuous measure of polysemy.

Turning to the relationship between semantic diversity and other measures of polysemy, diversity correlates positively with number of senses ($r = .36$, Hoffman et al., 2013). However, this modest correlation is not as high as we might expect it to be if both measures were capturing the same thing. It is important to acknowledge the differences between these values. Semantic diversity is based in co-occurrence, allowing it to capture the nuanced differences in how words are used in contexts. This graded measure allows for more sensitivity than counts of senses or meaning. This arguably leads to a more accurate measure of polysemy than number of senses, which according to Hoffman et al. (2013) can be largely dependent on the subjective ruling of the lexicographer. In this view of semantic diversity, diversity still captures linguistic variation and therefore still relates to the

broad variability discussed by Raviv et al. (2022). However, while Raviv et al. proposed that variability itself facilitates learning, the polysemy account would suggest that this relationship is more indirect. According to this view, semantic diversity gives a measure of variation that reflects polysemy, and it is the polysemy itself which influences lexical processing. However, others have criticised the view of semantic diversity as a measure of ambiguity.

Cevoli et al. (2021) questioned the extent to which semantic diversity captures distinct meanings of ambiguous words. Following Hoffman et al.'s method, they calculated semantic diversity values from the British National Corpus. They used these values to replicate previous studies that investigated the effects of ambiguity on lexical decision. In their simulations, semantic diversity could not account for differences between homonymous, polysemous and unambiguous words. They concluded that semantic diversity instead captures general textual information of the context in which words appear. They therefore argued that *textual diversity* would be a more appropriate name. In response, Hoffman et al. (2021) demonstrated that Cevoli et al.'s result could be attributed to a difference in calculation method. Cevoli et al. used scaled vectors in their latent semantic analysis, while Hoffman et al. argued that unscaled vectors are preferable (see Chapter 2 for further discussion). When Hoffman et al. (2021) repeated Cevoli's simulations using unscaled vectors, they found that the measure was closely related to variations in polysemy. It did not account for differences between homonyms and non-homonyms; therefore, semantic diversity should be viewed as a discrete measure of polysemy, rather than capturing the unique cases of words with multiple unrelated meanings. Note also that Hoffman et al.'s measure was a better fit to lexical decision data than Cevoli et al.'s measure.

Beekhuizen et al. (2021) further investigated the relationship between distributional semantic representations and ambiguity. They explored contextual similarities in the occurrence of three groups of words: monosemes (words with only one sense), polysemes (words with multiple

related senses), and homonyms (words with multiple unrelated meanings). They found a general trend of greater dissimilarity between contexts for homonyms than polysemes, which in turn had greater contextual dissimilarity than monosemes. Therefore, words with only one sense occurred in more uniform contexts than words with multiple senses or meanings. From this, it is apparent that diversity in semantic contexts does capture ambiguity information as well as textual diversity.

RELATIONSHIPS WITH OTHER VARIABLES

So far, I have reviewed research into contextual variation as a lexical variable alongside some discussion of word frequency and polysemy. However, there are many other factors important to lexical organisation. An array of semantic constructs have been proposed to impact lexical processing, including concreteness (Brysbaert et al., 2014), semantic neighbourhood density (Fieder et al., 2019), part of speech (Melinger & Koenig, 2007), and emotional valence (Kensinger & Corkin, 2003). Here, I discuss where semantic diversity sits amongst other meaning-related constructs in lexical-semantic space.

SEMANTIC DIVERSITY AS A COMPONENT OF SEMANTIC RICHNESS

As mentioned previously, many lexical measures are closely related. If a word is high frequency, it is also likely to have been acquired early in life, and to have high levels of flexibility in its contextual usage. It likely has multiple meanings or senses, can function as more than one part of speech, and will be well connected in a semantic network (Hills, 2009; Rodd, 2020). Some of these measures might tap into similar constructs. Pexman et al. (2007) described ‘semantic richness’: a principle reflecting the amount of semantic information that is elicited by a concept or word. They argued that semantic richness can be measured in multiple ways, citing number of features, number

of meanings, and semantic neighbourhood density as examples. Using behavioural data and fMRI, Pexman et al. demonstrated that semantically rich words require less effortful processing.

Pexman et al. (2008) compared the effects of three measures of semantic richness (semantic neighbourhood size, number of features, and contextual dispersion) on lexical decision and semantic categorisation performance. Here, contextual dispersion was taken from Zeno et al. (1995) and reflected the distribution of word occurrences across nine subject areas in a corpus of educational materials. Measured using relative entropy, this is similar to the previously discussed measures of contextual distinctiveness and document count. The researchers found that the three measures each predicted unique variance in task performance. Between lexical and semantic decision tasks, there was a different pattern of effects for each variable. This challenges the assumption that 'richer is always better'. The researchers went on to discuss the differing effects of contextual dispersion. For example, Steyvers & Malmberg (2003) asked participants to study a list of words, and then tested the extent to which participants recognised having studied them. Participants' recognition was better for words that lower contextual variability based on a corpus-derived measure. Pexman et al. argued that the direction of contextual dispersion effects is task-dependant, with tasks that rely on isolating one semantic representation having a processing disadvantage (e.g., semantic categorisation). This representation can be thought of as a particular sense, or as a particular way the word might be used within a phrase. In lexical decision, the processing of multiple different meanings or usages simultaneously leads to a processing benefit, as all representations share the same form. But in semantic judgement tasks, one representation must be isolated in order for the participant to make the decision. In this case, the other meanings or usages may make the decision more effortful. This is the same result reported by Hoffman and Woollams (2015), who found that semantic diversity was associated with faster lexical decision but slowed semantic decision. Extending this explanation to Steyvers and Malmberg's (2003) findings, episodic memory might rely on semantic information.

Yap et al. (2011) further built upon this work, reporting a distinction between 'control' lexical variables in the context of their study (e.g., word frequency, number of letters, phonological neighbourhood size) and semantic variables. While control variables accounted for more variance in lexical decision and speeded pronunciation, semantic variables accounted for more variance in semantic classification. Within the semantic variables, there were also task-dependant differences in the direction of effects. Words with multiple senses were associated with faster lexical decision response times, but less accurate semantic categorisations.

In teasing apart the different components of semantic richness measures, Vigliocco et al. (2009) drew a distinction between two types of semantic representation: experiential information and linguistic information. Experiential information involves how we interact with the word concept, and includes sensory, motor, and affective information. Linguistic information is derived from the linguistic context in which words appear and is based in co-occurrence patterns and syntactic structures. The researchers applied this distinction to explain differences between concrete words (high in experiential information) and abstract words (high in linguistic information). Taken further, this view is useful for understanding the complex relationships between the multifaceted measures of semantic richness. It is also helpful in framing semantic diversity. Given how this is calculated (e.g., Jones et al., 2012; Hoffman et al., 2013), it is based in linguistic experience and aims to capture how words are distributed across contexts. It might therefore capture different information to variables that aim to capture sensory, motor, and affective experience.

DIFFERENCES BETWEEN SEMANTIC MEASURES

Semantic richness is considered to be a multifaceted construct reflecting the meaning-related information associated with words (Pexman et al., 2007). However, there are some clear differences between semantic richness measures. One way of teasing apart semantic variability is to

use factor analysis. Hoffman et al. (2013) performed a principal components analysis on a range of lexical and semantic variables. From this, they identified a two-factor structure. Factor 1 (which the researchers described as 'frequency') saw positive loadings from semantic diversity, word frequency, document count, number of senses and contextual distinctiveness; while Factor 2 (described as 'abstractness') saw positive loadings from concreteness, imageability and context availability, and negative loadings from semantic diversity. This then shows that there are (at least) two types of semantic constructs within semantic richness.

The cluster of variables identified by Factor 2 in Hoffman et al.'s analysis appears to capture a different facet of semantic information than measures of polysemy and contextual variability. These relate to the mental imagery associated with a word. Examples include concreteness, imageability, context availability, and body-object interaction. Unlike semantic diversity, imagery and embodiment variables are typically based on subjective ratings by participants. Imageability is rated on a Likert scale for the extent to which a word arouses a mental image (Cortese & Fugett, 2004). Concreteness measures the extent to which participants rate a word as being concrete or abstract (Brysbaert et al., 2014; Muraki et al., 2022). To obtain a measure of context availability, participants rate the extent to which a word can be associated with a context (Altarriba et al., 1999). Body-object interaction ratings capture how easily the referent of a word can be physically interacted with (Tillotson et al., 2008).

Concreteness and imageability are closely related (r = .75, Hoffman et al., 2013). Close relationships have also been reported between concreteness and body-object interaction (r = .75, Pexman et al., 2019), imageability and body-object interaction (r = .7, Pexman et al., 2019), and between context availability and concreteness (r = .69, Schwanenflugel et al., 1988). These measures are all associated with performance, with studies reporting more efficient lexical access for words

that are concrete (Walker & Hulme, 1999), highly imageable (Strain et al., 1999), easy to place in a context (Columbo & Burani, 2002), and easily interacted with (Siakaluk et al., 2008).

The correlation between semantic diversity and concreteness is negative (r = -.48), and for imageability too (r = -.26; Hoffman et al., 2013). This means that words which appear in a wide variety of contexts are more likely to be abstract and harder to associate with an image. Regarding the relationship between semantic diversity and context availability, there have been mixed findings. Hoffman et al. (2013) reported a negative correlation (r = -.26) based on 279 words, while Moffat et al. (2014) reported a positive correlation (r = .36) for abstract words (N = 200) and no correlation (r = 0.00) for concrete words (N = 200). A larger itemset is needed to determine the relationship between semantic diversity and context availability.

Importantly, in Hoffman et al.'s (2013) principal components analysis, semantic diversity loaded onto both factors, but in opposite directions. It loaded positively onto Factor 1 alongside positive loadings from frequency and polysemy measures; and it loaded negatively onto Factor 2 alongside positive loadings from concreteness measures. To some extent, therefore, semantic diversity measures capture how abstract a word is. This pattern can also be seen in behavioural studies reporting the different effects semantic diversity exerts on processing compared to concreteness. As discussed, words high in both concreteness and semantic diversity typically enjoy a benefit in lexical processing tasks (Hoffman & Woollams, 2015). However, this is not true across all tasks. While concreteness has been associated with efficient processing in semantic judgement tasks (Pexman et al., 2007), Hoffman and Woollams (2015) found that participants made slower judgements for high diversity words. This could be in part because semantic diversity is capturing a word's abstractness.

The effects of semantic variables on lexical processing have also been investigated using event-related potentials (ERP). ERP allows for exploration of the neural activity associated with

performing behavioural tasks. Vergara-Martinez et al. (2017) found that ERP activity associated with contextual diversity was in the opposite direction of activity typically associated with word frequency. From this, the researchers concluded that the processing of frequency and contextual diversity is different. The pattern of activity observed in response to high vs low contextual diversity words was similar to that of semantic richness, supporting the idea that contextual diversity relates to a word's semantic properties. The researchers argued that both concreteness and diversity were associated with similar neural processes, but that the ERP activity within this processing was distinct. Vergara-Martinez et al. explained the effect of semantic richness by suggesting that long-term memory is enhanced for richer semantic representations. Therefore, words associated with more diverse contexts will be associated with more activity.

Differences have been found between concrete and abstract word processing, and this might link to the brain's organisation of semantic memory. Adorni and Proverbio (2012) found that abstract words were associated with prefrontal activity (similar to that of contextual diversity) while concrete words were associated with visual regions. The authors argue that these regions form a network that regulates semantic knowledge. As semantic diversity negatively correlates with concreteness and imageability (Hoffman et al., 2013), these concreteness effects could explain the importance of pre-frontal activity in processing contextually diverse words. Concrete words may recruit visual imagery processes, while abstract words (typically contextually diverse) rely on the top-down control of semantic memory that has been associated with prefrontal brain areas (Martin, 2007). However, Vergara-Martinez et al.'s experiment matched words for concreteness and imageability, meaning that their findings cannot be attributed to the concreteness effects previously found. Rather, contextual diversity appears to be independent of concreteness.

SUMMARY

The contexts in which words appear over time is important for acquiring and retrieving these words in the mental lexicon. Corpus-based semantic diversity measures capture this contextual flexibility, with multiple studies demonstrating that such metrics explain significant variance in lexical processing. Research from language processing studies demonstrates that there is often a processing benefit for words that have diverse contextual histories; however, due to competition between multiple semantic representations, these words may be associated with slower processing during tasks that rely on semantic information. Semantic diversity is positively correlated with measures of ambiguity and semantic connectivity; and negatively correlated with concreteness and imageability. It can be seen as one part of a multifaceted semantic richness, but research is yet to investigate the relationships between these semantic variables in a large itemset.

Human memory is proposed to be organised to prioritise information most likely to be needed in the future (Anderson & Milson, 1989). According to this theory, items are more accessible if they have been retrieved before. This might also be relevant in explaining the effects of semantic diversity. As noted by Johns et al. (2016a), in planning for an unknown future context, there is more likelihood of needing words that have been retrieved in multiple diverse contexts than words that only appear in limited contexts. As a result, diverse words will be more readily accessible. This is also shown in the Semantic Distinctiveness Model (Jones et al., 2012), in which words that appear in multiple, diverse contexts have higher lexical strength. However, as semantically diverse words also have high levels of polysemy, these words experience a processing disadvantage when the task requires a high level of semantic precision. The two constructs captured by diversity measurements (contextual histories and polysemy) lead to two theoretical explanations of semantic diversity's effects. On the one hand, Jones et al. (2012) propose that high diversity words are associated with speeded lexical decision because of the rich lexical history associated with a word, produced by updated representations across multiple varied encounters. On the other hand, Hoffman et al. (2013) propose that high diversity words are associated with speeded lexical decision because the

contextual variation on which diversity is based gives a continuous measure of polysemy, and it is this polysemy that drives diversity's effects.

THE AIMS OF THIS THESIS

A central aim of this thesis is to develop our understanding of the complex relationships between lexical variables which correlate with each other. In particular, I have focused on semantic measures. In Chapter 2, I discuss the relationships between different measures of contextual variation previously proposed in the literature. I investigate correlations between two different measures of semantic diversity (both calculated using latent semantic analysis), and two variations of semantic distinctiveness, as derived from the Semantic Distinctiveness Model (Jones et al., 2012). I compare the item-level correlations between these variables and other semantic measures. I then explore how these measures relate to performance by testing the fit of the semantic diversity measures to lexical decision data. Next, I use structural equation modelling to investigate the extent to which semantic diversity can be explained by other lexical variables. This builds upon the work of Hoffman et al. (2013), who explored semantic diversity's place within lexical-semantic space using principal components analysis. I test the fit of a model based on Hoffman et al.'s factor structure to investigate what proportion of variance in semantic diversity values is unique, and what proportion can be explained by other lexical and semantic measures. This helps us to understand what semantic diversity measures capture.

Chapter 3 introduces a new set of ratings for two variables: context availability (referring to how easily a word can be placed within a context) and sentence availability (a new variable reflecting the ease with which words are placed in sentences). Context availability facilitates word naming and lexical decision, and its effects are not readily explained by other variables (Schwanenflugel et al., 1988; Schwanenflugel & Noyes, 1996; Colombo & Burani, 2002). Despite this, the largest set of

context availability norms openly available prior to our study was the 325 words collected by Altarriba et al. (1999). The initial motivation for collecting more availability ratings was to test whether Hoffman et al.'s principal components analysis results would replicate in a larger itemset in order to assess semantic diversity's place in lexical-semantic space. Hoffman et al.'s analysis was limited to 279 words, primarily due to the limited number of words which had been normed for context availability. Collecting ratings for more words allows us to replicate Hoffman et al.'s analysis with a much larger itemset. We also explore the nature of sentence availability, proposing that this measure might relate to the textual information captured by semantic diversity. We investigate relationships between lexical and semantic variables using correlations and factor analysis.

Another aim of this thesis is to investigate the effects of recent experience on subsequent lexical processing. As discussed, measures like frequency and semantic diversity capture a word's lexical history, and therefore reflect long-term linguistic experience. We aim to determine how this interacts with more recent linguistic experience. This work is situated within the framework of MacDonald's (2013) Production-Distribution-Comprehension hypothesis, in which recent experience makes words more readily available. Chapter 4 begins by reviewing this hypothesis before testing whether encountering a word in a production task will affect subsequent comprehension of the same word, and vice versa. Words in this experiment vary in sentence availability, allowing us to also test the effects of lexical history on linguistic processing.

Chapter 5 also investigates the effects of lexical history and recent experience, using words that vary in imageability and semantic diversity. This experiment builds upon studies reporting the task-dependant effects of semantic diversity on lexical behaviour. Hoffman and Woollams (2015) reported that high semantic diversity values were associated with faster performance in lexical decision, but slower performance during semantic relatedness judgements. High imageability values were associated with faster performance in both tasks. In Chapter 5, participants encounter words in

either a lexical decision task or a semantic decision task. Then, after a distractor task, participants make lexical decision responses to the same words. Our motivation in this experiment is twofold. First, we aim to replicate Hoffman and Woollams' (2015) results for the effects of imageability and semantic diversity on lexical and semantic decision, with our analysis focusing on the first task that participants took part in. Second, we test whether these two types of processing exert different effects on subsequent lexical decisions. This allows us to explore how different types of recent lexical experience shape subsequent processing.

Together, this thesis aims to further our understanding of semantic diversity and its relation to other semantic measures. It aims to determine how recent linguistic experience interacts with long-term lexical histories. These results should help us to understand why some words are easier to process than others. This contributes to the literature investigating the mechanisms involved in word processing.

CHAPTER 2
COMPARING MEASURES OF SEMANTIC DIVERSITY

INTRODUCTION

Semantic diversity is a lexical characteristic which quantifies the contextual information a word is associated with across multiple linguistic encounters. This measure has been found to affect lexical decision performance (Hoffman et al., 2013) and word learning (Jones et al., 2012), and captures more variance in lexical processing than frequency (Brysbaert et al., 2018).

Semantic diversity has been conceptualised and measured in different ways. Theoretically, there are two explanations of why diversity affects processing, as discussed in Chapter 1. First, as proposed by Jones et al.'s (2012) Semantic Distinctiveness Model, semantic diversity reflects the contextual experience that a word has accrued via a lifetime of experience. Second, as proposed by Hoffman et al. (2013), semantic diversity can be seen as a measure of polysemy which captures the many nuanced meanings words activate when experienced in different contexts. Both these explanations predict similar effects on lexical processing, but for different reasons. According to Jones et al., diversity is associated with better lexical decision because high diversity words are associated with stronger lexical histories due to the multiple, varied contexts in which they have appeared. According to Hoffman et al., high diversity also reflects varied contextual usage, but the variation itself is not what drives processing. Rather, the variation is a product of polysemy, and it is this polysemy that explains the differences in processing of high and low diversity words. Alongside these conceptual differences, there are also large differences in how these measures are calculated. This chapter reviews differences between semantic diversity measures in the literature. It then compares four diversity measures (Hoffman et al., 2013; Cevoli et al., 2021; Johns et al., 2020; and Johns, 2021) using correlations, linear mixed effect models and structural equation modelling. These

methods allow us to investigate the extent to which diversity measures differ, and how they relate to lexical-semantic space. This aims to further our understanding of what semantic diversity measures capture, which is important when applying semantic diversity's effects on lexical processing to theories of lexical organisation.

Semantic diversity was proposed as a measure of polysemy by Hoffman et al. (2013). In calculating values for diversity, the researchers used latent semantic analysis to give a statistic ranging from 0.08 (occurring in uniform contexts) to 2.41 (occurring in diverse contexts). To give an example from their paper, the words *perjury* and *predicament* both have one core meaning yet differ greatly on the variability of the contexts in which they occur: *perjury* tends to be restricted to legal contexts, whereas *predicament* could appear in a wide range of contexts (e.g., being in a predicament because you have accepted two dinner invitations on the same night; or having a financial predicament due to low funds). This differs from other polysemy metrics like number of senses (often based on dictionary definitions) as it gives a continuous measure which considers nuanced differences in context.

Hoffman et al.'s (2013) measure was calculated by splitting the British National Corpus (BNC) into contexts of 1,000 words. Latent semantic analysis was then performed on these passages to give vectors for words and for contexts (see Chapter 1 for further explanation). This created a matrix, which was log-transformed and weighted by word entropy using the equation:

$$Entropy = \sum_{context} p_c \log(p_c)$$

where p_c is a word's frequency within a context divided by its overall frequency. Singular value decomposition was used to reveal latent relationships between words based on co-occurrence. Cosines between vectors were used to compute similarity. For each word, the mean cosine was log-transformed and reversed to give a semantic diversity value. As such, Hoffman et al.'s measure is not

set up to capture polysemy specifically, but rather captures contextual similarity which is proposed to give a gradient of linguistic flexibility which reflects polysemy.

Studies have employed Hoffman et al.'s diversity measure to observe the effects of meaning-related variance on lexical processing. Hoffman et al. (2014) found that patients with semantic dementia were less likely to use words with low semantic diversity, therefore demonstrating the importance of contextual information in the organisation of the mental lexicon. Hoffman and Woollams (2015) found that diversity is associated with faster performance in lexical decision, and slower performance in semantic relatedness decision. This means that words which typically appear in diverse contexts are more easily processed than words that typically appear in uniforms contexts when the task requires reflection on word form; but when a task requires reflection on word meaning, words that typically appear in less diverse contexts are easier to process. This finding distinguishes diversity from other measures of semantic richness like concreteness or imageability, which are associated with faster response times for both lexical and semantic processing.

According to connectionist models (e.g., Kawamoto, 1993; Rodd et al., 2004), when words have multiple semantic representations (i.e., words that are associated with multiple facets of meaning-based information), these words are activated simultaneously, weighted by the dominance of a particular meaning or usage (see Chapter 1 for review). Polysemous words will therefore be more likely to activate multiple meanings and receive a processing benefit in word naming and lexical decision (Hino & Lupker, 1996). They are also associated with slower semantic processing (Pexman et al., 2004), as this relies on isolating a particular semantic representation. Semantic diversity mirrors these findings, with high diversity words processed faster in lexical decision but slower in semantic decision (Hoffman & Woollams, 2015). This supports the idea that diversity captures the rich and nuanced meanings a word has in a similar way to other measures of polysemy.

Another set of diversity statistics calculated using latent semantic analysis comes from Cevoli et al. (2021). Cevoli et al. followed Hoffman's methodology using a lemmatised version of the same corpus. Unlike Hoffman et al., they used scaled singular values, which they argued were a key feature of latent semantic analysis. This means that, in every vector, each element was multiplied by its singular value. Cosines were the computed from scaled rather than original vectors. The result of scaling is that dimensions with high singular values are weighted more heavily (Hoffman et al., 2021), which can reduce noise from small terms whilst still accounting for distinct documents (Ando, 2000). Cevoli et al.'s norms had a correlation of 0.73 with Hoffman's measure, demonstrating that small differences in the calculation method do change diversity statistics. Cevoli et al. went on to test how semantic diversity captures differences between homonymous, polysemous and unambiguous words. In previous studies, homonyms were associated with slowed reading times, while polysemous words were associated with faster reading (Rodd et al., 2002). Cevoli et al. tested whether diversity statistics could explain differences in the stimuli of Rodd et al. (2002) or Armstrong and Plaut (2016), two studies investigating the effects of semantic ambiguity on lexical processing. They found that semantic diversity could not account for differences between ambiguity-class in the stimuli. Their semantic diversity measure was not able to capture the distinct meanings of ambiguous words, with no difference in semantic diversity values found between words with many meanings and senses compared to words with few meanings or senses. Rather, semantic diversity appeared to capture general information about the contexts in which words appeared. From this, Cevoli et al. argued that semantic diversity measures a word's spread across linguistic contexts, rather than diversity in meaning. They concluded that 'textual diversity' is a more appropriate term for semantic diversity.

In response to Cevoli et al., Hoffman et al. (2021) presented the decision to scale values as an empirical choice rather than an analytical one. They argued that scaling is important to reconstruct a data matrix, but this importance does not necessarily extend to computations of

semantic similarity. Hoffman et al. further demonstrated that the use of unscaled vectors was preferable, as this provided a better fit to human semantic judgements. When replicating Cevoli et al.'s analysis, Hoffman et al. found that their own measure was a better fit for word recognition performance and better captured contextual variability. Their measure also captured polysemy, with semantic diversity reliably covarying with polysemy in both the stimuli sets of Rodd et al. (2002) and Armstrong and Plaut (2016). This suggests that, supporting Hoffman et al.'s (2013) proposition, semantic diversity can be seen as a continuous measure of lexical polysemy – though the researchers acknowledged that this variable might not capture the distinct meanings of homonymous words.

Previous research calls into question how semantic diversity should be conceptualised. Semantic diversity could reflect a word's linguistic dispersion across contexts (according to Cevoli et al.); or it could additionally reflect the semantic usage of the words, as argued by Hoffman et al. While related, these two conceptualisations have different implications. Cevoli et al.'s 'textual diversity' might be more theoretically aligned with other constructs of linguistic variation, like frequency and document count. On the other hand, Hoffman's 'semantic diversity' might appear more similar to constructs of ambiguity, like number of senses. Correlations between these two diversity values and other lexical variables have not been compared.

Semantic diversity has also been described within the framework of the Semantic Distinctiveness Model (e.g., Jones et al., 2012; Johns et al., 2016a). According to this model, representations of words in memory are updated and developed with contextual experience. Semantic distinctiveness was not calculated using latent semantic analysis, as Jones et al. (2012) described how word-document matrices used in LSA (see Figure 1.2) may not reflect lexical access so well as weighted counts of semantic distinctiveness. Rather, semantic distinctives considers a word's distinctive use by capturing its lexical strength, based on a measure of linguistic dissimilarity. Each

word is assigned a vector representation, which can be thought of as the accrued contextual information associated with that word. The first time a word is encountered, this vector will be the context in which it appears. When a word is encountered in a different context to previous encounters, the word's vector is updated to include this information, and this increases the word's lexical strength. The more distinct a new context is from the prior representation, the more updating is required. Calculations of semantic distinctiveness therefore consider the diversity of contexts by calculating dissimilarity between documents a word appears in. As reported by Jones et al. (2012), similarity is the intersection of two contexts, divided by the smaller of the contexts. This is then subtracted from 1 to give a measure of dissimilarity, as follows:

$$Dissimilarity(doc_i, doc_j) = 1 - \frac{|doc_i \cap doc_j|}{\min\{doc_i, doc_j\}}$$

where doc_i and doc_j refer to two contexts in a word appears. Dissimilarity is then used to give a measure of lexical strength, which is boosted when words appear in contexts distinct from their prior uses. This gives a semantic distinctiveness value, which can be seen as the sum of the word's vector elements, based on contexts.

$$Semantic\ Distinctiveness = \sum_{i=1}^{n}\sum_{j=1}^{i} quantile\left(Dissimilarity(doc_i, doc_j)\right)$$

As such, high semantic distinctiveness values are assigned to words which appear in more semantically distinct contexts as opposed to redundant or predictable contexts. This is therefore similar to Hoffman et al.'s construct, in that both diversity measures capture flexibility of a word's contextual use based on contextual similarity. However, while Hoffman et al. (2013) propose that semantic diversity effects are due to the aggregated experience that results in a discrete measure of

polysemy, Jones et al. (2012) suggest that these effects occur due to the dynamic updating of lexical histories.

Alongside these theoretical differences, there are key differences in how diversity is calculated. One such difference is the corpora from which the measures are based. Hoffman et al. and Cevoli et al. both used variations of the British National Corpus. Measures based on the Semantic Distinctiveness Model have used different corpora. For example, Johns et al. (2020) and Johns (2021) both used language samples much larger than the BNC, which reflected different types of language use.

Johns et al. (2020) explored a corpus of fiction and non-fiction books. Two types of contextual diversity were initially computed: 'book prevalence', counting the number of books that a word appeared in; and 'author prevalence', counting the number of authors who used a word. Book prevalence closely resembled traditional measurements of document count, with each book treated as one document. Author prevalence shared similarities with word prevalence (Brysbaert et al., 2017), a variable that captures the number of people in a population that know the word. Words scoring highly for author prevalence were used by a large number of authors, therefore giving some indication of how well known it was. These measures were then transformed using a semantic diversity modification, calculated using the method outlined by Jones et al.'s (2012) Semantic Distinctiveness Model. This gave 'book prevalence diversity', which reflected the uniqueness of word usage across books, and 'author prevalence diversity', which reflected the uniqueness of usage across authors. Higher diversity values indicated larger variability between contexts.

Book prevalence, author prevalence, and the two semantic diversity modifications of these measures were fitted to lexical decision data from the British Lexicon Project (BLP; Keuleers et al., 2012) and the English Lexicon Project (ELP; Balota et al., 2007), and naming data from the ELP. All measures explained more variance in the data than word frequency, with 'author prevalence

diversity' predicting most variance overall. However, as Johns (2021) cautioned, the language within books might not be a valid reflection of linguistic experience: it is written, not spoken; and authors can be viewed as 'experts' of language use and therefore different to the average speaker. These limitations led Johns (2021) to take contextual and semantic diversity measurements from the social media platform Reddit. From a corpus of over 55 billion words produced by 334,345 users, two measures of contextual diversity were derived: 'discourse contextual diversity' (the number of semantic discourses or 'subreddits' a word appeared in) and 'user contextual diversity' (the number of individual accounts that used a word). These measures provided a better fit for lexical decision, naming, and word prevalence data compared to word frequency and document count; but no conclusive difference in fit was observed between discourse and user contextual diversity. Semantic diversity transformations were then performed on both measures, using two methods of updating the vector representation: 'word representation' calculated a vector of word count used by a user or within a discourse; and 'population representation' calculated a vector based on the distributional properties of discourses and users across the corpus. The measure most successful in explaining behavioural data was UCD-SD-PR (user contextual diversity, modified with population representation semantic diversity). From this, the researchers argued that lexical organisation is dynamic and uses both social and discourse informational cues to optimise according to context.

The measures collected by Johns et al. (2020) and Johns (2021) demonstrate that the impact of contextual information operates on a much larger scale than previously explored. Previous measures of diversity have defined 'context' as a narrower window of words. For example, contextual distinctiveness (McDonald & Shillcock, 2001) measured similarity between contexts in a five-word window. When Jones et al. (2012) calculated semantic distinctiveness values, the documents which constituted as a linguistic context had an approximate mean length of 300 words. Hoffman et al.'s (2013) measure can be seen to widen this window, with diversity calculated from contexts on 1,000 words. However, Johns et al.'s (2020) and Johns' (2021) measures operate on a

much larger contextual scale. Books in Johns et al.'s (2020) corpus contained an average of 6,000 unique words, while subreddit discourses explored by Johns (2021) contained an average of 1,800,000 words. These measures therefore capture the broad contextual information surrounding a word, not just the information within close proximity to the word. This broad contextual information might operate differently to the closer information captured by Hoffman et al.'s (2013) and Cevoli et al.'s (2020) measures.

Regardless of how it is defined or operationalised, semantic diversity effects remain significant above and beyond frequency and document count when calculated based on broad, widespread contextual discourses (Johns, 2021). These findings have important implications for our understanding of lexical organisation. They suggest that widespread linguistic co-occurrences are important to building lexical knowledge; not only the narrow window of context occurs alongside words, or frequency of occurrence alone. A question remains as to whether broad contextual information is different to that collected in a narrower, specified context, such as the 1,000-word contexts explored by Hoffman et al. If Johns' (2021) measure is a better fit to lexical decision data than those collected by Hoffman et al. and Cevoli et al., we might conclude that a broad network of semantic connectivity influences lexical processing. Conversely, if Hoffman et al.'s and Cevoli et al.'s measures provided a better fit, this might signal that the specific, close network of semantic connections is what induces effects on lexical processing. In addition to Johns et al.'s (2020) book prevalence and author prevalence diversity measures, the authors also reported a semantic diversity statistic based on the Semantic Distinctiveness Model. This measure is theoretically similar to other diversity values reported in the paper but was calculated using a window of 20 words, following the method outlined by Jones et al. (2012). By comparing this metric to Johns (2021), we can begin to tease apart the differences between these measures. If the narrow-window semantic distinctiveness measure calculated by Johns et al. (2020) and the broad-context metric of Johns (2021) explain similar variance in lexical processing, we might attribute this to the theoretical nature of these

variables. If Johns (2021) is a better predictor than Johns et al. (2020), we might conclude that broad context is preferable.

Differences in the size and content of the corpus might also be important. Both Hoffman et al. and Cevoli et al. used the British National Corpus (BNC). The BNC is a collection of written and spoken texts from a wide array of sources including newspaper articles, academic and fiction books, essays, and transcripts of conversations. This has over 100 million words from over 4,000 texts. By comparison, Johns et al.'s (2020) corpus contained 2.1 billion words from 26,000 fiction and non-fiction books; and Johns (2021) used a corpus with over 55 billion words from 30,000 subreddits. The extent to which corpus size influences semantic diversity's fit to lexical processing data is unclear. However, evidence from word frequency suggests that a large corpus is important in calculating occurrence counts for low- and medium-frequency words (Burgess & Livesay, 1998). The same might also be true of semantic diversity.

The sources from which these corpora are based are likely to influence diversity statistics. For example, written language is more lexically diverse and syntactically complex than spoken language (Roland et al., 2007; Dawson et al., 2021). While the BNC contains transcripts of spoken language, this only makes up 10% of the corpus. It also contains academic texts that might not reflect day-to-day language use, thus not producing a reliable reflection of the average person's linguistic experience. Johns et al.'s (2020) corpus was produced by authors who can be seen as experts of language. While Johns (2021) used a corpus based on written language, this differed from other corpora: first, its producers were not published authors, meaning that they were likely a more diverse sample and thus more representative of the population at large; and second, the producers were conversing with other users online rather than writing for publication. Different corpora provide different fits to lexical processing measures, suggesting that some samples of language might be more akin to typical language experience than others. For example, word frequency

derived from SUBTLEX-UK (a corpus based on UK TV subtitles) provides a better fit to lexical decision data than frequency measures derived from written corpora (Van Heuvan et al., 2014). Roland et al. (2007) further discussed these differences in exploring structural frequencies of different corpora. They noted that there are different registers of language, with scripted speech having structural differences to spontaneous speech. While Johns' (2021) Reddit corpus is based on written language, it likely reflects a more spontaneous and unplanned register of language compared to corpora based on published texts. As most of our language knowledge comes from spontaneous, unplanned encounters, Johns' (2021) measure might therefore provide a better fit to lexical processing data, as it is most likely to reflect everyday language experience.

In the literature, direct comparison of the measures of diversity discussed is limited. Item-level analyses are needed to determine whether words that score highly on one type of semantic diversity also score highly on other measures of diversity. Cevoli et al. reported a correlation of .73 between their 'textual diversity' and Hoffman et al.'s semantic diversity, with differences likely arising due to the scaling of vectors. To our knowledge, Johns et al.'s and Johns' measures of 'semantic distinctiveness' have not yet been compared to other types of diversity.

We aimed to determine the extent to which different measures of semantic diversity are similar. To do this, we investigated the item-level relationships between different semantic diversity metrics taken from Hoffman et al. (2013), Cevoli et al. (2020), Johns et al. (2020) and Johns (2021). We aimed to determine the relationships between the measures of diversity, and whether the measures would pattern similarly in their correlations with other lexical variables. As all diversity measures are based on calculations of contextual similarity, we predicted that all four measures would positively correlate with each other. We also predicted that the four measures would pattern similarly in their correlations with other lexical variables at the item-level. However, due to

theoretical and methodological differences, we expected some differences between the strength of these correlations.

We then tested the fit of these measures to behaviour, using lexical decision data from the BLP and ELP, two large datasets from British and American speakers of English. We tested whether each diversity measure would explain variance in lexical decision accuracy and response times when fit separately. In line with past research, we predicted that each measure would explain a significant proportion of variance after other key lexical variables were controlled for. We then compared each model to a base model (without semantic diversity) to assess the extent to which diversity improved the model fit. This allowed for a comparison of the effects of different measures of diversity on lexical processing. Together with item-level correlations, these analyses aimed to reveal the extent to which different measures of diversity are tapping the same underlying construct.

In addition to comparing different types of diversity, we tested how much variance in semantic diversity could be explained by other lexical measures. This is important for our theoretical understanding of semantic diversity's effects. Here, we employed a similar approach to Baayen (2010). As outlined in Chapter 1, Baayen explored how much variance in word frequency could be explained by other lexical variables. His aim was to determine whether frequency effects reflect the repetition of words or capture the summed co-occurrence of other lexical variables. His analyses found that multiple morphological and lexical characteristics predicted variance in word frequency, including morphological family size, inflectional and syntactic entropy, noun-to-verb ratio, contextual diversity and corpus dispersion. Further, Bayaan analysed the structure of lexical variables using principal components analysis and then fitted these loadings to lexical decision data. The factor with highest loadings from repetition frequency was not the best fit for lexical decision data; rather, variables that captured word context were the best predictors. From this, Bayaan

argued that learning from past contextual experiences is more important to lexical processing than repetition alone.

Semantic diversity was not included in Bayaan's analysis, but a similar question can be asked about the effects of this variable. Semantic diversity effects might be driven by the summed effects of co-occurring variables, like document count and number of senses. Or, semantic diversity could reflect flexibility in a word's linguistic variation and semantic use. This is important for understanding why diversity effects occur, and more broadly, how words are organised in the lexicon. We tested how well semantic variables could explain variability in two measures of diversity using structural equation modelling. Hoffman et al.'s measure was chosen due to its prevalence in the literature, and to remain consistent with the principal components analysis on which our model was based. This was then compared to Johns' (2021) user-based semantic diversity measure. As semantic diversity has been reported to correlate with the measures included in our analysis, we hypothesised that the model would explain some variance in both measures of diversity. However, as we expected diversity to uniquely capture words' contextual histories, we predicted that a larger proportion of variance would be left unexplained.

STUDY 1: EXPLORING MEASURES OF SEMANTIC DIVERSITY

Using correlations, we explored the relationships between four measures of semantic diversity, and their relationships to other variables. We predicted that all diversity measures would be positively correlated with each other. Due to their similarities in calculation method and the corpus used, we predicted that strongest relationships would be seen between that Hoffman et al.'s and Cevoli et al.'s measures; and between Johns et al.'s (2020) and Johns' (2021) measures. We also

explored correlations between these measures and frequency, age of acquisition, and a range of semantic variables; namely, number of senses, document count, concreteness and imageability.

According to Hoffman et al. (2013), semantic diversity correlates positively with word frequency (r = .49), document count (r = .56) and number of senses (r = .36). It correlates negatively with concreteness (r = -.51) and imageability (r = -.48). Hsiao and Nation (2018) reported a weak negative correlation between age of acquisition and Hoffman's semantic diversity (r = -.1), but this relationship was stronger for their diversity calculation based on the Oxford Children's Corpus (r = -.4). Semantic diversity therefore seems to share similarities with measures capturing occurrence and polysemy information but differs from variables reflecting the experiential component semantic richness. According to these correlations, words high in diversity typically have frequent occurrence, appear in a wide range of documents, are learnt early in life and have numerous senses. They are also typically abstract and are not easily associated with imagery. As semantic diversity measures are all based upon the similarity of contexts in which words appear, we expected the broad pattern of results to be the same across diversity measures. However, as there are key theoretical and methodological differences between the measures, we expected there to be some differentiation between them. Identifying where these differences occur should inform our knowledge of what semantic diversity measures capture.

We also investigated how each of these measures fit lexical decision data. We used linear mixed effect models to fit each semantic diversity measure discussed to lexical decision data taken from the BLP and ELP. Our aim was to replicate previous results demonstrating that each of the measures explains unique variance in lexical decision when other key lexical variables (e.g., frequency, age of acquisition) are controlled for. We then compared the fit of each diversity model to a base model using analysis of variance in order to determine differences between measures of semantic diversity.

MEASURES

We compared four measures of semantic diversity, as summarised in Table 2.1:

Hoffman, 2013 (N words = 31,739). This measure was computed using latent semantic analysis on the BNC, as outlined in Chapter 1. Each document within the corpus was divided into contexts of 1,000 words. Analysis computed vectors for words and contexts. The similarity of contexts in which any given word appeared were then compared. This gave a semantic diversity statistic for each word, reflecting the diversity of contexts in which words appear.

Cevoli, 2021 (N words = 28,555). Cevoli et al. followed similar methodology to Hoffman et al., with two key differences. First, they used a lemmatised version of the BNC whereas Hoffman et al. used an inflected version of the same corpus. Cevoli et al. also used scaled singular values in their latent semantic analysis. By comparison, Hoffman et al. used unscaled values.

Johns, 2020 (N words = 78,033). We included the measure of variation derived from the Semantic Distinctiveness Model, reported by Johns et al. (2020). This measure built upon the semantic distinctiveness count first proposed by Jones et al. (2012) and tested by Johns et al. (2016a). In calculating this measure, Johns et al. divided the intersection between two words by the size of the smaller document. This similarity value was deducted from 1 to give dissimilarity of contexts in which a word appeared. Dissimilarity was standardised into a quantile, and the sum of this quantile gave each word's semantic distinctiveness count.

The 2020 measure was based on a corpus of over 26,000 fiction and non-fiction books from 3,200 authors. Unlike 'book prevalence diversity' and 'author prevalence diversity' from the same paper (described in the introduction), diversity was calculated based on a moving window of 20 sentences. Each word was assigned a vector which was updated with every context in which the

word appeared. The strength of this update was dependant on how redundant a word was, based on previous contextual experience: the less similarity between the new context and the prior contextual vector, the more distinctive a word becomes.

Johns, 2021 (N words = 80,346). We included the measure 'UCD-SR-PR' from Johns (2021). This is user-based contextual diversity, with a semantic diversity modification using population representation. This was derived from Reddit data, with over 30,000 semantic discourses (or subreddits) and 334,000 users. The number of individuals who used a word was counted. The semantic diversity transformation took into account distributional properties of individuals and discourse, allowing for a distinction to be made in lexical strength. Words with high lexical strength were those used by large numbers of users and were not specific to particular contexts. A key distinction between this measure of diversity and previous measures is that it considers how words are used across different individuals, not just different texts.

Other lexical variables. Alongside these measures of diversity, we considered Zipf frequency based on the SUBTLEX UK corpus (Van Heuven et al., 2014), document count (Hoffman et al., 2013), age of acquisition (Kuperman et al., 2012), concreteness (Brysbaert et al., 2014), imageability (Cortese & Fugett, 2004), and number of senses (Wordsmyth; www.wordsmyth.net). These variables were chosen as they are associated with lexical processing performance (e.g., Grainger, 1990; Johnston & Barry, 2006; Romani, Mcapline & Martin, 2008). They were selected to reflect the range of variables of lexical constructs that appear related to semantic diversity in the literature. The number of items included in our analysis for each variable is shown in Table 2.1, along with the mean, range and standard deviation.

Behavioural measures. We investigated the relationship between semantic diversity measures and lexical decision performance. Accuracy and response time data were taken from the British Lexicon Project (BLP; Keuleers, Lacey, Rastle & Brysbaert, 2012) and the English Lexicon

Project (ELP; Balota et al., 2007). The BLP is a large-scale, openly accessible dataset containing lexical decision data from British English speakers for 28,730 English words. Similarly, the ELP contains lexical decision data from American English speakers for 40,481 words. These allowed for a thorough investigation of semantic diversity effects in a large itemset.

RESULTS

We first tested correlations between semantic variables. Our analysis used all available words within the itemsets in order to maximise the N for each calculation. Correlations were therefore calculated for the number of words which had a value for both variables tested. Descriptive statistics for these measures can be seen in Table 2.1.

Table 2.1: Descriptive statistics for the measures of semantic diversity.

Measure	Min	Max	Mean	SD	N
Hoffman (2013)	0.08	2.41	1.51	0.368	31,739
Cevoli (2021)	0.17	3.20	2.09	0.416	28,555
Johns (2020)	0.00	11.70	5.62	2.610	78,033
Johns (2021)	1.00	6,557.34	439.13	949.924	80,346
Frequency	1.173	7.67	2.031	0.859	160,022
Document count	40.00	87,627.00	1,100.00	4,386.667	31,715
Age of acquisition	1.58	25.00	11.00	3.044	31,105
Concreteness	1.04	5.00	3.04	1.039	39,954
Imageability	1.20	6.90	4.14	1.445	3,000
Number of senses	1.00	44.00	2.51	2.426	25,350

Note: This table presents diversity measures calculated by Hoffman et al. (2013), Cevoli et al. (2021), Johns et al. (2020), and Johns (2021), alongside Zipf frequency (Van Heuven et al., 2014), document count (Hoffman et al., 2013), age of acquisition (Kuperman et al., 2012), concreteness (Brysbaert et al., 2014), imageability (Cortese & Fugett, 2004), and number of senses (Wordsmyth; www.wordsmyth.net).

First, we tested the relationships between the different measures of diversity (Figure 2.1). All diversity measures correlated significantly with each other (Figure 2.2). Of these measures, the strongest correlations were between Hoffman and Cevoli ($r = .68$), and the weakest correlations between Cevoli and Johns 2020 ($r = .24$).

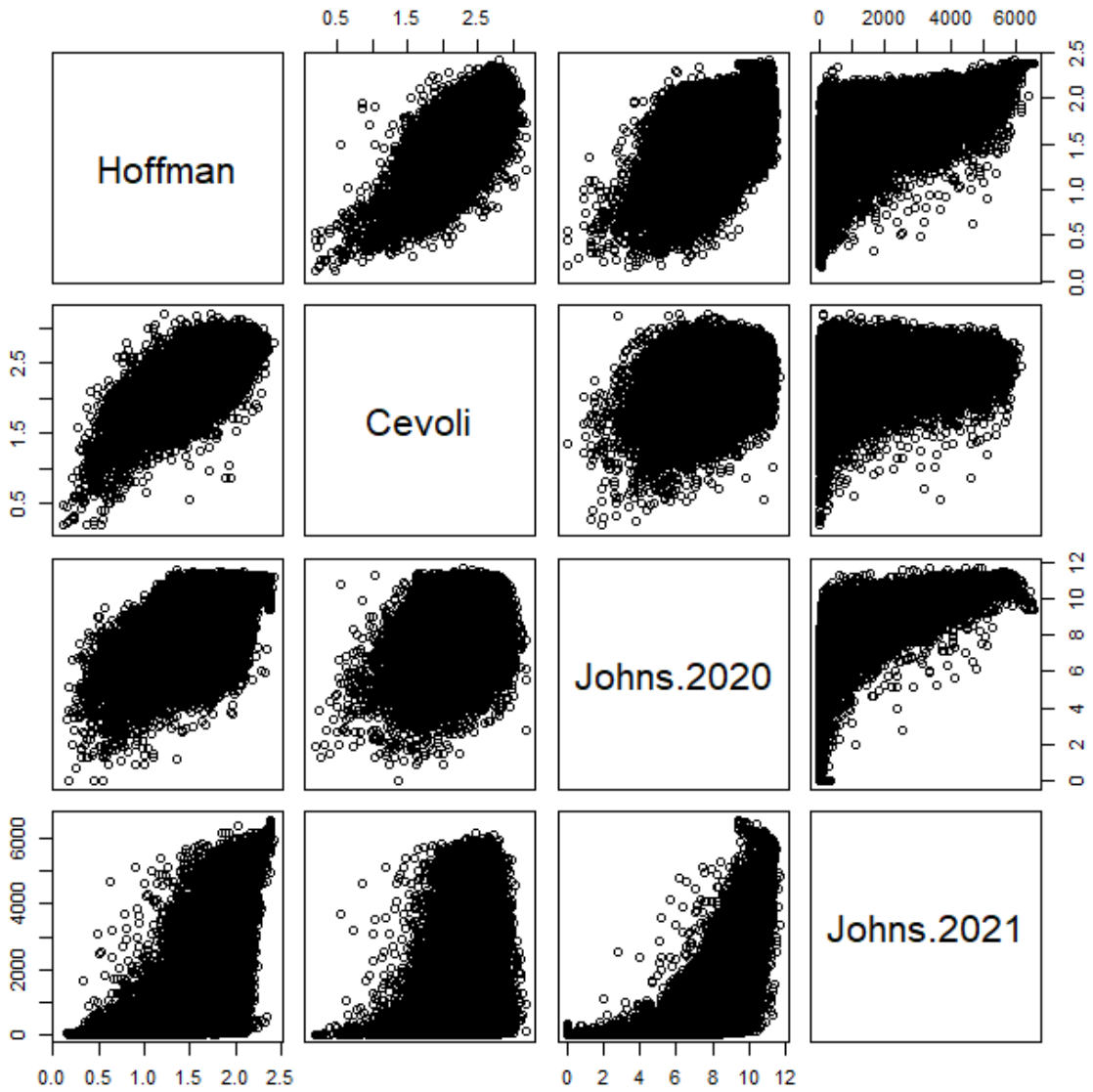


Figure 2.1: Scatterplots showing the relationship between semantic diversity measures.

Note: Diversity values are shown in the rows (top to bottom) and columns (left to right) in the following order: Hoffman et al. (2013), Cevoli et al. (2020), Johns et al. (2020). Johns (2021). The x axis value shows the value assigned to the column, and the y axis shows the value assigned to the row. Relationships between variables can be found by following where the two measures meet. For

example, the bottom left panel shows Hoffman et al. (2013) on the x axis, and Johns (2021) on the y axis.

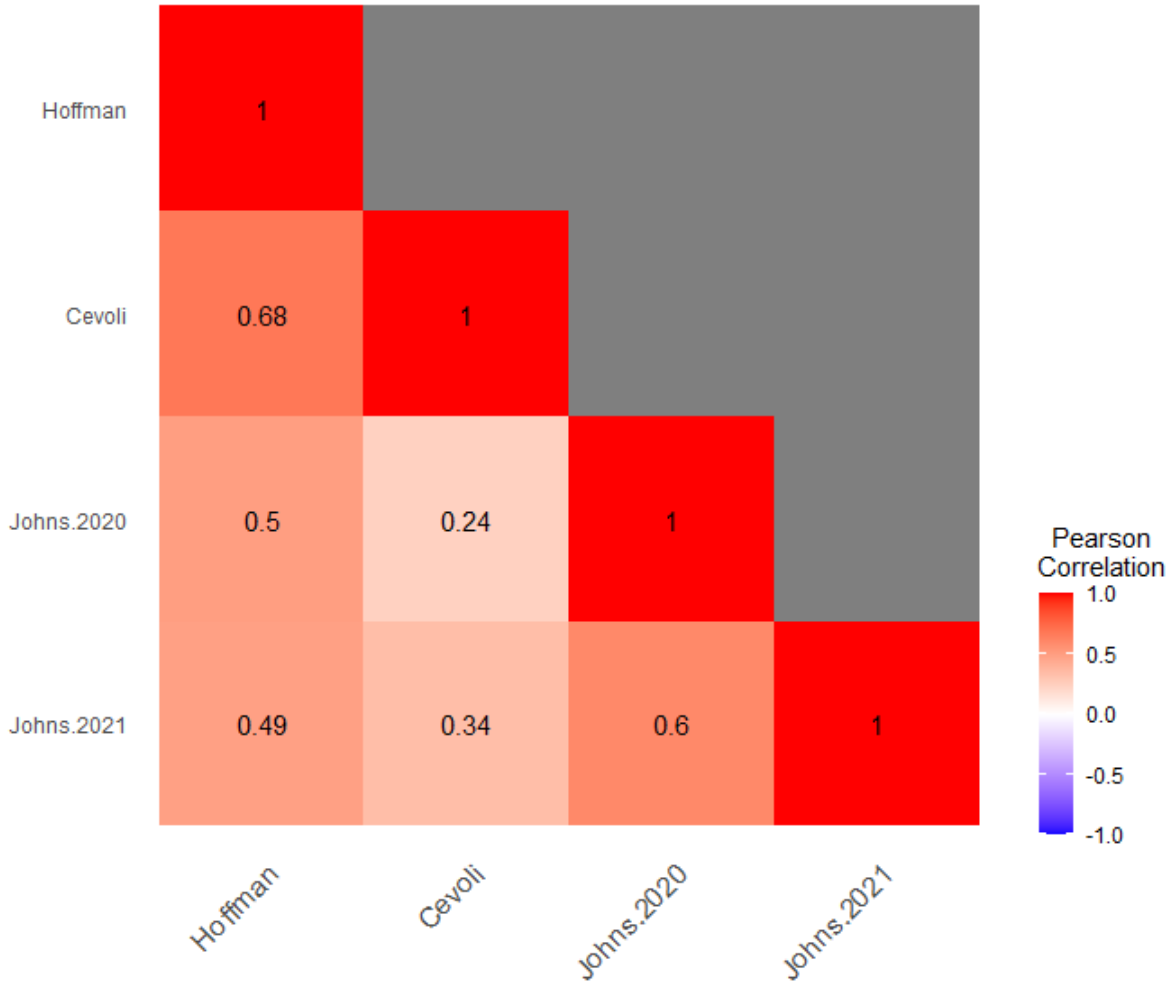
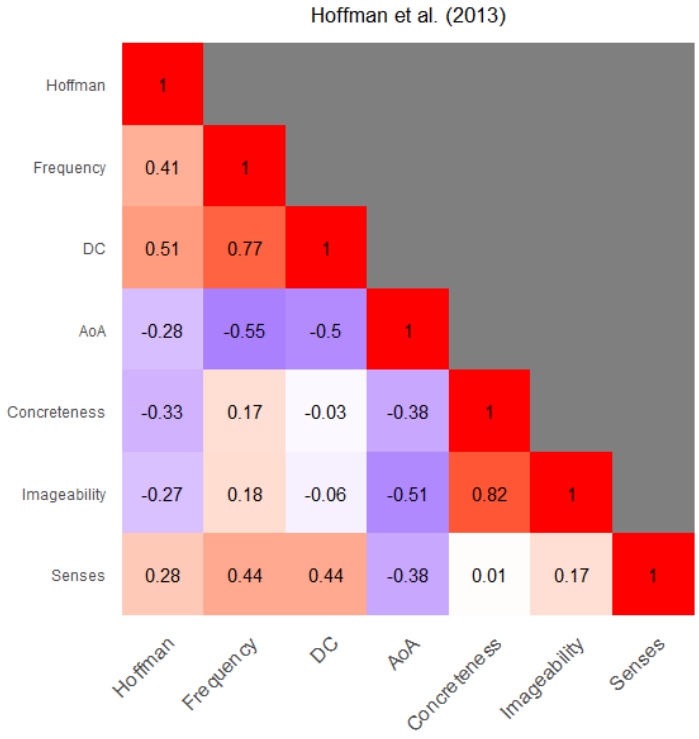
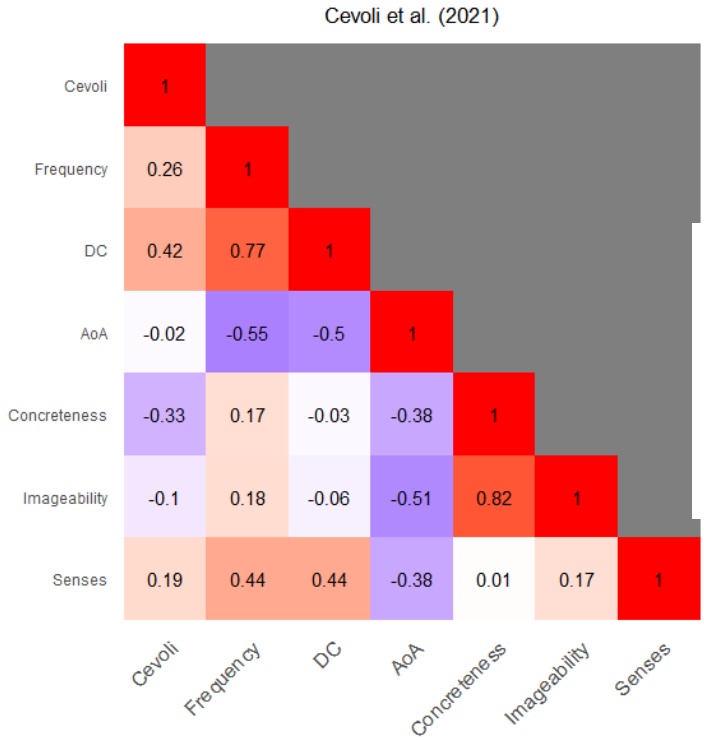
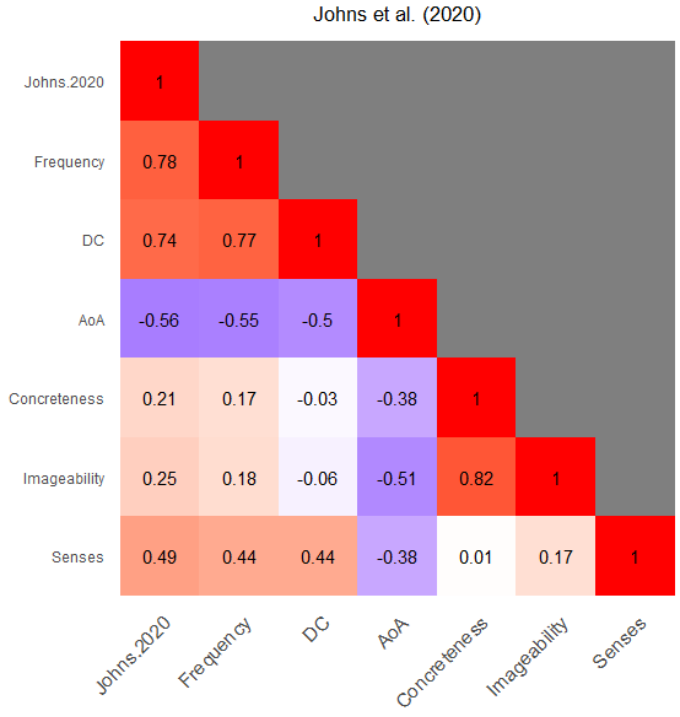


Figure 2.2: Heatmap showing correlation coefficients between the four diversity measures. All correlations were statistically significant.

We then tested the relationships between semantic diversity measures and other lexical variables. Hoffman's measure of semantic diversity correlated positively with frequency, document count, and number of senses; and negatively with age of acquisition, imageability and concreteness. Cevoli's measure patterned similarly, except the relationship between age of acquisition and diversity was non-significant. Johns et al. (2020) correlated positively with frequency, document count, concreteness, imageability and number of senses; and negatively with age of acquisition. Johns (2021) patterned similarly, with non-significant correlations between diversity and both concreteness and imageability. These correlations are shown for each semantic diversity measure in Figure 2.3.


Hoffman et al. (2013)
Hoffman
Frequency
DC
AoA
Concreteness
Imageability
Senses
1
0.41 1
0.51 0.77 1
-0.28 -0.55 -0.5 1
-0.33 0.17 -0.03 -0.38 1
-0.27 0.18 -0.06 -0.51 0.82 1
0.28 0.44 0.44 -0.38 0.01 0.17 1
Hoffman
Frequency
DC
AoA
Concreteness
Imageability
Senses


Cevoli et al. (2021)
Cevoli
Frequency
DC
AoA
Concreteness
Imageability
Senses
1
0.26 1
0.42 0.77 1
-0.02 -0.55 -0.5 1
-0.33 0.17 -0.03 -0.38 1
-0.1 0.18 -0.06 -0.51 0.82 1
0.19 0.44 0.44 -0.38 0.01 0.17 1
Cevoli
Frequency
DC
AoA
Concreteness
Imageability
Senses
Pearson Correlation
1.0
0.5
0.0
-0.5
-1.0


Johns et al. (2020)
Johns.2020
Frequency
DC
AoA
Concreteness
Imageability
Senses
1
0.78 1
0.74 0.77 1
-0.56 -0.55 -0.5 1
0.21 0.17 -0.03 -0.38 1
0.25 0.18 -0.06 -0.51 0.82 1
0.49 0.44 0.44 -0.38 0.01 0.17 1
Johns.2020
Frequency
DC
AoA
Concreteness
Imageability
Senses

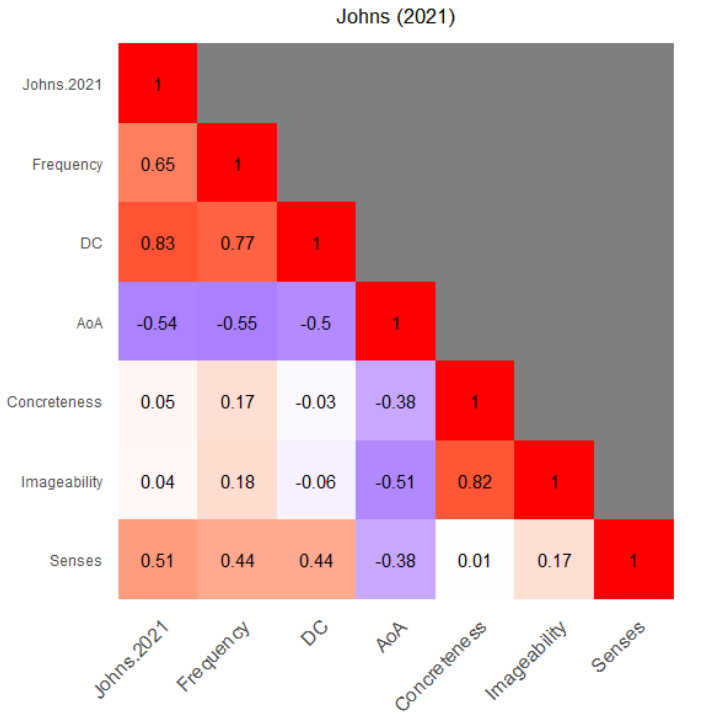

Johns (2021)
Johns.2021
Frequency
DC
AoA
Concreteness
Imageability
Senses
1
0.65 1
0.83 0.77 1
-0.54 -0.55 -0.5 1
0.05 0.17 -0.03 -0.38 1
0.04 0.18 -0.06 -0.51 0.82 1
0.51 0.44 0.44 -0.38 0.01 0.17 1
Johns.2021
Frequency
DC
AoA
Concreteness
Imageability
Senses
Pearson Correlation
1.0
0.5
0.0
-0.5
-1.0

Figure 2.3: Heatmap showing Pearson correlations between the measures of semantic diversity calculated by Hoffman et al. (2013, top left), Cevoli et al. (2021, top right), Johns et al. (2020, bottom left), and Johns (2021, bottom right) and other lexical-semantic variables.

Having established the pattern of associations between the different types of semantic diversity, we next explored how the different measures were associated with lexical decision, using data from the BLP and ELP. We controlled for Zipf frequency, age of acquisition, and concreteness. These variables were chosen to account for notable predictors of lexical decision in the literature, while avoiding multicollinearity. Multicollinearity was tested for using variance inflated factor (VIF) statistics on all linear mixed-effect models. For all models, VIF values were between one and five, thus lower than the 'rule of thumb' threshold of ten (O'Brien, 2007). Datasets were combined in R, such that each word had a valid entry for semantic diversity, Zipf frequency, age of acquisition and concreteness. Full data were available for 14,706 words for Hoffman, 15,497 words for Cevoli, 22,043 words for Johns (2020), and 22,045 words for Johns (2021). As Johns (2021) was on a much larger scale to the other variables and had a larger range, this was centred and scaled.

These variables were then combined with the BLP dataset and, separately, the ELP dataset. These two datasets contained trial-level data for lexical decision, allowing us to test whether semantic diversity was a unique predictor of lexical processing. To investigate this, we used linear mixed effect models. For predicting accuracy data, the function 'glmer' was used, and for predicting reaction time data 'lmer' was used, both from the package 'lme4' (Bates et al., 2007). P-values were computed using 'lmerTest' (Kuznetsova et al., 2015). Random effects of item and participant were also controlled for. By-participant slopes were first included for all lexical variables, but the model failed to converge. These were then removed one-by-one in order to identify the most complex model which would converge. All combinations including by-participant slopes for lexical variables

did not converge. Removing by-item slopes did not resolve this. Therefore, random intercepts were included for items and participants to give the following model:

Lexical decision ~ frequency + age of acquisition + concreteness + semantic diversity + (1|Word) + (1|Participant)

This was tested for accuracy and response time for both the BLP and the ELP. All measures of semantic diversity explained significant variance for BLP data, showing a facilitatory effect for both accuracy and response time. Table 2.2 shows the model fit for each of the semantic diversity measures predicting lexical decision response time from the BLP. The models patterned similarly when fitted to ELP data (see Appendix 1). Here, Hoffman et al.'s measure of semantic diversity explained unique variance in accuracy but did not explain variance in reaction time once other variables were controlled for. The measures from Cevoli et al. (2021), Johns et al. (2020) and Johns (2021) all explained significant variance in accuracy and response time for both BLP and ELP data. Full results fitting both BLP and ELP accuracy and response time data can be found in Appendix 1.

Table 2.2: Results from linear mixed effect models showing the proportion of variance in BLP response time explained by the model.

Hoffman et al. (2013)					
	Estimate	Std Error	Df	T value	P value
Intercept	664.4904	13.2186	61.2268	50.269	<.001
Frequency	-31.3921	0.8437	7489.9023	-37.206	<.001
AoA	6.3518	0.2674	7582.9412	23.753	<.001
Concreteness	-3.1507	0.6168	7454.8514	-5.108	<.001
SemD (Hoffman)	-9.0012	2.0263	7737.8722	-4.442	<.001

Cevoli et al. (2021)					
	Estimate	Std Error	Df	T value	P value
Intercept	678.7059	13.1388	56.3730	51.656	<.001
Frequency	-34.5278	0.8186	7806.0964	-42.178	<.001
AoA	6.4987	0.2612	7892.5614	24.876	<.001
Concreteness	-1.7840	0.5618	7702.0829	-3.176	<.01
SemD (Cevoli)	-10.7723	1.6864	7857.4648	-6.388	<.001

Johns et al. (2020)					
	Estimate	Std Error	Df	T value	P value
Intercept	738.4564	13.2409	50.8119	55.771	<.001
Frequency	-26.9112	0.9464	9412.2099	-28.435	<.001
AoA	4.7008	0.2435	9509.2119	19.308	<.001
Concreteness	-2.0211	0.5168	9249.7726	-3.911	<.001
SD (Johns, 2020)	-10.5982	0.5528	9537.3908	-19.172	<.001

Johns (2021)					
	Estimate	Std Error	Df	T value	P value
Intercept	681.8434	17.2198	2785.6905	39.597	<.001
Frequency	-4.3779	3.2131	2456.1140	-1.363	0.173
AoA	7.9260	0.7387	2450.1335	10.730	<.001
Concreteness	-11.2231	1.5620	2449.6815	-7.185	<.001
SD (Johns, 2021)	-34.2671	2.9113	2453.7106	-11.770	<.001

Note: This tested the model: Lexical decision response time ~ frequency + age of acquisition + concreteness + semantic diversity + (1|Word) + (1|Participant). Each table shows the model fitted with a different measure semantic diversity, in the order: Hoffman, Cevoli, Johns (2020), Johns (2021).

We then compared each of the four models containing semantic diversity to the base model using ANOVA. This approach was taken as restricted likelihood values cannot be directly compared using ANOVA or Chi-square when models have different fixed effects (Bolker, 2015). The base model therefore allowed the contribution of each semantic diversity measure to be assessed via comparison to a common model. Within this approach, Bayesian Informational Criterion (BIC) scores were calculated, allowing for comparison between model fits. To allow for such comparisons, only words with valid entries for all variables were analysed. Analysis of variance tests revealed that all models including semantic diversity provided a better fit to BLP response time than the base model. For ELP data, including Hoffman et al.'s semantic diversity did not improve the model fit, but the other semantic diversity values all improved model fit. According to BIC scores, Johns (2020) provided the best fit to BLP data and Johns (2021) to ELP data. However, differences between models appeared very small. Comparisons are shown in Table 2.3.

Table 2.3: Comparisons between the base model and semantic diversity model for BLP (above) and ELP (below) response time data.

			BLP rt			
	AIC	BIC	deviance	Chi-square	df	p
Base	*1594770*	*1594838*	*1594756*			
Hoffman	1594752	1594830	1594736	19.778	1	<.001
Cevoli	1594733	1594811	1594717	38.943	1	<.001
Johns 2020	1594607	1594685	1594591	164.44	1	<.001
Johns 2021	1594665	1594743	1594649	106.87	1	<.001

			ELP rt			
	AIC	BIC	deviance	Chi-square	df	p
Base	*1131549*	*1131614*	*1131535*			
Hoffman	1131548	1131622	1131532	2.9549	1	0.0856
Cevoli	1131537	1131611	1131521	14.17	1	<.001
Johns 2020	1131465	1131540	1131449	85.48	1	<.001
Johns 2021	1131452	1131527	1131436	98.235	1	<.001

Note: The base model is shown first, followed by the comparison of the base model with the base + each semantic diversity model.

DISCUSSION

We tested the relationships between different measures of semantic diversity and investigated how these measures correlated with other lexical-semantic variables. We then tested the ability of each diversity measure to predict lexical decision performance. All semantic diversity measures were positively correlated with each other. Of these, the strongest correlations were

between Hoffman's and Cevoli's measures, and the weakest between Cevoli's and Johns' (2020) measures. All measures of diversity correlated positively with frequency and document count, and negatively with age of acquisition (although this relationship was non-significant for Cevoli's measure). For concreteness, Hoffman and Cevoli's measures showed a negative relationship, while the measures of Johns et al. (2020) and Johns (2021) showed a weaker positive relationship. This might relate to the theoretical differences between the measures. Hoffman et al.'s conceptualisation of diversity is grounded in polysemy, which is a distinct construct to concreteness (Vergara-Martinez et al, 2017), while the Semantic Distinctiveness Model does not propose a relationship to concreteness. Alternatively, this might be due to differences in the corpora from which diversity measures are based.

To build upon the debate between Hoffman et al. (2013; 2021) and Cevoli et al. (2021) regarding semantic diversity's ability to capture ambiguity, we observed a difference between these measures in their correlations with number of senses. While both measures correlated positively with number of senses, this was stronger for Hoffman et al. ($r = .28$ compared to $r = .19$, $z = 11.0964$, $p<.001$). This supports Hoffman et al.'s (2021) claim that calculations based on unscaled vectors do capture information relating to ambiguity. We observed a correlation of .68 between Hoffman et al. and Cevoli et al.'s diversity measures. This is lower than Cevoli et al. report for their values based on the lemmatised corpus ($r = .73$). The reason for this is unclear. We used both Cevoli et al. and Hoffman et al.'s metrics that are openly available online, and we tested the correlation for the maximum number of words with a valid entry for both metrics.

When predicting variance in behavioural performance, all measures of diversity facilitated lexical decision accuracy for both the BLP and ELP datasets. For reaction time, Hoffman's diversity explained variance in BLP lexical decision latencies, but not in ELP latencies. All other diversity measures predicted both sets of response time data. ANOVA comparisons demonstrated that all

semantic diversity values improved the model fit in lexical decision response time. Johns et al.'s (2020) and Johns' (2021) values appeared a slightly better fit than Hoffman et al.'s (2013) and Cevoli et al.'s (2020) measures, but these differences were small.

These results lend support to the Semantic Distinctiveness Model. They demonstrate that lexical strength of word representations across a lifetime of experience is important to lexical processing. This suggests that the word representations updated through varied contexts could be more important to lexical processing than the continuous measure of polysemy. This is fitting with Jones et al.'s (2012) findings that the Semantic Distinctiveness Model provides better estimates to lexical performance than latent semantic analysis. This might suggest that semantic similarity is better when measured by comparing the similarity of whole texts (as outlined by the Semantic Distinctiveness Model) rather than by counting rates of occurrence in a word-document matrix (as is integral to latent semantic analysis).

However, it is also important to consider how the measures of diversity were calculated. Johns et al. (2020) and Johns (2021) both used much larger corpora than Hoffman et al. and Cevoli et al. Results from word frequency measures demonstrate that, while frequency effects do not differ based on corpus-size for high frequency words, low- and medium-frequency words provide better fit to word naming latencies when calculated from a larger corpus (Burgess & Livesay, 1998). The same could be true for diversity: large samples might be better at providing accurate information about a word's spread across contextual use, particularly for words that do not frequently appear. Further research is needed to determine whether a larger corpus leads to more accurate semantic diversity values.

In the introduction, we proposed that the contextual window might be an important aspect of how diversity is calculated. This window is much broader in Johns' (2021) measure (defined as a whole subreddit) than the other measures reported in our analysis. The good fit of this measure to

lexical processing suggests that a word's lexical history may be influenced by broad contextual information. However, as Johns et al.'s (2020) metric included here used a contextual window of 20 words and appeared a good predictor of lexical decision, the breadth of context is likely not the most important factor influencing fit of diversity values. Similarly, the introduction proposed that individual word usage might be important. This factor is considered by Johns' (2021), but not by the other diversity values. As with the contextual window, it is unlikely that this explains why Johns' (2021) measure provided such a good fit to lexical decision latencies, and Johns et al.'s (2020) measure provided a similar fit without considering the spread of a word's usage across individuals.

Rather, similarities between Johns et al.'s (2020) and Johns' (2021) methodologies might explain why these variables provided a better fit to lexical decision data than Hoffman et al.'s and Cevoli et al.'s The corpora used by Johns et al. and Johns were arguably more reflective of typical linguistic experience: Johns et al.'s corpus of fiction and non-fiction books is likely a good proxy for written language experience, while Johns' reddit corpus reflects more casual language use. By comparison, a large proportion of the British National Corpus (used by Hoffman et al. and Cevoli et al.) is based upon informative written texts, such as books and periodicals on applied science, commerce and finance, and world affairs (BNC Consortium, 2007), and therefore might be less representative of everyday linguistic experience.

Together, these results show that the broad pattern of results between different measures of semantic diversity and other lexical variables is similar. Slight differences in correlations were observed; most notably between diversity values and concreteness and imageability, where the direction of the relationship changed. Small differences were seen between measures in their effects on lexical processing, with Johns et al. (2020) and Johns (2021) best fitting the data. These differences might be driven by theory supporting the Semantic Distinctiveness Model or might demonstrate the importance of using a representative language sample.

STUDY 2: PREDICTING SEMANTIC DIVERSITY

As discussed, diversity is reportedly a better predictor of lexical processing than word frequency (Hoffman et al., 2013). To explain this result, the literature proposes two possible explanations. First, semantic diversity effects could function in a similar way to other measures of polysemy, like number of senses. Congruent with connectionist models of semantic representation, words high in diversity would activate multiple nuanced representations of meaning, resulting in boosted lexical processing during some tasks, and slowed processing in others. Second, semantic diversity effects could arise due to the importance of experiencing words in diverse and varied contents. According to this explanation, the contextual variability captured by semantic diversity is important to how words are accessed from memory. This would fit with the explanation of lexical organisation proposed by Jones et al.'s (2012) Semantic Distinctiveness Model, and is supported by the results of Study 1. In both the Semantic Distinctiveness Model and the polysemy account of semantic diversity, diversity values reflect a history of varied linguistic experience. If semantic diversity effects occur due to the dynamic contexts a word has previously appeared in, we would expect semantic diversity to be somewhat independent from other word-level measures of semantic richness (e.g., concreteness, imageability) and occurrence (e.g., word frequency, document count). Similarly, if these effects occur due to polysemy, we might expect semantic diversity to be independent from semantic richness and occurrence measures, but we might expect some variance in semantic diversity to be explained by number of senses.

However, an alternative explanation for semantic diversity effects exists. As discussed in Chapter 1, diversity correlates moderately with a wide range of lexical and semantic variables, including frequency, age of acquisition, concreteness, and number of senses. The effects associated with semantic diversity could be due to these covarying measures. For example, take a word that is

high in semantic diversity and low in concreteness. This word's slowed performance in semantic decision could either be attributed to semantic diversity, but it could also be explained by its abstractness. The extent to which other lexical variables might explain the effects of semantic diversity is unclear. Regression-type analyses give an opportunity to investigate this. If the effects of diversity on lexical processing are due to overlapping variance from other measures, we would expect these variables to explain a large proportion of variance within semantic diversity values.

While Study 1 showed that semantic diversity measures share moderate correlations with a wide range of variables, they are not the same as any one variable. Therefore, it is likely that a mixture of these two above propositions is true. We know that semantic diversity is related to abstractness and polysemy, and so we might expect measurements of these constructs to explain some variance in diversity values. But, as semantic diversity is unique in tracking a word's semantic information across contexts via co-occurrence, we might also expect a large proportion of semantic diversity to be unaccounted for by other semantic predictors.

The relationship between diversity and other variables was explored by Hoffman et al. (2013) using principal components analysis, reproduced in Table 2.4. Here, semantic diversity was linked to two latent factors: it loaded positively onto frequency and ambiguity information (factor 1), and negatively onto information relating to semantic richness (factor 2).

Table 2.4: Principal components analysis reported by Hoffman et al. (2013), using varimax rotation.

	Factor 1	Factor 2
Semantic diversity	.58	–.61
Log word frequency	.93	
Concreteness		.93

Imageability		.82
Context availability		.77
Contextual diversity	.93	
Contextual distinctiveness	–.83	
Log number of senses	.64	

Note: Values of < .3 are omitted for ease of interpretation.

We aimed to build upon this work using structural equation modelling. This method allowed us to verify the structure reported by Hoffman et al., and to further test the extent to which this factor structure could explain variance in semantic diversity values using regression. The model structure was based on Hoffman et al.'s two factors, but our model had three differences. First, we excluded contextual distinctiveness as these values were not openly available. Second, we excluded context availability as the current norms available (N = 235) limited the size of the itemset. Finally, we included age of acquisition based on the negative correlations with diversity reported in Study 1.

To remain consistent with prior work, we tested this model in fitting Hoffman et al.'s (2013) measure of semantic diversity. This analysis allowed us to firstly test how much variance in semantic diversity could be explained by other key lexical and semantic variables. Then, we fitted the same model to Johns' (2021) measure. We predicted that two latent factors would explain a moderate proportion of variance in both measures of semantic diversity. However, as semantic diversity is proposed to capture contextual variation in a way distinct from other measures, we predicted that a larger proportion of variance in semantic diversity would be left unexplained.

Our analysis also gave an opportunity to compare these two diversity measures. This allowed us to investigate differences between the polysemy account of semantic diversity (as captured by Hoffman et al., 2013) and the Semantic Distinctiveness Model (as captured by Johns, 2021). These measures are theoretically different, yet both rely on calculating similarity between the

contexts in which words appear. Our analysis allowed us to investigate whether these measures are tapping into different lexical constructs.

MEASURES

Our model included the same lexical variables included in Study 1. Firstly, we tested this model to explain variance in Hoffman et al.'s (2013) semantic diversity values. We then compared this to a model predicting Johns' (2021) UCD-SR-PR diversity measure.

ANALYSIS

We used structural equation modelling to test the factors that semantic diversity has previously loaded onto. This analysis used the same itemset as Study 1. Our structure followed that of Hoffman et al. (2013). The first factor reflected 'occurrence', with loadings from Zipf frequency, document count, number of senses, and age of acquisition, and was based on Hoffman et al.'s Factor 1. The second factor reflected 'richness', with loadings from concreteness, imageability, and age of acquisition, and was based on Hoffman et al.'s Factor 2. All variables were centred and scaled. Due to their abnormal distribution, frequency, document count, and number of senses were log transformed.

First, confirmatory factor analysis (CFA) was used to validate the factor structure. Schreiber et al. (2006) suggested that the root mean square error of approximation (RMSEA) should be less than .06-.08, and the comparison fit index (CFI) should be above .95. Initially, while the chi-square (x^2 = 64.41, p<.001) and CFI (0.977) were acceptable, the RMSEA (0.108) value was higher than indicative of a good fit.

In order to improve the model fit, we made changes as suggested by the modification indices. These indices are an output of confirmatory factor analysis which give suggestions for how the model can be improved based on the data. Our decision process here took two steps. First, changes were identified by the modification indices. Next, we assessed the extent to which the modifications were theoretically sound. Changes that were supported by both the data and theory were implemented.

Based on the modification indices, number of senses was entered as an independent predictor, rather than loading onto 'occurrence'. We justified this theoretically as senses is proposed to tap polysemy information, which differs from the frequency-type information reflected by word frequency and document count (which together with age of acquisition made up 'occurrence'). We specified that senses would covary with 'occurrence' and with concreteness. We also specified covariance between frequency and age of acquisition, and between document count and semantic diversity. These decisions were supported by our results from Study 1 and previous literature.

The modifications led to an acceptable model fit. The chi square was significant, $x^2(7) = 32.58$, $p<.001$, and both CFI (.992) and RMSEA (.072) were acceptable. The path diagram can be seen in Figure 2.4. Occurrence, richness, and number of senses all significantly predicted semantic diversity. All specified covariances were significant. Hoffman et al.'s semantic diversity was positively predicted by occurrence and number of senses, and negatively predicted by richness. As the model fit was only acceptable, this suggests that a large proportion of variance in diversity was unexplained. Regression statistics and covariance can be seen in Table 2.5. The upper part of this table shows regression statistics for the predictor variables (occurrence, richness, and number of senses) in explaining variance in Hoffman's semantic diversity. The lower part of this table shows covariance statistics between the latent factors and lexical variables for the relationships we

specified in our model. These were all significant, which is unsurprising as these decisions were largely data-driven.

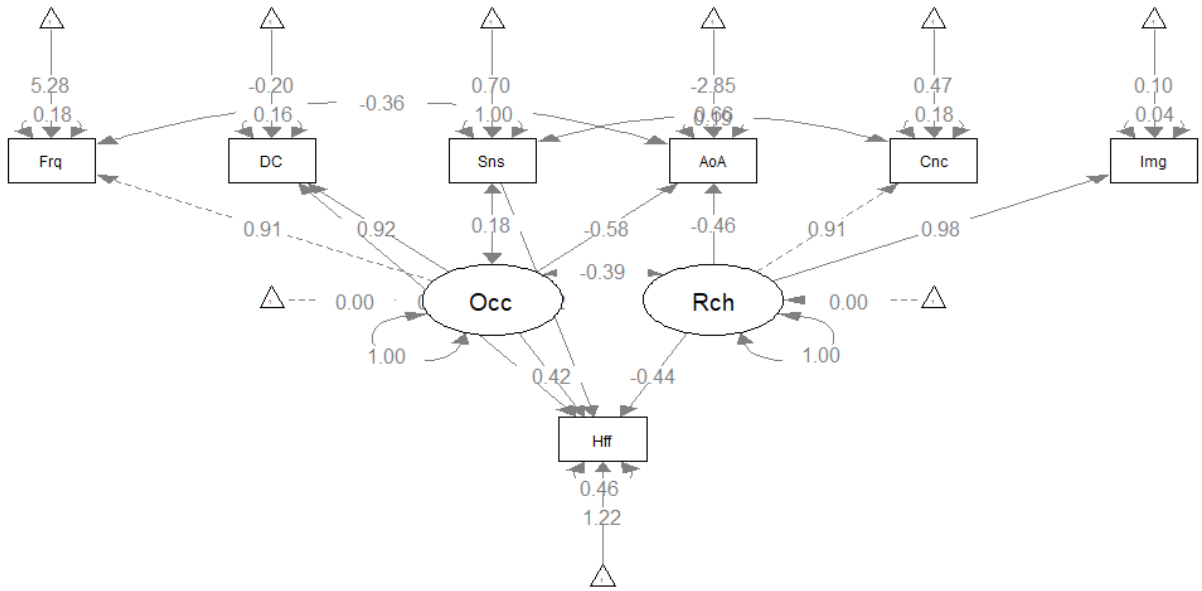


Figure 2.4: Path diagram for the adjusted structural equation model, showing the regression: Hoffman's semantic diversity ~ 1 + Occurrence + Richness + senses.

Note: Standardised parameter estimates are labelled. Loadings fixed to 1.0 (frequency and concreteness) are shown by a dashed line. Frqn = Zipf frequency (Van Heuven et al., 2014), DC = document count (Hoffman et al., 2013), Sns = number of senses (Wordsmyth; www.wordsmyth.net), AoA = age of acquisition (Kuperman et al., 2012), Cnc = concreteness (Brysbaert et al., 2014), Img = imageability (Cortese & Fugett, 2004), Hff = semantic diversity (Hoffman et al., 2013).

Table 2.5: Regression statistics (upper) and covariances between factors and variables (lower) for the model: Hoffman's semantic diversity ~ 1 + Occurrence + Richness + number of senses.

Factor	Estimate	Std Error	z	p	Std lv	Std all
Occurrence	1.447	0.132	11.003	<.001	0.299	0.422
Richness	-0.345	0.022	-15.590	<.001	-0.309	-0.436
Senses	0.088	0.017	5.044	<.001	0.088	0.123

Factor	Estimate	Std Error	z	p	Std lv	Std all
Occ ~~ Rch	-0.072	0.008	-8.640	<.001	-0.388	-0.388
Occ ~~ senses	0.037	0.008	4.834	<.001	0.181	0.181
Frequency ~~ AoA	-0.017	0.004	-3.890	<.001	-0.017	-0.362
DC ~~ diversity	0.119	0.030	3.974	<.001	0.119	0.377
Concrete ~~ senses	0.078	0.017	4.578	<.001	0.078	0.189

Note: 'Occ' = occurrence; 'Rch' = richness.

We then fitted the same model to Johns' (2021) measure of semantic diversity. Here, the chi square was significant, $x^2(7) = 84.169$, $p<.001$, and the CFI (.978) was acceptable. The RMSEA value (.125) was indicative of a good fit. The path diagram can be seen in Figure 2.5. Occurrence, richness, and number of senses all significantly predicted semantic diversity. Occurrence shared significant covariance with richness and senses; and concreteness shared significant covariance with senses. The other specified covariances were not significant. The model fit appeared similar when predicting Johns as to when predicting Hoffman et al., in that all variables were significant predictors but the model fit was only acceptable according to the CFI, suggesting a large proportion of variance was left unexplained. However, the strengths of predictors differed across models: while Hoffman et al. was moderately predicted by both occurrence and richness, variance in Johns et al. appeared to be largely associated with occurrence. Regression statistics for Johns' predicting semantic diversity can be seen in Table 2.6, alongside statistics for the covariances specified in the model.

Table 2.6: Regression statistics (upper) and covariances between factors and variables (lower) for the model: Johns' (2021) semantic diversity ~ 1 + Occurrence + Richness + number of senses.

Factor	Estimate	Std Error	z	p	Std lv	S td all
Occurrence	5.527	0.368	15.007	<.001	1.140	0.888
Richness	-0.102	0.038	-2.693	.007	-0.092	-0.072
Senses	0.125	0.027	4.680	<.001	0.125	0.097

Factor	Estimate	Std Error	z	p	Std lv	Std all
Occ ~~ Rch	-0.073	0.008	-8.672	<.001	-0.391	-0.391
Occ ~~ senses	0.040	0.008	5.175	<.001	0.194	0.194
Frequency ~~ AoA	-0.006	0.005	-1.301	0.193	-0.006	-0.144
DC ~~ diversity	-0.143	0.091	-1.571	0.116	-0.143	-0.465
Concrete ~~ senses	0.077	0.017	4.510	<.001	0.077	0.190

Note: 'Occ' = occurrence; 'Rch' = richness.

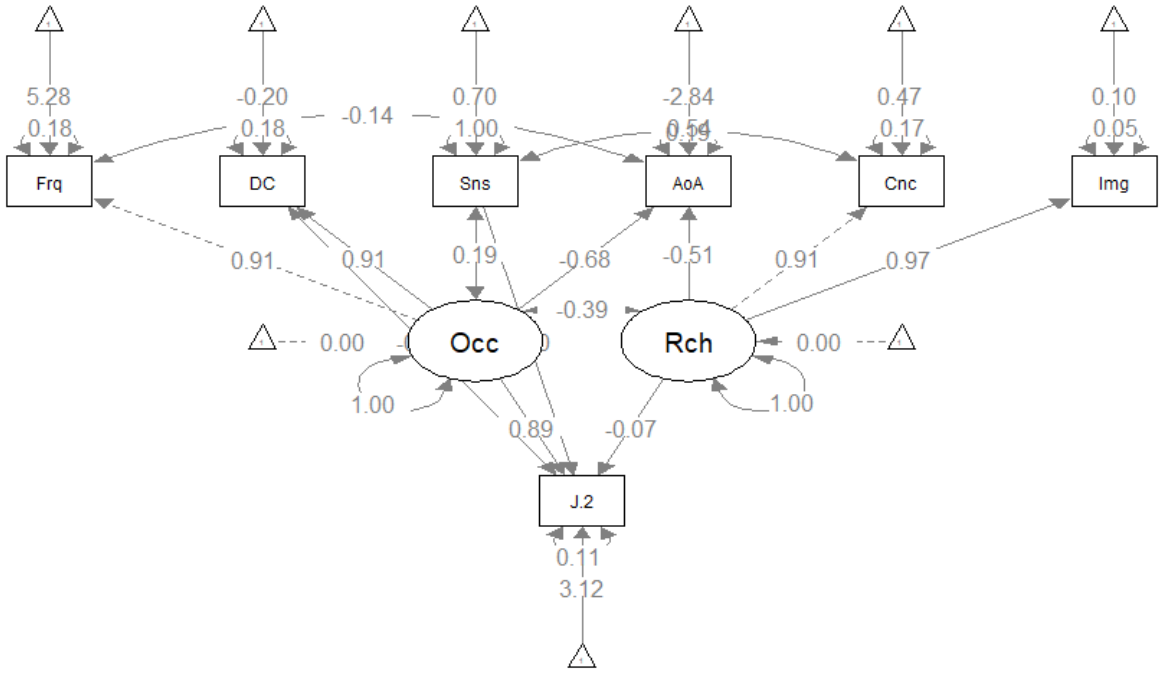


Figure 2.5: Path diagram for the adjusted structural equation model, showing the regression: Johns' (2021) semantic diversity ~ 1 + Occurrence + Richness + senses.

Note: Standardised parameter estimates are labelled. Loadings fixed to 1.0 (frequency and concreteness) are shown by a dashed line.

DISCUSSION

We investigated the underlying structure of semantic diversity. We hypothesised that two latent factors (occurrence and richness) and number of senses (entered independently based on modification indices) would explain a significant proportion of variance in both measures of semantic diversity. We also predicted that a large proportion of variance would not be accounted for. By fitting models to semantic diversity as operationalised by both Hoffman et al. (2013) and Johns (2021), we explored whether there would be differences in the model fit for two different diversity measures.

Our analysis found that the two factors, and number of senses entered independently, did contribute to variance in both measures of diversity. The model was an acceptable fit according to the CFI and RMSEA values for Hoffman et al.'s measure and had an acceptable CFI and good RMSEA when fit to Johns' measure. As the model statistics did not indicate a robustly good fit for either diversity measure, this suggests that a large portion of variance was left unexplained. This unexplained variance supports the idea that semantic diversity is measuring unique linguistic information that is not fully captured by the factors included in this study. Applied to semantic diversity's effects over lexical processing, our results suggest that these effects are attributable to semantic diversity itself and not due to co-occurring constructs.

However, we should note the limitations of our model, as we cannot rule out other explanations. The relationship between variables included in our analysis is likely more complex than specified by our model. Lexical variables are often highly correlated. For example, Hoffman et al.

(2013) found that frequency was significantly correlated with document count (r = 0.99), number of senses (r = 0.52), semantic diversity (r = 0.49), and imageability (r = -0.20). While our model assumed a correlation between the latent factors, and covariance was stated based on modification indices, we did not specify the correlations for all variables. It is also likely that variables not included in our analysis explain variance in semantic diversity. Baayen's (2010) analyses predicted word frequency from a large array of orthographic and morphological variables, such as word length and syntactic family size. Our analyses were instead led by semantic factors previously identified by Hoffman et al. (2013). A more complex model with a wider array of predictors could lead to a better model fit. Further research is needed to establish the relationship between diversity and non-semantic factors.

The tested model explained a similar proportion of variance in both measures of semantic diversity. However, we observed differences in the contributions of the latent variables between the two measures. For Hoffman's values, occurrence and richness had similar coefficients in predicting semantic diversity, with occurrence a positive predictor (0.42) and richness a negative predictor (-0.44). This replicates the pattern of loadings reported by Hoffman et al. (2013) in their principal components analysis. For Johns' measure, the difference between coefficients was much larger. Occurrence was a strong positive predictor (0.89) and richness a small but significant negative predictor (-0.07). This suggests that Hoffman et al.'s measure might capture both frequency-type information and abstractness, while Johns' measure might be more based in linguistic experience, fitting with the differences observed through correlations in Study 1.

Number of senses gave a similar fit to both diversity measures and appeared to explain slightly more variance in Johns' (2021) measure. This contradicts the assumption that Hoffman et al.'s diversity provides a closer measure of polysemy than Johns' diversity. Rather, both measures appear to capture some element of polysemy, with Johns' perhaps capturing more polysemy. This demonstrates that there might not be a one-to-one relationship between the theoretical

explanations of semantic diversity and the measures proposed. That is to say, while Hoffman et al. propose that polysemy drives semantic diversity, this does not mean that Hoffman et al.'s measure does not capture the updating of word representations as detailed by the Semantic Distinctiveness Model; and while Johns frames his measure within the Semantic Distinctiveness Model, this does not prevent it from also capturing polysemy. Rather, both variables might be explained by either (or both) theoretical explanations.

In summary, this research suggests that semantic diversity effects are not readily explained by other semantic variables. Diversity effects can therefore be interpreted as the effects of a word's contextual history, rather than a summed effect of other co-occurring constructs. The model fit was similar for both Hoffman's and Johns' (2021) measures. Differences were observed between these measures in the coefficients of occurrence and richness, suggesting that the two measures capture different types of semantic variance.

General Discussion

The results from Study 1 demonstrate that diversity measures broadly pattern similarly in their relationships with other measures. However, some differences were observed both in their relationship with other variables and in their effects on lexical processing. This was confirmed in Study 2, in which lexical-semantic variables explained a similar proportion of variance in both Hoffman et al.'s and Johns' measures, but different contributions from these variables were observed.

These differences could be due to the method with which the measures were calculated. Hoffman et al. and Cevoli et al. employed similar methods of calculating semantic diversity: latent semantic analysis was used to compare the lexical co-occurrence of words in their surrounding

context. Diversity therefore reflected dissimilarity between context for each use of a given word. Johns' (2020) measure also calculated dissimilarity, but their method was not based in latent semantic analysis, and measured similarity across the text rather than using a frequency-type count of word-document occurrence, as is implemented in LSA (see Figure 2.1 for example). Johns' (2021) measure had the additional user-based information, setting it aside from the other metrics. These differences in computation could affect the nature of the measurement.

Another explanation for observed differences could be the corpora used to derive the measures. Hoffman et al. and Cevoli et al., whose measures were most strongly correlated (r = .68), both used variations of the BNC, which contains approximately 100 million words from a range of newspapers, journals, fiction and non-fiction books, letters, essays, and unscripted conversations. By comparison, Johns (2020) used a corpus of 2.1 billion words from fiction and non-fiction books, while Johns (2021) used a corpus of 55 billion words from online Reddit posts. Both the corpus size and the differences in sources could contribute towards differences in the measures calculated. Johns' (2020) diversity reflected 'book' language which might be more formal and descriptive than the widely variable Reddit language. Despite this, Johns et al.'s and Johns' measures patterned similarly to each other in correlative analyses. To determine why differences between diversity measures occur, future research could compare diversity calculated as described by Hoffman and diversity as described by Johns using the same corpus.

Comparisons between Hoffman et al.'s and Cevoli et al.'s measures allows us to assess the influence of the calculation method more directly, as both values were based on the same corpus. Hoffman et al.'s and Cevoli et al.'s measures patterned similarly to each other, but Hoffman's measure was more strongly correlated with both Johns et al. (2020) and Johns (2021), and with frequency, document count, and age of acquisition. Cevoli's measure explained more variance in lexical decision than that of Hoffman, though this difference was very small. This was likely due to

Cevoli et al.'s use of scaled singular values, which the researchers argued were important for latent semantic analysis (Cevoli et al., 2021). Hoffman et al. (2021), however, demonstrated that diversity measures derived using unscaled vectors better fit semantic judgements. The differences between our findings and theirs could highlight a difference between lexical and semantic judgements. Hoffman's measure is more reliably related to polysemy, which might be more important when reflecting on the meaning of words during a semantic judgements task. However, textual diversity (as captured by Cevoli et al.) might be more important to lexical decision, in which processing the visual wordform is important.

Study 2 used structural equation modelling to test the proportion of variance in semantic diversity that can be explained by other lexical-semantic factors. We tested a model consisting of two latent factors, occurrence (with loadings from frequency, age of acquisition, and document count) and richness (with loadings from age of acquisition, concreteness and imageability), and number of senses entered independently. This model was an acceptable fit for both Hoffman et al.'s (2013) and Johns' (2021) measures of semantic diversity but did not provide a robust fit to either measure. This suggests that both measures of diversity cannot be explained by the other variables in our model, supporting the idea that semantic diversity's effects on lexical processing are due to the contextual variability captured through this measure, rather than due to overlapping variance with other lexical and semantic measures.

The decision to enter number of senses as an independent predictor was data-driven, informed by the modification indices in order to improve the model fit. However, this gave an opportunity to further explore differences between the two measures in their relationship to polysemy. Hoffman et al. (2013) proposed diversity as a continuous measure of polysemy. We would therefore expect number of senses to explain a significant proportion of variance in Hoffman's diversity; perhaps to a larger extent than when fit to Johns' measure. Number of senses was a

significant predictor to both measures of diversity. In fact, it explained a larger proportion of variance in Johns' measure than in Hoffman's. This pattern is also seen in Study 1, where the correlation between number of senses and semantic diversity is stronger for Johns' measure (r = .51) than for Hoffman's (r =. .28).

The closer relationship between Johns' measure and number of senses does not necessarily mean that Johns' measure is a better predictor of polysemy. Number of senses correlated strongly with frequency (r = .44) and document count (r = .44), both of which had stronger correlations with Johns' than Hoffman's diversity. Further, Johns' diversity appeared to be largely driven by occurrence information in Study 2, which correlated with number of senses. Hoffman et al. (2013) discuss the limitations of counting senses as a discrete measurement, arguing instead in favour of continuous contextual variation. In this view, number of senses is likely not sensitive enough to act as the proxy of polysemy by which we should compare diversity measures. Nonetheless, the observed relationship between number of senses and Johns' (2021) measure brings us to question whether the theoretical distinctions made by Hoffman et al. and Johns map onto the diversity values presented by the researchers. In the introduction, we argued that Hoffman et al.'s diversity statistics reflect polysemy, while Johns' reflect the Semantic Distinctiveness Model. Our results suggest that this might be an oversimplistic view. Rather, both theoretical explanations of diversity might hold true for both types of diversity measure, and differences between diversity measures might be driven by methodological differences rather than theoretical ones. This makes sense as latent semantic analysis as outlined by Hoffman et al. (2013, and see Chapter 1) is designed to capture semantic co-occurrence rather than polysemy; and the equation used by Jones' (2012) Semantic Distinctiveness Model also captures lexical similarity.

We turn, then, to methodological differences to explain the results of Study 2. While Hoffman's measure had similar coefficients for occurrence and richness information, Johns' measure

was more closely related to occurrence information. One explanation for this difference could be the breadth of contextual information considered in the calculations. As Johns (2021) based his measure on much larger contexts which contained more words, information captured by the close-proximity context could be lost. This close context could contain experiential information that drives the negative relationship between semantic diversity and concreteness.

To illustrate this with an example, the word 'star' has a moderate diversity value according to Hoffman et al. but scores highly according to Johns. This word is contextually flexible and has numerous senses, but for this example, we will consider one particular sense: applying the word 'star' as a descriptor of celebrity, fame, or excellence. Using this example, Hoffman's contextual window of 1,000 words would likely pick up on specific information relating to this particular star (e.g., a context about a *film star* might contain mention of red carpets, paparazzi, magazines or reviews). This specific semantic information might be useful for making concrete or abstract representations of a word. By comparison, Johns' subreddit discourses would contain a wider variety of words alongside information relating to this star; perhaps the star in question is only tangentially related to the much larger discourse topic. As such, this would be less useful for capturing experiential information that links us to the meaning of a word, but rather gives a diversity metric more akin to document count. Johns' measure therefore considers broad contextual information in a similar way to document count, potentially at the loss of specific contextual information relating to abstractness. This pattern can also be seen in the correlations. Johns (2021) has a strong positive correlation with document count ($r = 0.83$), while for Hoffman et al. this relationship is moderate ($r = 0.51$). By contrast, Hoffman et al. correlates negatively with concreteness ($r = -.33$) and imageability ($r = -0.27$), while for Johns' (2021) these correlations are non-significant ($r = 0.05$; $r = 0.04$).

These results relate to the two types of semantic representation outlined by Vigliocco et al. (2009): experiential information (how we interact with words) and linguistic information (how words

occur within language). The distinction seen between Hoffman's and Johns' could indicate that experiential information appears in close proximity to words, while linguistic information is more broadly distributed throughout language. However, this proposal is complicated by the correlations observed between Johns et al. (2020) and lexical variables. Johns et al. (2020) correlated strongly with document count (r = 0.74), but also correlated positively with concreteness (r = 0.21) and imageability (r = 0.25). While some measures of semantic diversity reported in Johns et al.'s (2020) paper were based on broad contextual windows (i.e., books), the measure included in our analysis used a moving window of 20 words. That this narrow window has produced a significant correlation with richness variables fits our suggestion that close-proximity context contains experiential information, but this does not explain why the correlation is in the opposite direction to that of Hoffman et al.'s diversity measure. Further, if broad contexts are important to developing linguistic information similar to document count, then we would expect Johns et al.'s (2020) correlation with document count to be weaker. More research is needed to determine whether, and how, the contextual window used to calculate semantic diversity affects the nature of these measures.

Another key difference was that Johns' measure considered the producer of the word, by measuring distributional properties across users (individual Reddit account holders) and discourse (subreddits). High diversity words therefore reflect a large spread of individuals producing this word. This user-based component of semantic diversity might capture different constructs compared to Hoffman's availability. For example, it might be more closely related to frequency or familiarity, as we would expect highly frequent and familiar words to be produced by a large proportion of the population. This could be more akin to measures of frequency, as we would expect high frequency words to be distributed across the population, while low frequency words might be produced more by particular users. The correlations reported in Study 1 support this idea, with Johns (2021) correlating highly with Zipf frequency (r = 0.65) while Hoffman et al.'s measure showing a more moderate correlation (r = 0.41). However, Johns et al.'s (2020) measure, which did not consider

individual use of words, also correlated strongly with frequency (r = 0.78), suggesting that another factor could be responsible for the strong relationship between frequency and Johns (2021).

Considering the similarities in correlation patterns between Johns et al. (2020) and Johns (2021) despite their methodological differences, we can begin to pinpoint what might be driving these relationships. These two measures differ to Hoffman et al.'s and Cevoli et al.'s in the size and sample of their corpora. According to comparisons of linear-mixed effect models reported in Study 1, these measures provided the best fit to lexical decision performance. This highlights the importance of using a large and representative corpus, and is congruent with results that demonstrate the importance of both the size (e.g., Burgess & Livesay, 1998) and sample (e.g., Van Heuvan et al., 2014) on calculating other corpus-derived measures like word frequency.

Our results may also be taken as evidence for the differences between the polysemous account of diversity and the Semantic Distinctiveness Model. Both Johns et al. and Johns propose that semantic diversity reflects the lexical histories that words have developed through years of experience. Through correlations and structural equation modelling, this appears to be strongly related to measures of occurrence and shares a weak relationship with concreteness. By comparison, the continuous measure of polysemy proposed by Hoffman et al. appears to capture both occurrence and abstractness to a similar extent. However, our analyses also demonstrated that the relationship between theoretical frameworks and diversity statistics is complex. Applying this to diversity's effects, it is likely that both the polysemy captured by this measure and the updating of semantic representations are important; but semantic diversity measures might not be able to discriminate between these theoretical accounts. Further exploration is warranted to determine whether the differences reported between these conceptualisations of diversity are theoretical or methodological. In order to control fully for differences in the size and sample of the corpora, future

researchers may wish to calculate diversity values based on Hoffman et al.'s method and the Semantic Distinctiveness Model method from the same corpora.

In conclusion, semantic diversity differs due to the method of calculation and the source on which it is based. These differences should be considered when investigating semantic diversity further, and when comparing results of studies which use different measures. Semantic diversity effects are attributable to diversity itself rather than other co-occurring semantic variables; but there are different types of diversity, which tap into different lexical constructs. Hoffman et al.'s (2013) measure captures both occurrence and abstractness information, while Johns' (2021) measure is strongly related to occurrence and appears to be a better predictor of lexical processing. Our results highlight areas for future research. Differences observed between Hoffman's and Johns' measures indicate a potential difference between the distribution of experiential and linguistic information in language.

CHAPTER 3
ESTABLISHING VALUES FOR CONTEXT AVAILABILITY AND SENTENCE AVAILABILITY

Note: This chapter was published in *Journal of Cognition* in 2022 by Ellen Taylor, Kate Nation and Yaling Hsiao as 'Context availability and sentence availability ratings for 3,000 English words and their association with lexical processing'. This work has been edited for this thesis, and random slopes have been added to the linear effect models where possible. Otherwise, the reported analysis remains the same.

INTRODUCTION

After exploring different measures of semantic diversity, we turned our focus to a wider array of semantic variables. This led us to collect norms for context availability; a metric that refers to how easily people can think of an imagined situation or circumstance for a word (Altarriba et al., 1999). In Hoffman et al.'s (2013) principal components analysis discussed above, the itemset analysed was limited to 279 words, in part due to the small number of openly available norms for context availability (N = 325, Altarriba et al., 1999). While collecting ratings for context availability, we also developed a new variable 'sentence availability' that we hypothesised would capture linguistic variation rather than experiential word knowledge.

Words high in context availability (CA) are advantaged in lexical decision and word naming (Colombo & Burani, 2002; Schwanenflugel & Noyes, 1996), and these effects are not readily explained by other psycholinguistic variables (Schwanenflugel et al., 1988). CA is closely related to concreteness and imageability (Van Hell & De Groot, 1998; Schwanenflugel et al., 1992). Variation in

these measures is associated with variation in semantic richness (Yap et al., 2015), but this cluster of related variables appears to capture a different type of semantic information to corpus-derived variables which measure contextual variation (e.g., contextual and semantic diversity; for reviews, see Caldwell-Harris, 2021; Jones et al., 2017). Adelman et al. (2006) used the term contextual diversity to describe document count (the number of unique texts a word appears in across a large corpus) and found this to be a better predictor of lexical decision and word naming than frequency. Other researchers have derived measures of semantic diversity that capture semantic variability in the contexts in which a particular word is used, not just the number of unique texts (Hoffman et al., 2013; Johns et al., 2012). While semantic diversity is also associated with lexical processing across a range of tasks (e.g., Hoffman & Woollams, 2015; Hsiao et al., 2020), there is some debate about its nature (see Chapter 2 for discussion). No matter what theoretical approach is taken, there is strong evidence that contextual experience shapes lexical organisation beyond both raw frequency and document count (Cevoli et al., 2021; Hoffman et al., 2013, Johns & Jones, 2022).

Our starting point in this chapter is with the relationship between context availability (CA) and semantic diversity. Both measures index contextual experience, but in rather different ways, with CA being a subjective measure of how easily a situational context can be activated, and semantic diversity being derived from large corpora based on linguistic context. The relationship between these variables is unclear in the literature. In their analysis of 200 concrete and 200 abstract words, Moffat et al. (2015) reported a positive correlation between rated CA and semantic diversity. In contrast, Hoffman et al. (2013) reported a negative correlation between the two variables, based on 279 words. These contradictory findings might be due to limitations in the number and range of words sampled. To date, the largest set of openly accessible CA norms in English contains only 325 words (Altarriba et al., 1999). Much larger itemsets are now commonplace for other variables, allowing researchers to explore relationships with other lexical statistics and with behavioural data from megastudies (e.g., Keuleers & Balota, 2015). Our first aim was to produce CA

ratings for a larger set of words, namely the 3,000 English words in Cortese and Fugett's (2004) imageability norms, to facilitate research on CA and its relationship with other lexical variables.

Further exploration of larger data sets is warranted, but it seems likely that the linguistic context captured by a word's semantic diversity is not the same as its rated context availability. Words may occur across diverse contexts in large language corpora for different reasons: as noted by Hoffman et al. (2020), high semantic diversity words are associated with a range of semantic states by virtue of their contextual promiscuity. Some words may be high in semantic diversity because they are polysemous (e.g., *glass*) and therefore have flexible meanings depending on the context. Other instances of high semantic diversity include function words that also depend on context for precise meaning. Given this semantic flexibility and openness, it is not surprising that there is a negative correlation between semantic diversity and measures of semantic richness such as imageability and concreteness (according to Hoffman et al. 2013, r = -.48 and r = -.51, respectively). In contrast, CA is positively associated with concreteness and imageability (Moffat et al., 2015) as words high in these variables can more easily activate situational contexts.

As well as indexing different aspects of context, it is important to note that semantic diversity and CA are very different types of measure. Semantic diversity, while objective, is derived from large corpora that do not reflect the language experience of any individual. By contrast, CA is obtained from participant ratings and is necessarily subjective. The two measures both reflect contextual experience but tap into two different aspects of context. Semantic diversity taps words' linguistic contexts, while context availability taps the imagined context words are associated with, which appears to be grounded in experiential representations. With this observation as a backdrop, our second aim was to investigate a new variable, sentence availability (SA). In contrast to CA in which participants rate how readily a context or circumstance comes to mind, SA invites people to indicate how easy it is to think of a sentence for each word. We theorised that, by changing

instruction from '*context or circumstance'* to '*sentence'*, participants would make reflections based on linguistic contexts rather than imagined contexts. Conceptually then, SA can be seen as a participant-based metric of linguistic experience and was proposed to capture lexical histories in a similar way to semantic diversity.

Our aim was to develop a measure of linguistic availability based on individual subjective reporting. This would allow us to directly compare situational and linguistic availability for the same words, and to consider how both relate to other lexical variables, and to lexical processing itself. This therefore allowed us to further explore the idea of distinct facets of semantic knowledge (for discussion, see Pexman et al., 2007).

As with our analysis in Chapter 2, we included a range of lexical variables to help understand how CA and SA operate in lexical-semantic space. Core variables included frequency and document count, both of which correlated positively with CA and semantic diversity in previous studies. Words learned early in life tend to have high CA (Hoffman et al., 2013; Moffat et al., 2015) and they also tend to be higher in semantic diversity. We therefore included age of acquisition (Kuperman et al., 2012). Including imageability and concreteness allowed us to test the relationship between both types of availability and semantic richness. For completeness, we also considered variables that tap emotion and embodiment, including measures of valence (a word's pleasantness), arousal (the intensity of emotion associated with a word), dominance (the extent to which a reader feels influential or in control in response to reading a word, ranging from *controlled* to *in control*) from Warriner et al. (2013), and body-object interaction (BOI, the ease with which a word can be physically interacted with) from Tillotson et al. (2008).

Figure 3.1 summarises the predicted relationships between our core variables of interest. The relationship between CA and concreteness is well established (Van Hell & De Groot, 1998; Schwanenflugel et al., 1992). We therefore expected CA to correlate positively with concreteness

and imageability. Sentence availability might not be so closely aligned with concreteness and imageability because it draws upon linguistic contexts. We therefore predicted that SA would be related to variables capturing linguistic variation, such as semantic diversity. To pre-empt our findings with an example, 'braid' had high CA relative to SA, suggesting that it can be associated with a context more easily than a sentence. In contrast, 'fleck' was more easily placed in a sentence than a context (see Figure 3.3 for further examples). 'Braid' is imageable and concrete, but only appears in a limited number of contexts. 'Fleck', however, is more difficult to associate with a particular context, but it can be placed in a sentence with relative ease. It is less imageable than 'braid' but is more diverse as it appears in a wider range of linguistic contexts.

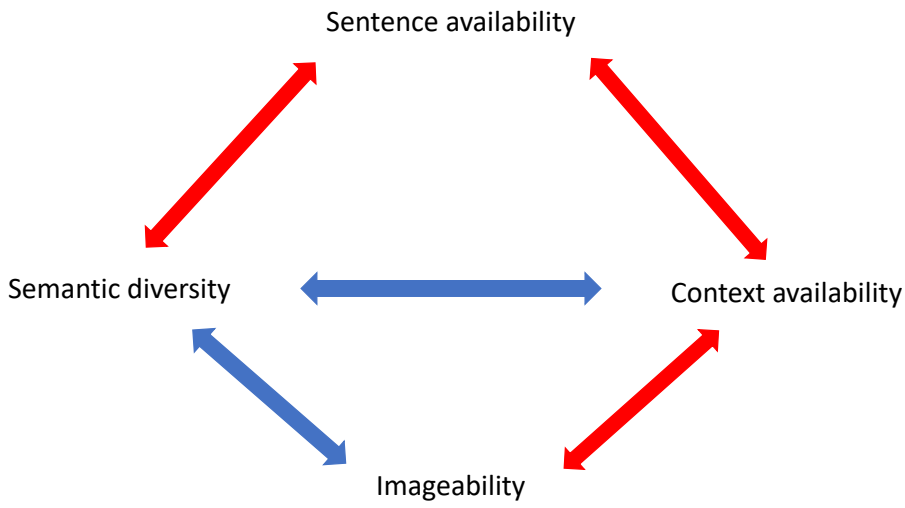


Figure 3.1: Predicted relationships between context availability, sentence availability, imageability and semantic diversity. Negative relationships are shown in blue, and positive relationships in red.

Based on findings reported by Hoffman et al. (2013) and Moffat et al. (2015), we expected CA to correlate positively with word frequency and document count, and negatively with age of acquisition. As discussed by Pexman (2020), semantic concepts are learnt largely through sensorimotor association and, in the absence of these associations, meaning can be grounded in felt experience. In line with this, Moffat et al. (2015) reported positive correlations between CA and body-object interaction and emotional experience. Given SA might be more associated with linguistic diversity, we anticipated that it would be less associated with emotion and embodiment variables than CA.

Finally, we tested whether CA and SA are associated with lexical processing, using lexical decision data from the British Lexicon Project (BLP; Keuleers et al., 2012) and the English Lexicon Project (ELP; Balota et al., 2007). We hypothesised that both types of availability would facilitate lexical decision, and both would explain unique variance in performance.

METHOD

PARTICIPANTS

Eight-hundred native speakers of British English were recruited via Prolific (Palan & Schitter, 2018) to provide either context availability or sentence availability ratings. After exclusions (see below) we analysed data from 359 people rating contexts and 365 rating sentences.

MATERIALS AND PROCEDURE

Ratings were collected for the 3,000 English monosyllabic words in Cortese and Fugett's (2004) imageability norms, programmed using the online platform Gorilla (www.gorilla.sc). We

created 20 150-word lists, sampled from the 3,000 itemset. Each list contained six non-words, randomly placed to check for attention. Participants in each version (CA or SA) were randomly assigned to one of the 20 lists. The words were presented in random order in six blocks of 26 words. They were asked to rate each word on a 7-point Likert scale. The lower end of this scale was labelled 'easy to think of a context' (CA) or 'easy to think of a sentence' (SA). The other end of the scale was 'difficult to think of a context/sentence'. Participants were told to skip words they did not know. Instructions for both tasks are shown in Appendix 2. Due to randomisation via Prolific, some word-lists were allocated additional participants. This resulted in each word being rated by a maximum of 21 people for context availability and 20 people for sentence availability.

RESULTS

Participants who provided ratings for more than 30% of the catch non-words or gave the same rating for 95% of trials were removed, N = 41 (CA) and 35 (SA). Words with less than 15 valid ratings for both variables (N = 509, 16.97% of data) were then excluded from the analyses reported here (but data for all 3,000 words are available at https://osf.io/gr73b/). The mean (SD) number of observations per word was 17.73 (1.70) for CA and 17.81 (1.48) for SA. Ratings were reverse coded to follow the same direction as Altarriba et al.'s norms, with higher scores reflecting higher availability. Both CA and SA were normally distributed (Figure 3.2) but skewed towards the upper end of the scale. They correlated with each other, $r = .643$, $p < .001$, and CA correlated with CA for the same items in Altarriba et al.'s dataset, N = 117, $r = .645$, $p < .001$ (for further information see Appendix 3).

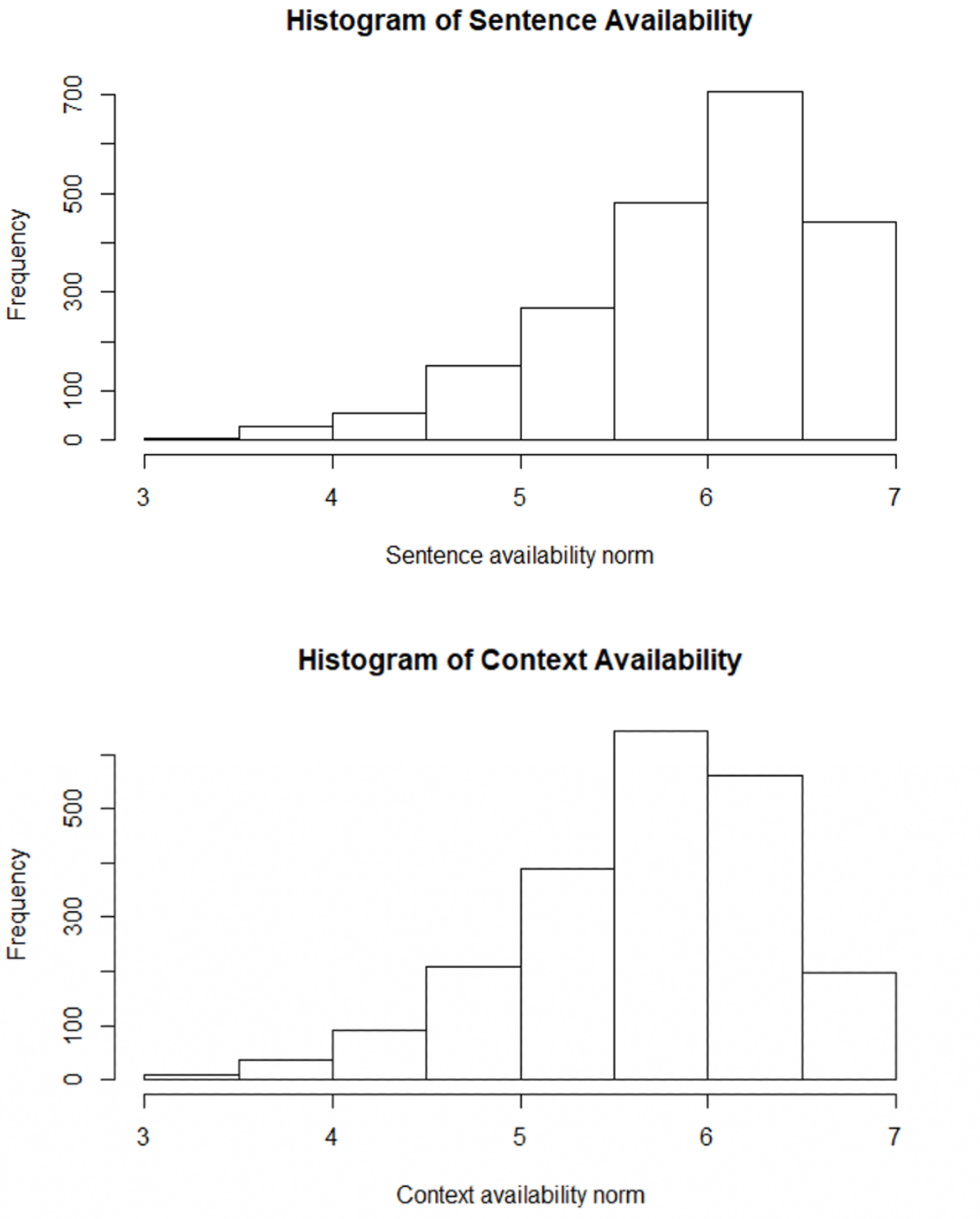


Figure 3.2: Histograms showing the distribution of CA (M = 5.69, SD = 0.68) and SA (M = 5.94, SD = 0.67).

Availability norms are plotted against each other in Figure 3.3, with examples labelled to illustrate the relationship between CA and SA. Figure 3.4 shows Pearson correlations between CA, SA, and a range of other variables. Of note, there was a positive correlation between CA and imageability, r = .59, p < .001, but not between CA and semantic diversity, r = <-.001, p = .945. In contrast, SA was positively correlated with both imageability, r = .28, p < .001, and semantic diversity, r = .34, p < .001. Both availability measures correlated positively with frequency (r = .36 for CA; r = .66 for SA) and negatively with age of acquisition (r = -.62 for CA; r = -.70 for SA).

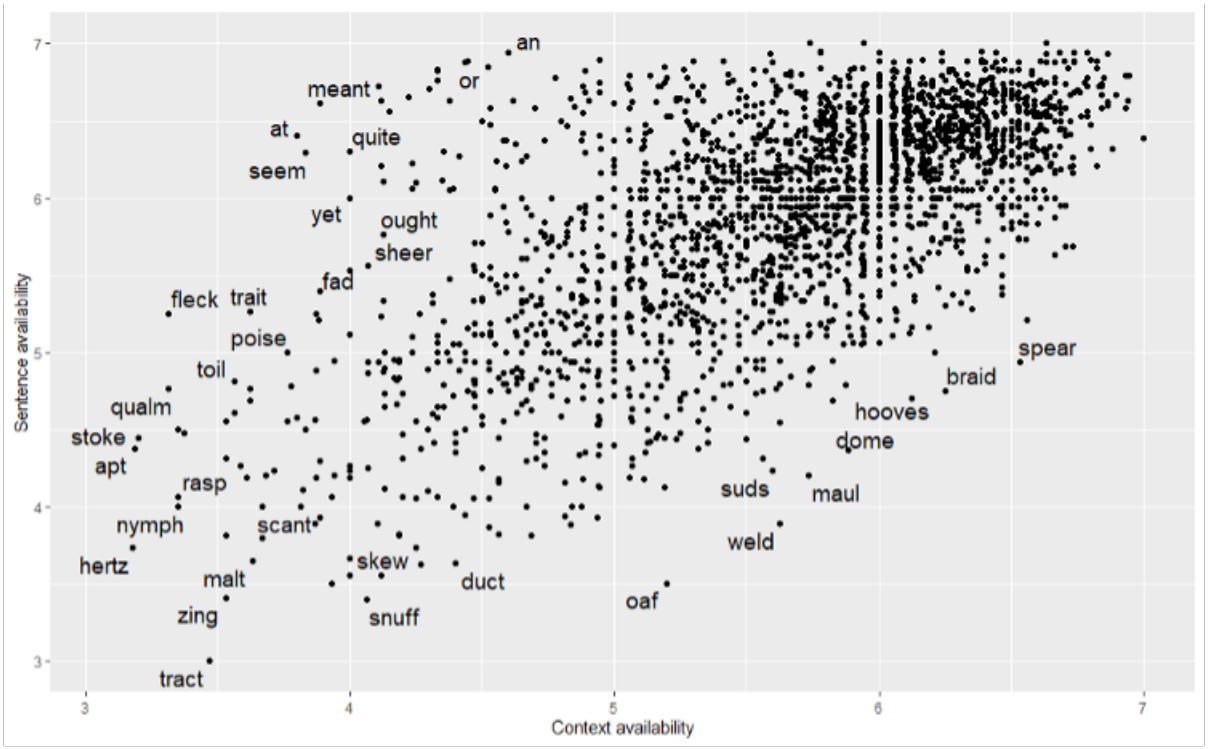


Figure 3.3: Scatterplot showing the relationship between context availability and sentence availability, with examples labelled.

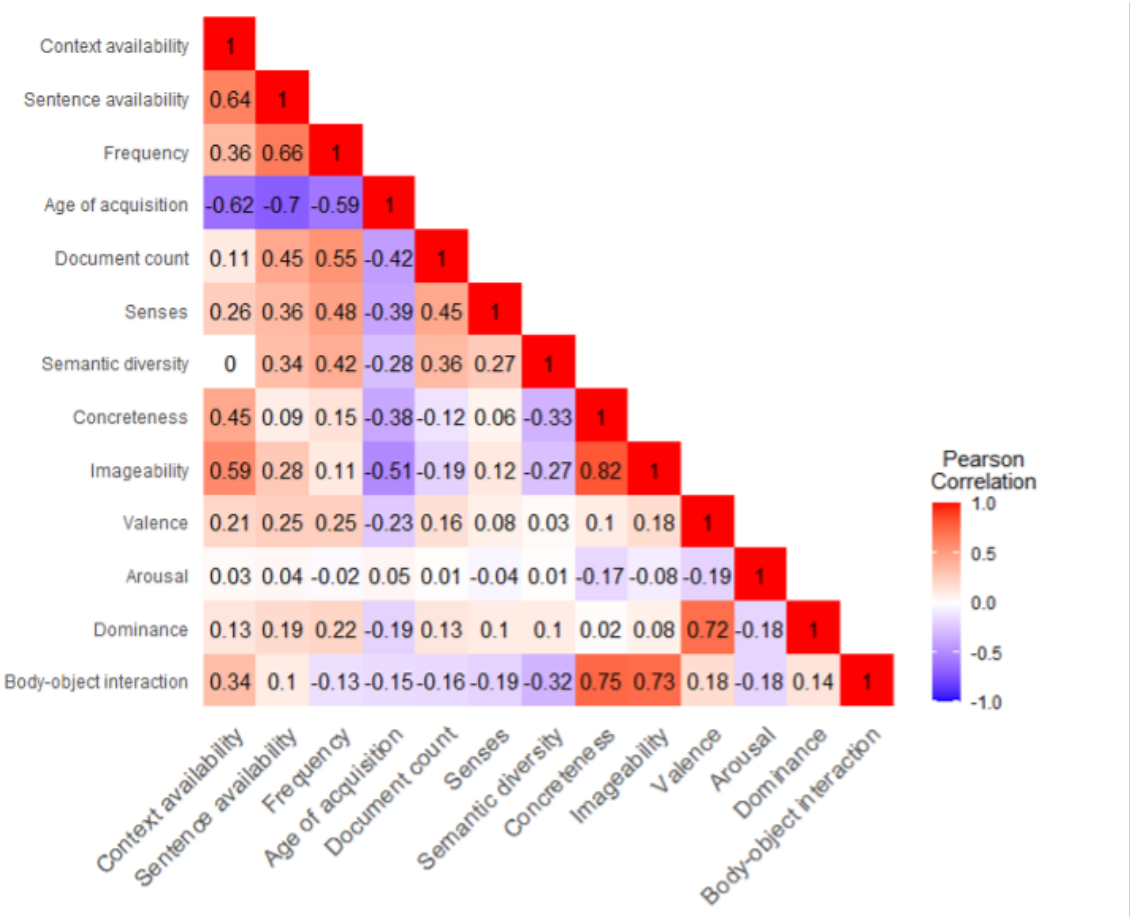


Figure 3.4: Pearson correlations for CA, SA and other lexical variables, based on N available for each pairwise correlation.

Having observed no correlation between CA and semantic diversity, we checked the relationship for those items also included by Hoffman et al. (2013). They reported a negative correlation (N= 325; r = -.26), an observation that replicated for the N=117 Hoffman words also included here, r = -.36. This observation indicates that the itemset used in Hoffman et al.'s analyses might be too small to accurately reflect the relationship between CA and semantic diversity.

To consider the relationships between the variables further, we used factor analysis with oblique rotation, using the function 'oblimin' in R Package 'Psych' (Revelle, 2017). Similar results were seen following varimax rotation. Parallel analysis identified three factors (Table 3.1; see Figure 3.5 for scree plot) that we labelled as 'occurrence' which captured variables relating to frequency and ambiguity; 'richness' which captured imageability and concreteness information; and 'affect', with positive loadings from valence and dominance. Occurrence explained 24.3% of the variance, richness 22.5%, and affect 11.5%. There were small but significant correlations between the factors. CA loaded positively onto both occurrence and richness, while SA loaded only on occurrence.

Table 3.1: Factor loadings for an exploratory factor analysis using oblique rotation.

	Occurrence	Richness	Affect
CA	0.501	0.487	
SA	0.801		
Zipf Freq	0.801		
AoA	-0.774		
Doc Count	0.626		
Senses	0.574		
SemD	0.497	-0.424	
Concreteness		0.874	
Imageability		0.936	
Valence			0.824
Arousal			
Dominance			0.857
BOI		0.808	

	Occurrence	Richness	Affect
Occurrence		.15	.24
Richness	.15		.12
Affect	.24	.12	

Note: Only words with data for all variables are included (N = 1,270). Loadings less than 0.3 are not displayed; lower part of table shows correlation between factors.

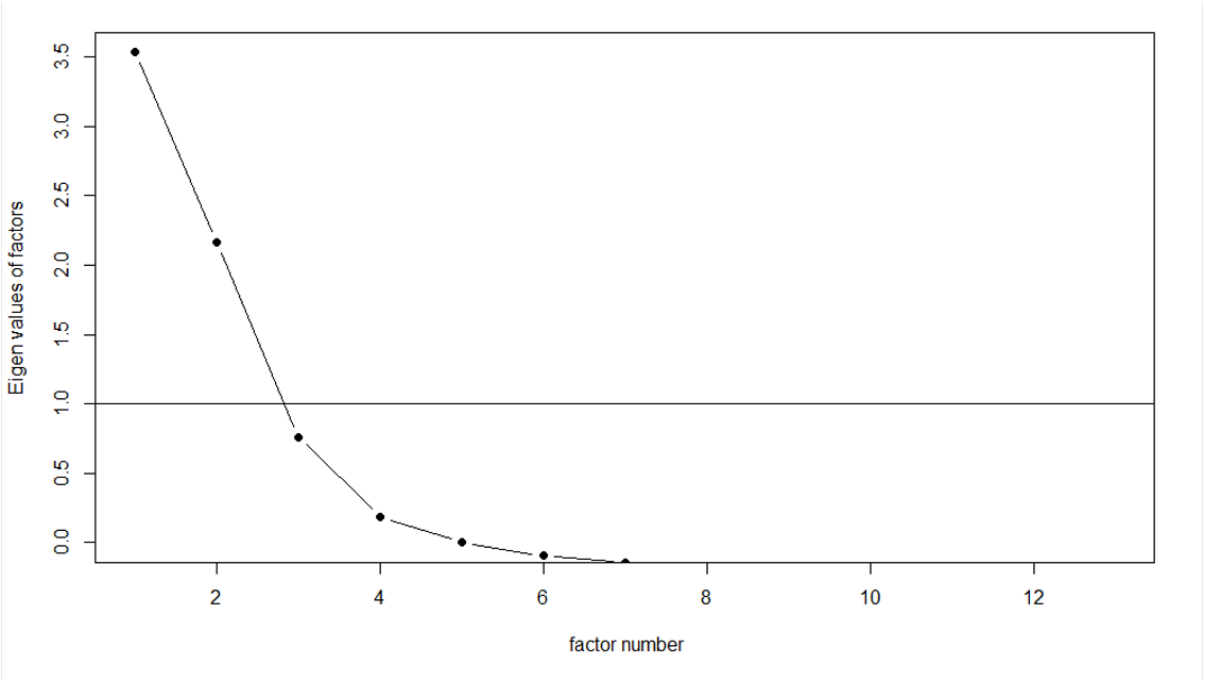


Figure 3.5: Scree plot showing the factors identified by parallel analysis. Note that only Factors 1 and 2 had Eigen values above 1.0.

We then investigated whether CA and SA predicted lexical decision, using linear mixed effects models (lme4 in R, Bates et al., 2007; p values from lmerTest, Kuznetsova et al., 2017) and data from the BLP (Keuleers et al., 2012). Accuracy was analysed using glmer and the RT analysis used lmer. Both models included frequency, age of acquisition, semantic diversity and imageability (variables that influence lexical decision) as well as CA and SA. Random intercepts of word were included, and all predictor variables were centered and scaled. For predicting BLP accuracy data, by-

participant slopes were included for CA and SA. Including slopes for other lexical variables resulted in the model failing to converge. This therefore tested the model:

Accuracy ~ frequency + age of acquisition + semantic diversity + imageability + CA + SA + (CA + SA | participant) + (1 | word)

Fixed effects of SA and CA were then removed from the model for comparison.

For predicting ELP accuracy and both BLP and ELP response time, it was not possible to include any by-participant slopes as the model failed to converge. Therefore, models predicting RT data instead included random interceptions:

Performance ~ frequency + age of acquisition + semantic diversity + imageability + CA + SA + (1 | participant) + (1 | word)

with either CA or SA omitted for comparison.

As shown in Table 3.2, both availability measures were associated with lexical decision performance. Comparisons demonstrated that this model explained significantly more variance than when only one of the availability measures was included (for RT: $x^2(1) = 43.301$, $p < .001$ compared

to CA alone; $x^2(1) = 58.062$, $p < .001$ compared to SA alone. For accuracy: $x^2(1) = 18.06$, $p < .001$ compared to CA alone; $x^2(1) = 12.70$, $p < .001$ compared to SA alone).

Table 3.2: Fixed effects predicting lexical decision accuracy (upper section) and RT (lower section in BLP data (N = 1724 words).

	Estimate	SE	z	p
Intercept	2.997	0.139	21.308	<.001
Freq	0.462	0.065	7.151	<.001
AoA	0.064	0.067	0.956	.339
SemD	0.147	0.048	3.025	.002
Imageability	0.274	0.051	5.425	<.001
CA	0.229	0.060	3.806	<.001
SA	0.319	0.069	4.594	<.001

	Estimate	SE	df	t	p
Intercept	586.037	11.259	37.502	52.052	<.001
Freq	-18.910	1.628	2054.430	-11.615	<.001
AoA	5.007	1.723	2042.398	2.905	.004
SemD	-3.538	1.325	2094.082	-2.671	.008
Imageability	-4.962	1.256	2039.830	-3.951	<.001
CA	-10.991	1.435	2068.772	-7.660	<.001
SA	-10.501	1.590	2092.945	-6.602	<.001

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + context availability + sentence availability + (1|participant) + (1|word).

The generalisability of these results was tested using data from the ELP (Balota et al., 2007). Once again, both availability measures were associated with lexical decision performance (Table

3.3).

Table 3.3: Fixed effects predicting lexical decision accuracy (upper section) and RT (lower section in ELP data (N = 2119 words).

	Estimate	SE	z	p
Intercept	-2.44269	0.45169	-5.408	<.001
Freq	0.42106	0.04631	9.092	<.001
AoA	-0.05054	0.01619	-3.122	.002
SemD	0.68378	0.10889	6.279	<.001
Imageability	0.28597	0.02624	10.896	<.001
CA	0.14105	0.05627	2.507	.012
SA	0.22429	0.05900	3.802	<.001

	Estimate	SE	df	t	p
Intercept	977.2004	25.2271	2236.1832	38.736	<.001
Freq	-28.7873	2.3931	2103.7453	-12.029	<.001
AoA	3.2613	0.8641	2099.2646	3.774	<.001
SemD	-29.0695	5.9157	2098.5337	-4.914	<.001
Imageability	-12.0402	1.3371	2098.8602	-9.005	<.001
CA	-7.3309	2.9978	2099.0068	-2.445	.015
SA	-14.1835	3.2280	2100.8502	-4.394	<.001

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + context availability + sentence availability + (1|participant) + (1|word).

DISCUSSION

Context and sentence availability values were collected for 3,000 monosyllabic words. While these variables correlated positively (r = .64), changing the instruction to focus on sentences rather

than contexts did change the nature of availability. The two variables showed a different pattern of correlations with other variables. CA was positively correlated with imageability but not correlated with semantic diversity whereas SA was positively correlated with both diversity and imageability. SA showed strong relationships with a range of variables (including frequency, age of acquisition, and document count), while CA was closely related to concreteness and imageability. Three factors were identified by factor analysis, reflecting 'occurrence', 'richness', and 'affect'. As anticipated, CA loaded onto richness, but also onto occurrence. Semantic diversity also loaded in the same direction as CA for occurrence, but the opposite direction for richness. This highlights a more complex relationship between CA and semantic diversity than apparent from the pattern of simple correlations. The observation that SA loaded onto occurrence supports the hypothesis that it might capture information relating to textual variation and thus reflects word usage in sentences. This is also consistent with its correlation with semantic diversity. In contrast, ratings of concreteness and imageability reflect the semantic richness of words, and this might be less directly related to linguistic diversity, where more abstract or underspecified words tend to occur across varying contexts. While we collected a large set of availability ratings, we did not ask participants to produce the imagined sentences they associated with target words. A sentence production task would allow for a more thorough investigation of this.

Both availability measures were negatively correlated with age of acquisition, with words learnt earlier in life being easier to contextualise (see Hills et al. (2010) for broader discussion of contextual diversity and early word learning). It is possible that other variables closely related to age of acquisition might be influencing how participants rate words for CA and SA. For instance, word familiarity is also negatively correlated with age of acquisition (Stadthagen-Gonzalez & Davis, 2006). Greater familiarity might make words easier to associate with sentences and contexts. Future work might therefore wish to explore the relationships between these norms and a wider range of lexical measures.

Both CA and SA predicted lexical decision performance. This finding held for both BLP and ELP datasets, demonstrating that high availability is associated with more efficient lexical processing. The contribution of SA could be driven by shared variance with other lexical variables (e.g., number of senses, valence, dominance, and arousal) which were not included in our model, or by the high SA ratings for function words such as 'or' (6.85), 'an' (6.94) and 'at' (6.40; see Figure 3.3). SA tends to be high both for function words which are contextually flexible, and for words which are semantically rich. CA is also associated with semantic richness, but tends to be lower for function words as these are harder to place in a situation context than content words. In a post-hoc analysis, we identified 83 function words in our dataset. All analyses patterned identically after excluding these words (for details, see Appendix 4). This suggests that function words are not skewing the pattern of results seen here. The finding that the two availability measures account for separate variance in lexical processing supports the conclusion drawn from the correlational analysis and factor analysis, namely that SA and CA reflect different types of availability information.

While SA was associated with semantic diversity, there are clear differences between these measures: SA was correlated with CA and imageability, while semantic diversity was uncorrelated with CA and negatively correlated with imageability. In other words, words that easily arouse imagery may be easy to place within sentences, but they might not be semantically flexible across contexts. Although SA did not load onto 'richness' alongside imageability, it is nonetheless positively correlated with this variable, suggesting that SA captures information other than linguistic variation. As SA is based on participant ratings, it might be more influenced by the experiential information attached to the to-be-rated words, relative to semantic diversity which is corpus-derived and therefore reflects linguistic information.

Before closing, we should note some limitations of our study. For both measures, participants gave high ratings, suggesting that most words were easily associated with a context or

sentence. While the norms collected by Altarriba et al. (1999) were also rated highly, this was more pronounced in our study. This might reflect participant differences, or differences across the item-sets; following Cortese and Fugett, our words were all monosyllabic, but as those authors noted, many studies tend to focus on monosyllabic words, making this a useful item-set for future experiments. That said, there is a clear need to extend the number and range of words considered, not least because multisyllabic words are more complex and may relate to availability within semantic space in different ways. We note too that we excluded words that received less than 15 ratings. There is little consensus in previous studies as to the number of participants needed to produce valid ratings. In Kuperman et al.'s (2012) age of acquisition norms, each word had 18 or more ratings, drawn from a large participant pool. A different approach is for a number of people to rate all words (e.g., 31 participants for Cortese & Fugett, 2004; 78 for Altarriba et al., 1999). In our study, participants had the option to skip words that they did not know, adding reassurance that the ratings were a valid reflection of availability.

In conclusion, CA and SA are two distinct measures that capture word knowledge and word usage in different ways. CA captures information similar to concreteness and imageability, while SA is more closely related to textual or linguistic variation. Words high in either type of availability showed processing advantages in lexical decision, and each explained unique variance when other key lexical variables were controlled.

CHAPTER 4
LONG-TERM AND RECENT EXPERIENCE ON PRODUCTION AND COMPREHENSION

INTRODUCTION

The experimental work so far in this thesis has focused on indices of long-term lexical experience. The linguistic and semantic information we associate with words is developed through contextual experience, with diverse encounters over time providing a word with rich and nuanced lexical histories (Nation, 2017; Jones et al, 2012). From an early age, humans are sensitive to patterns in language. Infants extract statistical information to aid learning of complex linguistic input (Romberg & Saffran, 2010). For both older children (Hills et al., 2009) and adults (Mak & Twitchell, 2020), words are more easily acquired when they are highly semantically connected to other words in the lexicon. Evidence from the literature on contextual and semantic diversity supports this view as variance in lexical processing performance is affected by the spread of contexts that words appear in (e.g., Hoffman & Woollams, 2015; Johns et al, 2020). Lexical statistics such as frequency, semantic diversity, and age of acquisition can therefore be seen as estimates of the linguistic knowledge built across years of language experience.

Research has also highlighted the importance of recent, short-term linguistic experience on lexical processing. MacDonald's (2013) Production-Distribution-Comprehension account argues that when producing language, we choose words for how easily they can be retrieved from the lexicon, and reuse utterances or structures that we have previously been exposed to. Readily available words or structures are often more frequent, shorter, or less syntactically complex. When we encounter language in our environment, we retrieve linguistic information from memory in order to comprehend the words, phrases or utterances. This retrieval makes the encountered language more

readily available. Therefore, when we produce words, phrases, or utterances, we are more likely to produce those we have recently encountered. This process works both ways, with recent comprehension aiding subsequent production, and recent production aiding the ease with which we can comprehend words. The effects of recent experience can be seen in priming studies, in which recently encountering words typically leads to more efficient subsequent processing (e.g., Tenpenny, 1995; Stark & McClelland, 2000).

The current study aimed to investigate the effects of both long-term and recent lexical experience on lexical processing. This builds upon Chapter 3, in which we proposed the new variable *sentence availability*. Sentence availability explained unique variance in lexical decision performance, but further investigation is warranted to understand how this variable relates to other lexical behaviours. We employed sentence availability as a measure of long-term linguistic experience. This allowed us to validate the measure, and to further develop our understanding of this variable's effects on lexical processing. We also used a priming paradigm to explore the extent to which long-term and recent experience interact. We tested whether the latency to make sentence availability ratings would change if words had been encountered before. We also tested whether reading latencies would be different for words with high vs low sentence availability.

LONG-TERM LEXICAL EXPERIENCE

As discussed in Chapter 1, long-term linguistic experience is captured by lexical variables. Word frequency reflects how many times we have encountered a word and is a known predictor of lexical processing (Brysbaert et al., 2018). Semantic diversity, capturing the similarity of contexts in which words have appeared over time, facilitates lexical decision (Hoffman & Woollams, 2015). Words with more semantic neighbours, more features and more associated contexts show benefits in processing across numerous lexical tasks (Yap et al., 2011). Lexical variables can be based on

analyses of language use as collected in large corpora (e.g., semantic diversity, Hoffman et al., 2013; Zipf frequency, Van Heuven et al., 2014), or compiled from participant ratings (e.g., imageability, Cortese & Fugett, 2004; concreteness, Brysbaert et al., 2014).

We introduced the concept of sentence availability in Chapter 3. It reflects how easily a word can be associated with a sentence and is measured on a Likert scale. This is similar to 'context availability' which reflects the ease with which words can be associated with contexts but, in sentence availability, the focus is on sentence context rather than an imagined circumstance. Sentence availability can be viewed as a marker of long-term lexical experience, in the same way as other rated variables such as age of acquisition, imageability and concreteness. Chapter 3 found that sentence availability correlated positively with measures of frequency and ambiguity, including semantic diversity; it also explained unique variance in lexical decision performance. We proposed that sentence availability likely captures information about a word's textual use in a similar way to corpus-derived measures like document count and semantic diversity.

While lexical variables provide a useful estimate of a word's history, the information we associate with words is not a static value, but a dynamic construct which is susceptible to change dependant on individual experience. As proposed by Jones et al.'s (2012) Semantic Distinctiveness Model, semantic representations are updated whenever words are encountered in a new context. Therefore, while corpus-derived metrics and participant-based ratings give an estimate of the population's experience, these values should not be viewed as unchangeable. Individual differences exist in the fit of lexical variables to processing: the effects of imageability vary with phonological coding skill (Strain & Herdman, 1999), and the effects of word length and frequency vary with vocabulary knowledge (Butler & Haynes, 1979). Rodd et al. (2016) reported that the interpretation of ambiguous words also varies dependent on individual experience: for homonyms with both rowing related and non-rowing related meanings (e.g., *feather*), rowing experience increased the

likelihood of interpreting the word in its rowing-related meaning. This demonstrates that lexical measures are not constant and universal measures but are largely dependant on individual experience.

As well as long-term differences in linguistic experience as indexed by measures like vocabulary, MacDonald's Production-Comprehension-Distribution theory asserts that words recently accessed from semantic memory will receive a processing benefit. This might also relate to Jones et al.'s (2013) Semantic Distinctiveness Model, in which words encountered in contexts give an 'updating' of vector representations. Though Jones et al. present this to explain the accrued effects of long-term experience, the active updating of word knowledge might result in small changes in subsequent experience. We therefore turn to the research investigating the influence of recent experience on lexical processing.

RECENT LEXICAL EXPERIENCE

Knowledge can be more easily accessed in memory if it has been recently retrieved (Ratcliff & McKoon, 1988). In the context of language processing, the effects of recent experience have been investigated through priming studies. Here, participants are primed to an item (e.g., encountering a target word via incidental reading or exposure to a related word in a lexical processing task) and are then tested on subsequent, primed behaviour. In priming paradigms, repetition tends to speed lexical decision response times, with responses faster for words previously encountered (Forbach et al., 1974; Scarborough et al., 1977). This effect also transfers to word reading. Lowder and Gordon (2017) found that repetition of words embedded in sentences led to shorter gaze durations, suggesting faster lexical processing. Priming effects were modulation by scores on the author recognition test, with participants scoring lower on print exposure showing greater boosts in

performance from repetition. These results suggest that recent linguistic experience influences comprehension of words; and that this experience interacts with prior reading experience.

According to MacDonald, both language production and comprehension affect and are affected by lexical processing. As well as the priming effects on processing during comprehension discussed above, recent experience appears to shape production. Using a syntactic priming paradigm, Branigan et al. (2006) found that language production was influenced by recent lexical experience. When participants read sentence fragments comprising a particular syntactic structure, they subsequently used the same structure in their own utterances. Loebell and Bock (2003) found that syntactic priming effects transfer across languages in bilingual speakers. Here, participants that read sentences in their first language were likely to use the same syntactic structure when producing another sentence in their second language. Phonological priming also appears to influence language production. For example, James and Burke (2000) found that participants were more likely to respond correctly to a general knowledge question when they had previously heard a word phonologically similar to the answer. This suggests that lexical retrieval is facilitated via the access of connected words, supporting connectionist models of lexical organisation.

Priming studies have also investigated the role of contextual information in processing ambiguous words (Rodd et al., 2013; Betts et al., 2018). Rodd et al. (2016) demonstrated that experience can affect retrieval of ambiguous word meanings in more naturalistic environments. Rowers were more likely to interpret rowing-related meanings for ambiguous words (e.g., *bump* or *feather*) if they had rowed earlier that day. This effect was mediated by participants' age and rowing experience, with rowers who had gained lots of experience at a young age most likely to interpret the ambiguous words with rowing-related meanings. This study demonstrates how differences in linguistic experience can lead to differences in lexical processing.

These results all provide general support for MacDonald's Production-Distribution-Comprehension theory. According to this theory, producing or comprehending a word will make that word easier to retrieve from memory, thus boosting the ease with which it can be subsequently produced or comprehended. This process could explain the effects of repetition priming. Words, phrases, or meanings that are encountered during priming become more readily accessible and are therefore more efficiently processed during subsequent language processing. Priming effects can also be understood through Ratcliff and McKoon's (1988) retrieval theory. This proposes that, during retrieval, both the prime and target are combined to make a compound cue. The more closely related the prime and target are, the stronger this cue is: therefore, an exact repetition serves as a stronger cue to recall than when prime and target are less strongly related. Lester et al.'s (2017) investigation of contextual similarities in prime-target pairs found that, the more similar the syntactic context the prime was to the target, the greater the priming effect was on visual lexical decision. This supports a probabilistic network account of language, in which connections between words or syntactic structures are fine-tuned through experience. Together, these studies suggest that people continuously accumulate linguistic experience and draw on it during the course of processing. As in the Production-Distribution-Comprehension account, recent experience shapes how we process the language that we subsequently encounter.

Similarities can be seen here between recent exposure and theories of long-term language experience relevant to lexical organisation. In Nation's (2017) lexical legacy hypothesis, lexical representations of words are developed via numerous encounters across contexts. Similarly, Jones et al.'s (2012) Semantic Distinctiveness Model proposes that word representations are 'updated' with every encounter, weighted by the similarities of each contextual use. The encounters with words within a priming paradigm might be viewed as opportunities for the 'live' updating of word representations. A key example of this is shown by Rodd et al.'s (2016) findings in which word meanings appear to have been recently updated to consider the latest contextual encounter.

However, there are also key differences between these theoretical accounts of lexical histories and the priming literature. The theoretical accounts of contextual histories (e.g., Jones et al., 2012; Nation, 2017) rely heavily on the varied contexts surrounding words, whereas many priming studies present words in a single context. As such, the live updating of lexical representations reflected during priming paradigms are limited and might not capture the complex contextual histories that are developed across time. Similarly, models of lexical organisation consider the knowledge a word has accrued over years of experience, while priming studies measure short-term, almost immediate effects. While there is evidence that some priming effects are more enduring (e.g., effects of ambiguity priming lasting 20 minutes in a lab experiment, Rodd et al., 2013, or 8 hours with rowers interpreting rowing-related ambiguous words, Rodd et al., 2016), the majority of word priming studies only measure the short-term influence of priming and do not consider whether this type of encounter may sum to lexical knowledge over a longer term.

While the research on recent experience is vast, less is known about how this interacts with long-term experience, as indexed by lexical variables. Small differences in the magnitude of priming effects are observed between high and low frequency words in fragment-completion experiments (MacLeod, 1989) and recognition memory tasks (Jacoby & Dallas, 1981). In these studies, the magnitude of priming effects is larger for low than high frequency words. Ostergaard (1998) argued that low frequency words are less available for retrieval, and therefore can receive more of a boost through priming. Seen through the lens of the Semantic Distinctiveness Model, this heightened priming effect for low frequency words could be because these words have had less opportunity to be updated, and their vector representations are therefore weighted more heavily by the last encounter than a high frequency word. However, Ostergaard (1998) reported that frequency-related differences in priming effects on word naming tasks are small and not always statistically significant; research is also lacking into the relationship between priming and other markers of long-term lexical experience, such as those which consider a word's lexical history. The current study therefore aimed

to investigate the effects of both long-term and recent experience on processing, and to test whether an interaction exists between these effects. This study was preregistered on As Predicted (aspredicted.org, see Appendix 6).

THE CURRENT STUDY

This aimed to test the effects of long-term and recent lexical experience using cross-model priming. We included a comprehension task (self-paced reading) and a production task (generation of sentence availability ratings). This allowed us to test whether there are differences in the effects of long-term and recent experience on different facets of language processing. Words differed in exsisting sentence availability, using ratings collected in Chapter 3. This gave a variation in long-term lexical experience. Short-term experience was measured by comparing responses to words during their first encounter (unprimed; either in a production task or a comprehension task) with responses to the same words in the alternative task (primed). To do this, participants encountered the same words in two tasks. The 'comprehension' task used a self-paced reading paradigm, allowing us to measure word-level processing in a naturalistic sentence-reading task. This technique has been implemented in previous priming experiments (e.g., Luke & Christianson, 2013; McNamara & Healy, 1988). During this, we measured reading times of the target word and the following 'spill-over' word, as this could reflect participants' delayed reading of the target (Mitchell, 1984). The 'production' task involved asking participants to rate words for sentence availability. Here, we instructed participants to mentally produce a sentence using the target word before making a judgement of how easy this sentence was to produce. This allowed us to collect response time as an index of production as well as rating scores.

To assess the influence of long-term experience, we tested whether there would be a difference in lexical processing for high vs low availability words during sentence reading and sentence production rating. As words differed in sentence availability, we expected high availability words to be rated as easier to associate with sentences than low availability words. As sentence availability influences lexical processing (see Chapter 3), we also predicted that high availability words would be read and rated more quickly than low availability words.

To assess the influence of recent experience, we tested whether participating in the first task (be it sentence reading or availability rating, manipulated between-participants) influenced performance when the words were encountered again in the other task, later in the experiment. We refer to the difference between processing in task 1 and task 2 as priming. Comparing performance between participants allowed us to compare unprimed vs primed behaviour for the same task and same items. We predicted that primed words would be read faster, rated faster, and rated as easier to associate with sentences than unprimed words.

Our study also gave opportunity to add to research into the differences between production and comprehension. Hopman and MacDonald (2018) argued that production provides more opportunity for learning than comprehension. In their study of artificial language learning, participants who had learnt via production learned better than those who learnt via comprehension. This could be because production is an active practice whereas comprehension is more passive. Conversely, Bock et al. (2007) found that presenting syntactic structures both via production and via comprehension lead to post-priming production benefits, and the magnitude of these priming effects were similar, suggesting that production is not necessarily better than comprehension. Both of these studies used auditory stimuli. To our knowledge, the priming effects of production and comprehension have not been directly compared in visual word processing. McKone and Dennis (2000) discussed the differences between visual and auditory word priming. In their study, priming

effects for auditory stimuli were longer-lasting than for visual stimuli. The researchers argued that differences in short-term priming between visual and auditory stimuli arise due to the relative signal-to-identity ambiguity of speech vs writing. In speech comprehension, there is more room for perceptual error, as single auditory sounds are more ambiguous than single written words. As a result, temporary retention of information in working memory, which can later be integrated with more context as it becomes available, is more crucial for auditory stimuli than for visual stimuli. Research is therefore needed to determine whether differences between production and comprehension are observed during visual word processing. According to MacDonald's account, both production and comprehension lead to words being more readily available, meaning that words primed in either production or comprehension should be facilitated in lexical processing. In line with Hopman and MacDonald's (2018) study, the facilitative effects of priming via production might be greater than those of priming via comprehension.

Our sentence production task tapped mental verbalisation rather than verbal output. Mental verbalisation may better reflect non-edited language production: though there is evidence that we self-monitor our speech internally before articulation (Özdemir et al., 2007), participants might be less likely to alter the sentences they produce if they are not asked to speak them aloud. By testing whether encountering words in a mental production task produces a priming effect on subsequent reading behaviour, this study will allow us to investigate production in the context of visual language processing.

We also explored interactions between long-term experience (as tapped by the sentence availability ratings of the words, taken from Chapter 3) and recent experience (as induced by the priming phase of the experiment). Here, our predictions were not as clear. Words that are difficult to place in context might have more opportunity for change, and therefore could be more susceptible to priming effects in a similar way to the greater priming benefits seen for low frequency words. On

the other hand, high availability words might benefit more from priming as they are likely to co-occur with a larger range of contexts and are better connected in the lexicon than low availability words. This study sought to determine the relationship between the effects of lexical variability (long-term experience) and priming (recent experience).

METHOD

PARTICIPANTS

Two hundred adult native speakers of British English took part in the study. Participants were recruited via Prolific (www.prolific.co, Palan & Schitter, 2018) and paid £10/hour.

DESIGN

All participants took part in a production task (sentence availability rating) and a comprehension task (self-paced reading). Participants were randomly assigned to a group, determining task order. Half read sentences first ('reading-prime'), and the other half rated words first ('rating-prime', see Figure 4.1). In both conditions, task 1 served to prime performance in task 2.

Within each group, participants received one of two wordlists. Participants encountered the same target words for both tasks. All participants also took part in a five-minute digit recall task between the two experimental tasks, which served as a distractor.

Sentence availability (SA) was compared within participants, with each participant reading and rating 10 high and 10 low availability words. Priming effects were compared between participants, as the order of tasks 1 and 2 were manipulated between the reading-prime and rating-prime groups. The experiment was designed in Gorilla (www.gorilla.sc, Anwyl-Irvine et al., 2018) and

administered via Prolific. This study was approved by the University of Oxford Ethics Committee R73573/RE001.

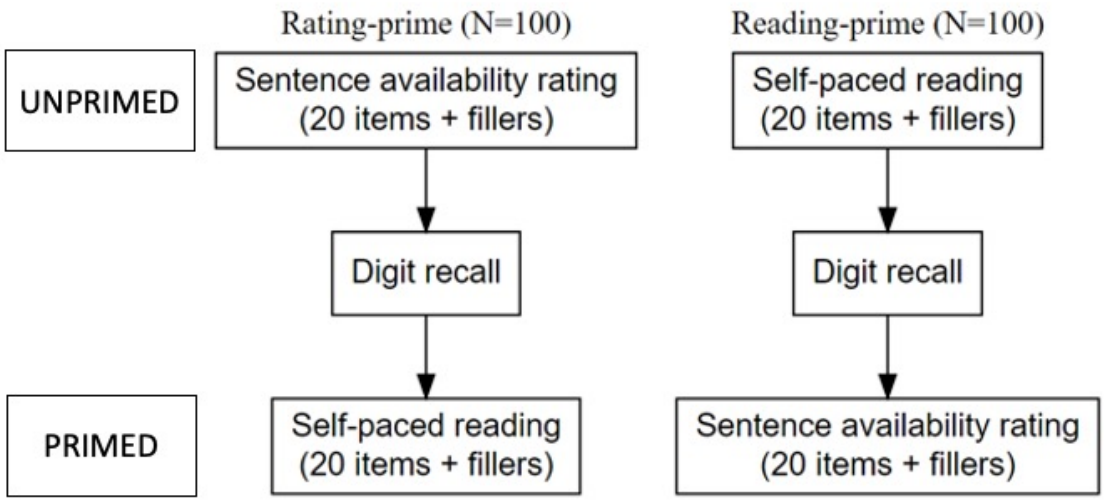


Figure 4.1: Flowchart showing between subjects' priming conditions.

Materials and Procedure

Stimuli consisted of 40 target words (20 high SA and 20 low SA), with availability scores being taken from the norms provided in Chapter 3 (and see Taylor et al., 2022). These varied between 1 (difficult to associate with a sentence) and 7 (easy to associate with a sentence). Using LexOPS (Taylor et al., 2020), the wordlists were designed to be matched for frequency and imageability, but to differ in sentence availability. To avoid confounds of ambiguity, none of the words were homophones. Descriptive statistics are shown in Table 4.1.

Table 4.1: Means and standard deviations for high and low availability wordlists.

	Sentence Availability		Frequency		Imageability	
	Mean	SD	Mean	SD	Mean	SD
High SA (N = 20)	6.390	0.226	4.105	0.544	5.028	1.227
Low SA (N = 20)	4.747	0.227	3.559	0.309	4.819	1.319

Note: Frequency was taken from the SUBTLEX-UK corpus (Van Heuven et al., 2014), and imageability from Cortese & Fugett (2004).

Sentence reading (task 1 or task 2, depending on condition)

Participants read pairs of sentences in a self-paced reading paradigm. Sentences were split into segments of either single words or two-word phrases which appeared on screen one at a time. The first sentence in each set allowed a context to be introduced which would make the following sentence plausible. The second sentence contained the target word. Here, the first four words/phrases were matched across the high and low availability conditions, such that every low availability word shared the same sentence opening with that of one high availability word. The spill-over word was also matched. Target words were always placed in the fifth word/phrase of the second sentence, with the spill-over word in the sixth position. Both target and spill-over words were always presented in isolation. Examples are shown in Table 4.2.

Table 4.2: Example sentence pairs for AWE (upper section, high availability) and YIELD (lower section, low availability).

					AWE					
S1	The	comet	shot	through	the	sky.				
S2	It	was	a	great	**awe**	*for*	all	who	witnessed	it.
					YIELD					
S1	The	carrots	filled	several	bags.					
S2	It	was	a	great	**yield**	*for*	the	farmers	that	harvest.

Note: The target word is shown in bold, and the spill-over word in italics.

Wordlists were divided so that each participant would only encounter one sentence with the same opening. For example, half of participants would see 'awe' but not 'yield', and vice versa. Each word or phrase appeared centred on screen until participants pressed any key on the keyboard, allowing them to move onto the next word/phrase. Response times were measured for the target word and separately for the spill-over word, from onset of the word's presentation to when participants pressed the keyboard. The length of sentence 2 did not differ significantly between the high and low sentence availability conditions, $t(37.83) = 0.80$, $p = .428$, and neither did the length of the prior sentence, $t(35.82) = -0.33$, $p = .743$. Materials were piloted to ensure that they were plausible to participants. A separate experiment tested predictability. Participants were presented with Sentence 1 and the first 4 words of Sentence 2. They were asked to type which word they expected to see next. Participants' responses were then checked against the target words to give a predictability score for each item. No difference in predictability score between sentence availability conditions was found (see Appendix 7 for full details).

To ensure that participants were attending and reading for meaning, we included 10 pairs of sentences that were similar in style and structure to the experimental sentences, but were followed by a forced-choice comprehension question. Participants who did not respond correctly to 80% of

trials were excluded from analysis. Experimental sentences were not followed by comprehension questions so as not to confound comprehension question with learning opportunity, but from the participants' point of view, the question probes were not predictable.

We also included filler words in separate sentence pairs to test whether priming effects were confounded by order. Here, our aim was to ensure that potential differences between primed and unprimed performance could not be attributable to order, as 'primed' performance was always measured during the last task that participants completed (see Appendix 8 for summary). Sentence pairs were of similar length and style to those containing target words, and thus did not appear different to experimental sentences from the participants' perspective.

Participants were given two practice trials before the task began. Both trials were followed by a comprehension question. Test trials, filler trials, and trials with comprehension questions were then randomised. This task took 10 minutes to complete.

Filler task: digit recall

This served as both a non-linguistic distractor task, and as a means of checking attention. Digits were presented one-by-one on screen in a string of three to six digits. Each digit appeared on screen for 500ms. After each string, participants were asked to type the string into a text box. There were 15 strings, taking five minutes to complete in total. At the end of the task, participants were told their score as a percentage. Participants who scored less than 50% were excluded from analysis.

Sentence availability rating (task 1 or task 2, depending on condition)

Participants were presented with a single word at a time and asked to rate how easily it could be associated with a sentence on a Likert scale of 1 (easy to think of a sentence) to 7 (difficult to think of a sentence). This task is the same as the sentence availability rating task reported in Chapter 3, with one difference: the instructions in the current study emphasised that participants should only make their rating once they had thought of a sentence using the target word. Response time per trial was collected as a measure of how long participants took to think of a sentence and make their rating.

Two examples were provided in the instructions, followed by three practice trials. Participants responded to 20 target words (10 high SA, 10 low SA) and 10 filler words, presented in random order. As with the reading task, filler words (different to those seen in sentence reading) were included to test whether potential priming effects were confounded by task order (see Appendix 8). This task took 10 minutes to complete.

RESULTS

Five participants were excluded for low performance on comprehension questions (<80% accuracy) and digit recall (<50% accuracy). We excluded reading times above 2,500ms per word, and rating times above 8,000ms per trial. Outliers more than 2.5 standard deviations from the participant mean for each task were then removed. These procedures reduced the data by 11.33%.

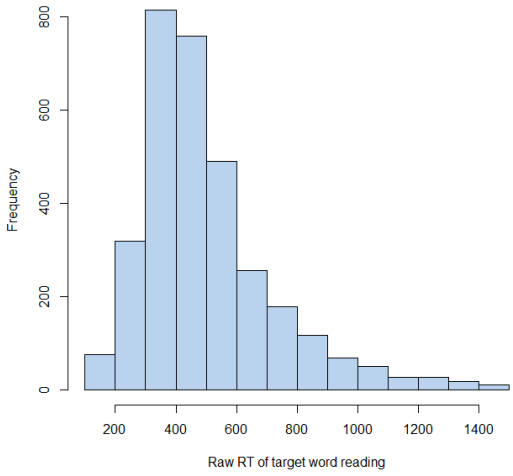
Reading times and rating times were log transformed. Likert rating scores were reverse coded so that a higher score would reflect higher sentence availability. After transformations, all data were normally distributed. Likert scores were slightly skewed, as participants tended to rate most words as easy to associate with a sentence (see Table 4.3). This is similar to the set of 3,000 sentence availability norms (Taylor, et al., 2022), in which the average rating was 5.94. Participants'

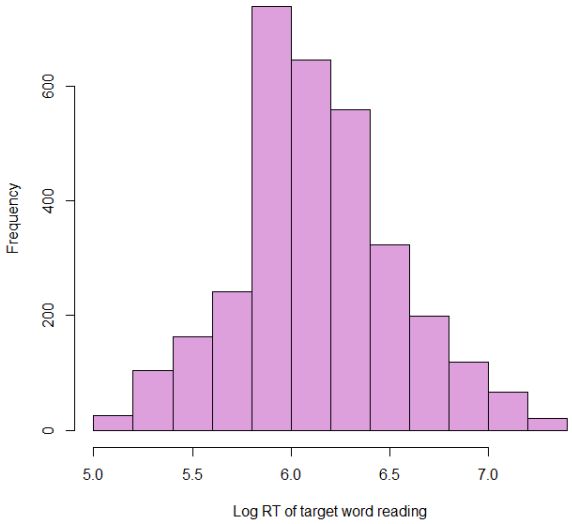
ratings in the current study were rated significantly lower than the same words when rated in Chapter 3, $t(4280.5) = 6.402$, $p < .001$. This could be due to differences in range of items presented between the studies. Distributions can be seen in Figure 4.2.

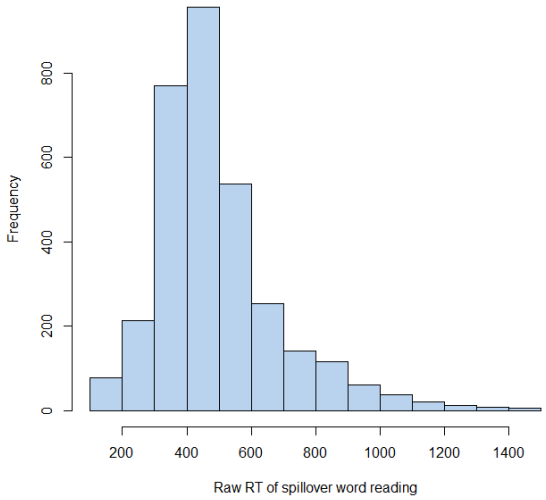
Table 4.3: Mean average, skewness and kurtosis of reading time (upper section; ms) and rating response time and score (lower section) collapsed across condition (N = 195 participants).

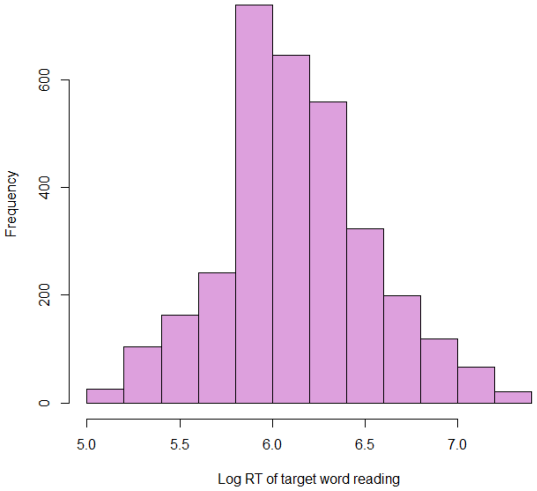
Reading measure	Mean	Skewness	Kurtosis
Raw RT (target)	500.71	1.45	2.54
Log RT (target)	6.13	0.20	0.11
Raw RT (spill-over)	493.78	1.47	3.08
Log RT (spill-over)	6.13	0.11	0.55

Rating measure	Mean	Skewness	Kurtosis
Raw RT	3211.39	0.87	0.10
Log RT	7.94	-0.11	-0.65
Likert score	5.41	-1.02	0.07


Frequency
0
200
400
600
800
200
400
600
800
1000
1200
1400
Raw RT of target word reading


Frequency
0
200
400
600
5.0
5.5
6.0
6.5
7.0
Log RT of target word reading


Frequency
0
200
400
600
800
200
400
600
800
1000
1200
1400
Raw RT of spillover word reading


Frequency
0
200
400
600
5.0
5.5
6.0
6.5
7.0
Log RT of target word reading

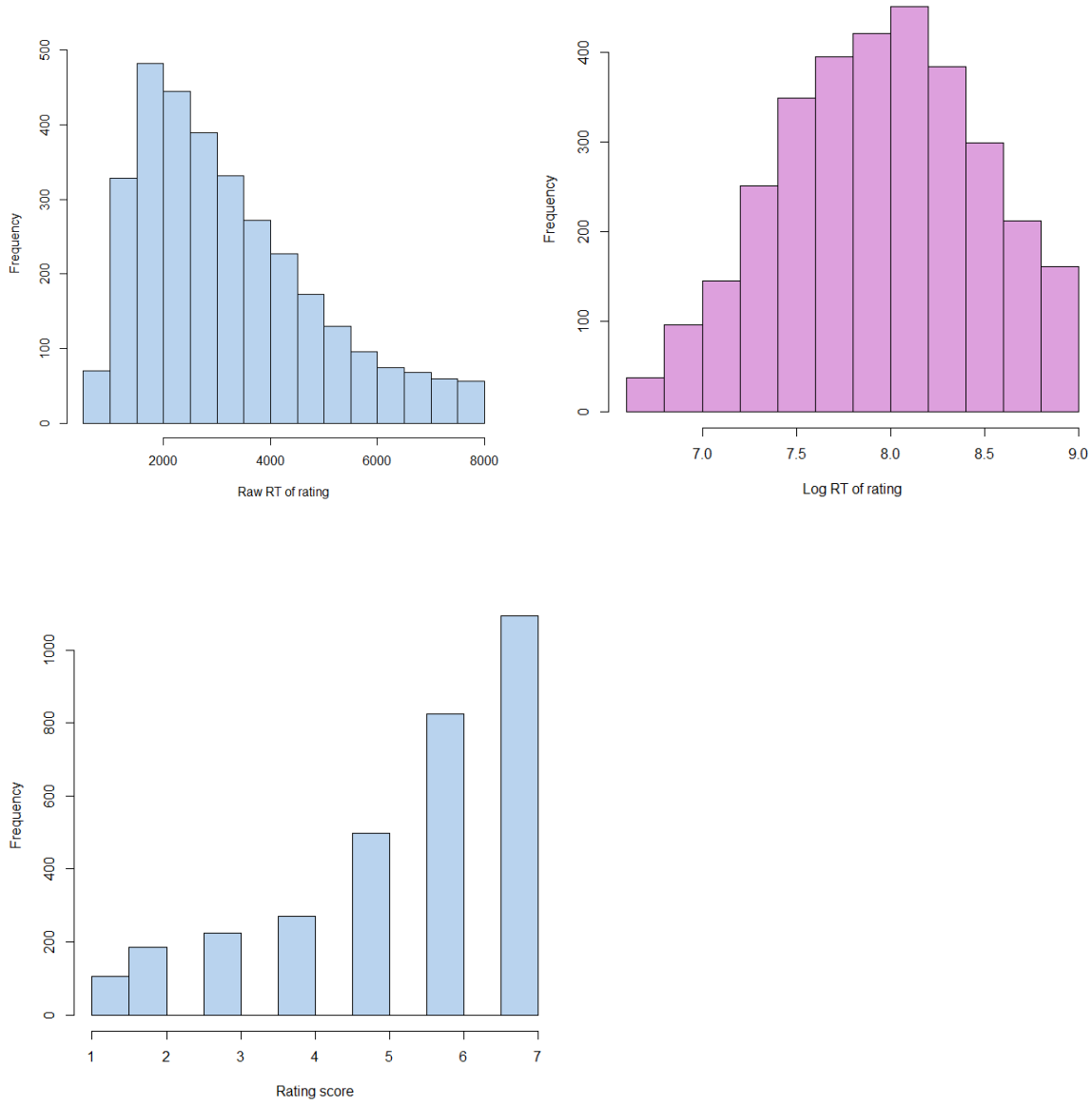


Figure 4.2: Histograms for reading time of target word and spill-over, Likert rating response time, and Likert score.

Note: Raw data is shown on the left and log transformed data on the right. Likert score was not log transformed.

Linear mixed effect models were performed to investigate the effects of sentence availability within groups and priming between groups. Log-transformed response time data were analysed using 'lmer' from the package 'lme4' (Bates, Sarkar, Bates, & Matrix, 2007), and rating scores using 'clmm' from the package 'ordinal' (Christensen, 2019). Random effects for item and participant were included, with the random effect explaining the least amount of variance removed if convergence issues arose. Details of the random effects removed and the final models tested can be seen in Appendix 9. Contrast coding was used for high SA (0.5) and low SA (-0.5), and for participants who rated words first (0.5, reflecting 'primed' performance in reading and 'unprimed' performance in rating) and read words first (-0.5, reflecting 'unprimed' performance in rating and 'primed' performance in reading).

Reading performance

First, we investigated reading times of target and spill-over words, comparing performance for those participants who read the sentences at the start of the experiment (unprimed) and those who read them at the end of the experiment, after first seeing the same words in the rating task (primed). Slopes were included by-item for priming order and by-participant for priming order and sentence availability in the following model:

Reading time ~ priming order * sentence availability + (1+priming order | item) + (1 + priming order + sentence availability | participant)

This model failed to converge. Slopes for priming order were removed by-item and then by-participant. Both models identified a singular fit. By-participant slopes were then removed for sentence availability. This gave the following model, for which we report results:

Reading time ~ priming order * sentence availability + (1 | item) + (1 | participant)

Descriptive data are shown in Figure 4.3 and Table 4.4. There was no main effect of priming or sentence availability (see Table 4.5 and Figure 4.3). For target word reading, there was a significant interaction between prime condition and SA.

Table 4.4: Raw (upper section) and log transformed (lower section) mean and standard deviations of reading times for target words, split by condition.

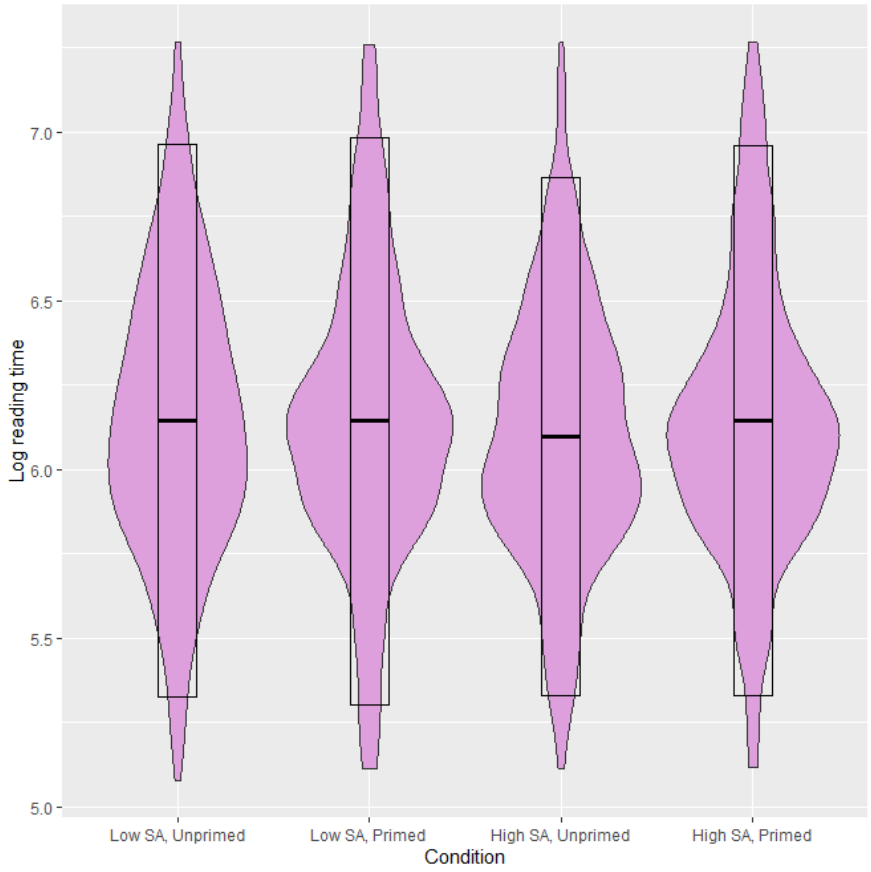
	Raw RT	
	Mean	SD
Low SA, Unprimed	507	219
Low SA, Primed	509	232
High SA, Unprimed	479	197
High SA, Primed	508	231

	Log RT	
	Mean	SD
Low SA, Unprimed	6.14	0.41
Low SA, Primed	6.14	0.42
High SA, Unprimed	6.09	0.38
High SA, Primed	6.14	0.41

Table 4.5: Fixed effects for predicting target word reading time (upper section) and spill-over word reading time (lower section).

	Estimate	Std Error	t	p
Intercept	6.17	0.03	181.68	<.001
Priming	-0.04	0.05	-0.87	.39
Sentence availability	-0.00	0.01	0.13	.90
Interaction	0.05	0.02	2.41	.02

	Estimate	Std Error	t	p
Intercept	6.17	0.03	199.41	<.001
Priming	-0.03	0.04	-0.80	.43
Sentence availability	0.00	0.01	-0.35	.73
Interaction	0.01	0.02	0.67	.50


7.0
6.5
6.0
5.5
5.0
Log reading time
Low SA, Unprimed
Low SA, Primed
High SA, Unprimed
High SA, Primed
Condition

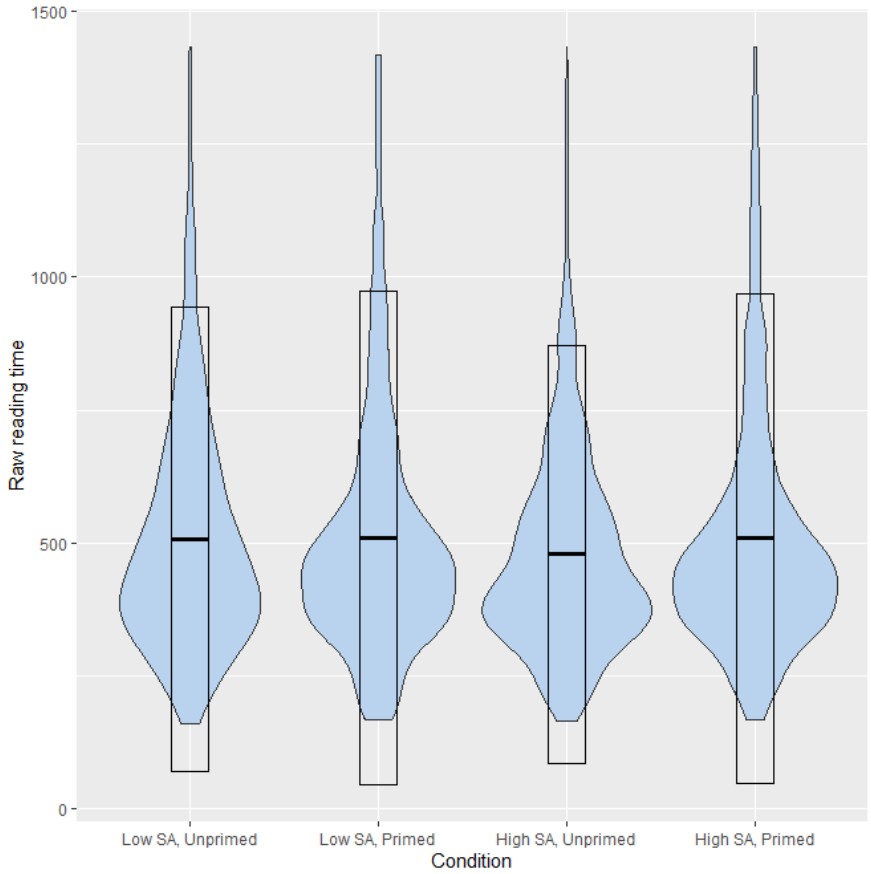

1500
1000
500
0
Raw reading time
Low SA, Unprimed
Low SA, Primed
High SA, Unprimed
High SA, Primed
Condition

Figure 4.3: Violin plot of log (top) and raw (bottom) reading time for target word reading, split by sentence availability condition and priming condition (with mean and SD).

Rating performance

Next, we investigated performance on the availability rating task. As for reading performance, we first included slopes by-item for priming order and by-participant for priming order and sentence availability in the following model:

Rating ~ priming order * sentence availability + (1 + priming order | item) + (1 + priming order + sentence availability | participant)

This model failed to converge. Slopes for priming order were removed by-item and then by participant. This gives the following model for rating time:

Rating time ~ priming order * sentence availability + (1 | item) + (1 + sentence availability | participant)

For rating score, the model had a singular fit, and we therefore removed by-participant slopes for priming order and then for sentence availability. This gave the following model:

Rating score ~ priming order * sentence availability + (1 | item) + (1 | participant)

Descriptive statistics are shown in Tables 4.6 and 4.7. As predicted, high SA words were rated as easier to associate with a sentence than low SA words. High SA words were also responded to faster than low SA words. Priming condition (whether the rating task was completed first, or after the same words were encountered in sentence reading) did not explain unique variance. No interactions were found between priming and sentence availability (see Table 4.8). Exploratory

analyses were also performed to investigate how the ratings given to the items in this experiment compared with those previously collected for the same items in the norming study (Chapter 3). These analyses are reported in Appendix 10. The correlation was significant, but lower than expected (r = .35).

Table 4.6: Raw (upper section) and log transformed (lower section) mean and standard deviations of response times for rating, split by condition.

	Raw RT	
	Mean	SD
Low SA, Unprimed	3301	1652
Low SA, Primed	3471	1685
High SA, Unprimed	2903	1594
High SA, Primed	3163	1616

	Log RT	
	Mean	SD
Low SA, Unprimed	7.98	0.51
Low SA, Primed	8.03	0.51
High SA, Unprimed	7.83	0.53
High SA, Primed	7.93	0.52

Table 4.7: Rating score, split by condition.

	Rating	
	Mean	SD
Low SA, Unprimed	4.93	1.87
Low SA, Primed	4.78	1.87
High SA, Unprimed	6.04	1.20
High SA, Primed	5.88	1.34

Table 4.8: Fixed effects for predicting rating response time (upper section) and Likert score (lower section).

	Estimate	Std Error	t	p
Intercept	7.87	0.04	175.50	<.001
Priming	0.08	0.06	1.13	.18
Sentence availability	0.15	0.02	7.61	<.001
Interaction	-0.04	0.03	-1.52	.13

	Estimate	Std Error	z	p
Priming	-0.27	0.21	-1.28	.20
Sentence availability	-1.47	0.10	-14.76	<.001
Interaction	-0.10	0.13	0.75	.46

DISCUSSION

Our experiment tested the effects of long-term and recent lexical experience on two different processing tasks. Long-term experience was indexed by prior sentence availability rating and was investigated by comparing task performance for high and low availability words. Recent experience was investigated by comparing performance between participants when words appeared unprimed (task 1) vs primed (task 2). We hypothesised that primed processing would be faster than non-primed, and that high availability words would be processed faster than low availability words.

Words in the high SA condition were processed faster in both the rating task, and they were rated as easier to associate with sentences than low SA words; however SA did not appear to affect word reading. These results validate the sentence availability measure, but the application of this measure to word processing in a reading task appears limited. This calls for more research into SA. A large literature demonstrates that long-term lexical experience (as captured by a lexical variable) influences how words are processed. On this view, we predicted that sentence availability would be

associated with more efficient reading. SA can be seen as a reflection of participants' prior language experience; but our results demonstrate that words rated as having more readily accessible linguistic contexts do not receive a speed benefit during sentence reading.

There was no main effect of priming for either task. That is, encountering a word in sentence reading did not change how people rated the same words subsequently, and rating a word did not influence subsequent word reading. Priming effects have been found in other cross-modal tasks (e.g., spoken sentences prime picture naming, Felser & Roberts, 2007; auditory words prime visual word recognition, Allen & Badecker, 2002). It could be that, in our design, a single encounter with a word in the context of self-based reading or rating was not sufficient to alter subsequent lexical behaviour. This could be due to the time course of our study; the time between tasks 1 and 2 might have been too long for a temporary change in performance to be maintained.

Turning to the interactions, we found that high SA words were read faster than low SA words if they had not been primed. The direction of this effect is the opposite to our predictions, as we predicted that priming would facilitate performance. It could be argued from our results thta previous exposure inhibits processing of subsequent encounters, but only for words which are already available. While inhibition was not predicted, it is not alone in the literature (e.g., Oltra-Massuet et al., 2017; Waszak et al., 2005; Mayr & Buchner, 2007). It could be that encountering words across the two different types of task induced cognitive demands that slowed processing. However, it is unclear from these results whether we can interpret the faster reading of unprimed high SA words as inhibition. The other conditions (primed high SA words; and both primed and unprimed low SA words) all had the same log mean reading time (see Table 4.4). This suggests that unprimed high SA words were not only faster than primed high SA words, but also from low SA words read in both conditions. This speeded reading for unprimed available words is therefore difficult to interpret.

A possible explanation rests on the assumption that high SA words would typically be processed faster than low SA words, in line with our theoretical account of availability. Placing a high SA word in a sentence (as required by the rating task) might induce a context that then interferes with reading a word in a different context. On this view, high SA words might be associated with a wider range of possible linguistic contexts. This idea is supported by the results reported in Chapter 3, where sentence availability correlated positively with semantic diversity. In the current experiment, the varied linguistic contexts available during task 1 might therefore lead to a delay during task 2, as participants might encounter the primed word in a context different to the one previously produced. Schuberth and Eimas (1977) reported that participants responded faster to words when they were preceded by a congruous context and slower when preceded by an incongruous context. Bainbridge et al. (1993) found that repetition priming effects were eliminated when polysemous words were presented in contexts that vary the word's perceived sense compared to contexts where the word sense is uniform. We controlled for ambiguity in that our analysis avoided homophones. However, most words are polysemous. Particularly for items high in sentence availability (which correlates positively with measures of polysemy like semantic diversity and number of senses, see Chapter 3), a wide range of sentence contexts might be produced. If the produced context is incongruous with the subsequently read context, this could cause slowed reading, especially for those words with high SA.

While we present this as a possible explanation of our results, our evidence is not sufficient to determine whether this is the case. The idea of polysemy explaining our interaction rests on the assumption that high SA words are typically read faster than low SA words; and this has not been established. As the significant interaction between sentence availability and priming is small and difficult to understand, it could be a spurious effect. A pre-registered replication is needed.

Polysemy might have also played a role in our other conditions, as low availability words still might be susceptible to such context effects. To give an example from our stimuli, 'ranch' scores low for polysemy, but might still be perceived in a different sense (e.g., a farm vs a salad dressing). A future study with a tighter control for polysemy might be useful in teasing apart these effects. For words primed via sentence reading, this explanation is less likely, as we would expect this given context to be made readily available and therefore made more accessible during sentence production. As we did not collect the sentences participants produced, it is unclear whether these were congruous with the context given during sentence reading.

In our experiment, participants switched between comprehension and production tasks. According to MacDonald (2013), producing words makes them more accessible and boosts subsequent comprehension. In turn, comprehension makes words more accessible, meaning that we are more likely to produce them. MacDonald's account would therefore predict more efficient processing in the second task, once words have already been read or rated. This is not what we found. Note that our study differs to previous studies investigating production. In the Production-Distribution-Comprehension account of language, production often reflects the likelihood of an individual producing a certain word, phrase or syntactic structure that may have been encountered in previous comprehension. In our experiment, however, participants were shown the target word. We instructed participants to think of sentences from target words that were given on screen, and then make a rating. Therefore, while other production measures reflect likelihood of word use, the current measure gave a proxy for the ease with which sentences could be produced. Mental production might also be different to verbal production. While the former requires self-monitoring and message planning, verbal production requires more elaborate planning of words and structure. Further research is needed to establish the effects of different types of production on comprehension, and vice versa.

It is also important to note that, while our task instructed participants to think of a sentence before making their rating, we did not check whether they mentally produced a sentence. It is possible that participants did not follow instructions exactly; or behaviour might have changed throughout the course of the rating task. Though rating response time did appear to capture the ease with which words were associated with a sentence, this reflection does not necessarily involve placing a word in a specific linguistic context. There might be differences between these two types of processing: sentence availability rating involves thoughtful reflection on a word but might not force participants to engage actively with its linguistic context. Therefore, while we framed our task as a measure of production, this framing should be interpreted with caution. To ensure participants engage in production, a future study could measure the sentence produced. For example, participants could be asked to speak this sentence aloud while their audio is recorded.

Butler and Berry (2004) discussed differences between repetition priming paradigms (in which participants typically take part in an active judgement task, e.g., lexical decision, word naming) and exposure tasks (in which participants make an affective judgement, e.g., rating, preference judgement). In discussing similarities and differences between the research outcomes of these two paradigms, the researchers questioned the status of exposure as a form of priming. Given this distinction, we might question where our availability task sits within these paradigms. Availability ratings are not affective as such, but might involve a similar type of judgement, as participants reflect on the ease with which they can place the stimuli. Butler and Berry went on to argue that there might be multiple mechanisms contributing to exposure effects; a specific component relating to perceptual fluency, and a diffuse component relating to generic positive affect. More research is needed to determine the nature of sentence availability ratings, and whether thinking about a sentence context influences processing in similar ways to making judgements about affect, as set out in Butler and Berry's framework.

In conclusion, in finding an influence of sentence availability on rating times, our results validate the metric devised in Chapter 3. We did not find any effects of priming: reading or rating the words in the first task did not influence the reading or rating of the same words in the second task. There was some evidence to suggest that, for high availability words, rating the words in the first task inhibited word reading in the second task. This might be due to the relationship between availability and polysemy, as words high in availability may be placed in a wider range of contexts. Further research is needed to investigate the relationship between polysemy and sentence availability, and the relationship between comprehension and production in the context of written language processing.

CHAPTER 5
TESTING THE EFFECTS OF RECENT EXPERIENCE ON SUBSEQUENT WORD PROCESSING THROUGH LEXICAL AND SEMANTIC DECISION TASKS

INTRODUCTION

The experiment reported in Chapter 4 attempted to test the effects of long-term and recent lexical experience on lexical processing. In this chapter, we test this general idea again but with a different design. Here, items varied in their existing semantic diversity and imageability values, giving proxies of long-term experience. Lexical and semantic decision tasks were used to prime words, and the effects of this on subsequent performance was measured through lexical decision. This allowed us to test the effects of long-term and recent experience on lexical processing using more robust design.

Semantic diversity (a corpus derived measure) explains unique variance in lexical processing (see Chapter 2). Hoffman and Woollams (2015) explored the effects of diversity on two tasks: lexical decision and semantic decision. During lexical decision, participants were presented with either a word or non-word and were asked to decide whether the letter string is a word or not. Words high in semantic diversity were processed faster than words low in diversity. The semantic decision task involved making relatedness judgements. Here, target words were presented alongside a word that was either semantically related (e.g. arc – circle) or semantically unrelated (e.g. arc – paper). Participants were asked to decide whether or not the two words on screen were related. Hoffman and Woollams reported that, during semantic relatedness decisions, higher diversity was associated with slower processing. This suggests that making decisions about word meaning is harder for words which tend to appear in diverse contexts. The task-dependant effects of semantic diversity follow a wider trend in the literature, in which higher levels of word ambiguity are associated with better

lexical processing (in lexical decision) but poorer semantic processing (e.g., Yap et al., 2011). This pattern is also seen when semantic diversity is induced experimentally. Johns et al. (2016a) presented participants with pseudowords in passages that were either uniform in theme (akin to low diversity) or had different themes (akin to high diversity). High diversity words were responded to faster in speeded recognition but were associated with slower processing during semantic judgement.

These task-dependant effects of semantic diversity likely reflect differences in the knowledge required to perform each task. In lexical decision, a participant reads a visual stimulus and decides on its lexicality. The emphasis of this task, therefore, is on the written form of the word. Words with frequent orthographic neighbours are associated with slowed lexical decision times (Grainger, 1990). Irregularly spelt words are responded to more slowly than words with regular spellings (Parkin, 1982), particularly when irregular words are difficult to pronounce. McClelland and Rumelhart's (1981) interaction-activation model has described the perceptual processing in visual word recognition. Here, recognition is the result of excitatory and inhibitory interactions between visual features. This model suggests that visually similar words compete for identification, with pronounceable pseudowords also activating visual feature detectors. During lexical decision, therefore, recognising the visual features of a word is key to accurately completing the task. Balota (1990) proposed that a similar process could account for processing advantages seen for semantically rich words. These words might be facilitated in lexical decision as they generate more activity at the semantic level, thus giving semantically rich words more opportunity for feedback (for review, see Yap et al., 2015).

Words with multiple meanings (Rubenstein et al., 1970) and fewer derivations of meanings (i.e., multiple related meanings or senses, Jastrzembski, 1981) are processed faster in lexical decision. Pexman et al. (2002) discussed the relationship between orthographic and semantic units

of word knowledge during word processing. Both words with numerous senses and numerous features are associated with a richer wealth of semantic information, and so activation from semantics to orthography will be stronger for these semantically rich words. Taken together, these findings suggest that form-based processing is most efficient when nodes are semantically connected. This idea is supported in the literature. Words with higher associative strength are associated with better performance in recall tasks (Steyvers et al.) and are learnt earlier in life (Hills et al., 2009). This semantic connectiveness might therefore explain the effects of semantic diversity on lexical decision.

Unlike lexical decision, semantic relatedness judgements rely on processing a word's meaning. Semantic decision is typically slower if the target word is ambiguous (Yap et al., 2011). This effect can be changed by disambiguating information, which can help the participant know which meaning the word should be processed in. For example, Gottlob et al. (1999) presented homographs (words with multiple meanings that have different pronunciations, e.g., *lead*), homonyms (words with multiple meanings that share a pronunciation, e.g., *spring*) or unambiguous words to participants. Each target word was followed by a related or unrelated word. Participants were slower to make judgements for homographs and homonyms when the paired word was presented after the target word. When paired words were presented beforehand, however, the inhibitive effect was not observed for homonyms. The researchers argued that this was because the paired word gave a context that informed which meaning of the ambiguous word should be processed. These results demonstrate that, while semantic decisions are slower for ambiguous words, this effect is dependent on context. The ambiguity account of slowed semantic processing can explain the results of semantic diversity as diversity can be viewed as a continuous measure of polysemy which captures nuanced differences in word meaning (Hoffman et al., 2013). Hoffman and Woollams' (2015) experiment therefore demonstrated that the ambiguity disadvantage for semantic processing holds when ambiguity is measured continuously rather than categorically.

Hoffman and Woollams' (2015) also found that lexical decision and semantic decision were faster for words high in imageability. This demonstrates that both form-based and meaning-based processing is more efficient for words that can be easily associated with an image and shows how different semantic variables can have different effects across types of task.

Hoffman and Woollams' (2015) findings are generally consistent, having been replicated by other studies. For example, Hsiao et al. (2020) replicated the effects of Hoffman and Woollams in both children and adults: semantic diversity was associated with faster lexical decisions. Diversity was also associated with slower semantic judgements, but only for trials where words were semantically unrelated. Imageability was associated with faster performance in both tasks. Though Hsiao et al. initially failed to find a facilitatory effect of diversity on lexical decision in adults, this effect was robust when tested using the same items as Hoffman and Woollams, when analysing data through ANOVA rather than linear mixed effect models, and when using largescale secondary data.

We employed lexical decision and semantic decision tasks to test the effects of recent experience on subsequent processing. In Chapter 4, there was no main effect of priming for either of the tasks used in that experiment, suggesting that recent encounters with words do not change subsequent processing. This contradicted predictions made in line with MacDonald's (2013) Production-Distribution-Comprehension account of language processing, in which words that are recently accessed are more readily available, and thus easier to produce or comprehend. One reason for the null results could be the experimental design (see Chapter 4 for discussion). In the current study, we once again tested the role of recent experience on subsequent processing but used lexical variables and tasks for which there is a more clear relationship in the literature. Basing our design around the materials and tasks used by Hoffman and Woollams (2015), we tested whether making a decision about a word would affect subsequent lexical processing. As in their experiments, words varied in semantic diversity and imageability. In the first phase of the experiment (serving as

'priming'), participants either took part in lexical decision or semantic relatedness decision. Then, after a short filler task, all participants responded to the same words in a lexical decision task. This allowed us to test whether encountering a word in one task would facilitate subsequent processing of the same word. In line with MacDonald's account, we predicted that post-test lexical decision performance would be significantly faster than pre-test lexical decision performance.

There also might be differences between the types of processing during pre-test. Reflecting on a word's meaning is more cognitively demanding than reflecting on a word's form and involves a deeper level of processing. As demonstrated by semantic richness effects, words associated with more semantic information are advantaged in lexical processing (Pexman et al., 2002). While processing advantages have been explained through theories of competition for semantic representation (Morton, 1979), they could also reflect a more general facilitatory process by which activating semantic information eases lexical retrieval (Balota, 1990). Considering semantic information might therefore give more opportunity for words to be prioritised in the lexicon, thus making them more readily available and easier to process (MacDonald, 2013). The importance of semantic information on subsequent lexical processing has also been investigated through ambiguity priming experiments, where the context in which an ambiguous word is primed impacts subsequent word association (Betts et al., 2018). While this type of priming is different to the semantic decision task of the current study, these results demonstrated the lingering influence of semantic experience on later performance. If meaning-based decisions involve deeper level processing, we would expect semantic decision to have a greater facilitative effect on subsequent processing than lexical decision.

However, an argument could be made for the opposite prediction. As discussed in Chapter 4, a wealth of literature has found that repetition facilitates subsequent production. This effect might differ depending on the degree of overlap between pre-test and post-test tasks. In this view, both semantic decision and lexical decision could constitute as repetition, as the same words appear

in both tasks; but while participants in the semantic decision condition repeat words only, participants who are primed via lexical decisions will repeat the whole task. We might therefore expect greater repetition effects for those in the lexical decision condition.

The aim of this experiment was to further investigate the role of recent experience on lexical processing, and how this interacts with the long-term lexical histories as indexed through semantic variables. We tested whether there would be a difference in the magnitude of priming effects between participants who responded to lexical decision vs semantic relatedness decision. This gave an opportunity to test whether form-based processing and meaning-based processing would affect subsequent lexical behaviour in a different way. We also explored the effects of semantic diversity and imageability on primed lexical decision performance. This repeated Hoffman and Woollams' analysis with data split by pre-test condition (i.e., whether participants had previously encountered words in the context of lexical decision vs. semantic decision); therefore allowing us to see how recent experience interacts with item-level variability in semantic properties developed across years of experience.

This experiment also gave an opportunity to replicate the results of Hoffman and Woollams' (2015) study using a between-participants comparison across the pre-test data. Using the same items as Hoffman and Woollams, we predicted that words high in semantic diversity would be associated with faster response times for lexical decision, and slower response times for semantic relatedness decision. We predicted that words high in imageability would be associated with faster response times for both tasks. This study was preregistered on As Predicted (aspredicted.org, see Appendix 11).

METHOD

PARTICIPANTS

Participants (N=100) were adult speakers of British English recruited online via Prolific (www.prolific.co, Palan & Schitter, 2018). Of these, 50 were randomly assigned to the 'lexical decision' pre-test condition, and 50 to the 'semantic decision' pre-test condition. This value was based on Hoffman and Woollams' sample size of 49 participants and was adjusted to allow sufficient power for between-group analyses.

Materials

Target words were taken from Hoffman and Woollams (2015). The same target words were used for both pre-test conditions (lexical decision and semantic relatedness decision), and in lexical decision at post-test. Items varied factorially for semantic diversity and imageability to form four lists (see Figure 5.1). They were matched for frequency, word length, syllable length, bigram frequency, and neighbourhood size.

	High imageability	Low imageability
High diversity	N = 60 I = 548 (57) D = 1.83 (.09)	N = 60 I = 341 (35) D = 1.84 (.08)
Low diversity	N = 60 I = 559 (56) D = 1.40 (.16)	N = 60 I = 343 (34) D = 1.41 (.18)

Figure 5.1: Factorial design of the target wordlist for lexical decision and semantic relatedness decision tasks, showing the N of items in each list, and the M (SD) for imageability (I) and semantic diversity (D).

In the lexical decision tasks, 240 pronounceable non-words (also taken from Hoffman and Woollams) were presented amongst the 240 words. One item was displayed per trial, centred on the screen. Trials were randomised. In the semantic relatedness decision task, each target word was paired with a related and an unrelated word, which were counterbalanced across participants such that one paired word was shown per trial. As described by Hoffman and Woollams, definitions of semantically relatedness "included shared category membership (e.g., stomach – chest), synonymy and antonymy (e.g., slept – woke), semantic association (e.g., thirst – drought) and action-recipient relationships (e.g., roll – dice)." Target words were displayed alongside a related word in 50% of

trials, and alongside an unrelated word in the remaining 50% of trials. This was counterbalanced such that, for each word, 25 participants saw a related pair and 25 saw an unrelated pair. In each task, trials were randomised.

PROCEDURE

This experiment was designed in Gorilla (https://gorilla.sc, Anwyl-Irvine et al., 2020). All participants completed two tasks, one at pre-test and one post-test. For the pre-test, participants were randomly assigned to either the lexical decision (N = 50) or a semantic relatedness decision condition (N = 50). For the post-test, all participants completed a lexical decision task. Between the pre- and the post-test phases, participants completed an Author Recognition Test. This measured print-exposure and provided a distractor between the two experimental tasks.

Pre-test lexical decision (N = 50). Words and non-words were presented individually on screen. In each trial, participants were asked to press a button on their keyboard to indicate whether it was a word (q) or a made-up word (p). The keys were displayed on-screen during practice trials, but not during experimental trials. For each word, response time and accuracy were measured. Words remained on screen until a decision was made. Trials were separated by a 500ms fixation cross.

Pre-test semantic relatedness decision (N = 50). 240 target words were presented in the semantic relatedness task, alongside either a related or unrelated word. Unlike Hoffman and Woollams, these words were presented simultaneously, shown side by side on screen. Participants pressed a key to indicate whether the words were related in meaning (q) or not related in meaning (p). As with lexical

decision, keys were displayed for practice trials only, and trials were separated by a 500ms fixation cross. Trials were randomised, and word-pairs remained on screen until a decision was made.

Author recognition (N = 100). This task was developed by McCarron (2022). 90 names were presented individually on screen. Of these, 60 names belonged to published authors, and 30 did not belong to authors. For each name, participants pressed a key to indicate whether they recognised the author's name (f) or not (j). Each trial automatically ended after 3,000ms if participants did not give a response.

Post-test lexical decision (N = 100). This task followed the same procedure as the pre-test lexical decision task.

RESULTS

We excluded all trials that were less than 100ms. We also excluded participants who took over 2,000ms (lexical decision) or 4,000ms (semantic relatedness decision) to respond to more than 5% of trials (N = 19). This differed from our preregistration plans, as the proposed 2,000ms cut-off was too strict for the semantic decision task, resulting in a large loss of data. This difference likely resulted in our presentation being different to that of Hoffman and Woollams: while they presented one pair after another, we presented words simultaneously (see Chapter 6 for discussion). We then trimmed each participant's data by excluding response times that were more than two standard deviations from the participant mean in that condition. After exclusions, 81.64% of data remained for analysis, with 39 participants in the lexical decision condition and 42 in the semantic decision

condition. Responses to non-words in the lexical decision task were not analysed, and only correct trials were analysed for response time analysis.

THE EFFECTS OF SEMANTIC DIVERSITY AND IMAGEABILITY ON LEXICAL AND SEMANTIC DECISION AT PRE-TEST

First, we tested whether Hoffman and Woollams' (2015) results would replicate. This analysis only concerned pre-test data. Descriptive statistics are shown in Table 5.1. Our preregistration did not specify whether imageability and semantic diversity would be entered as categorical or continuous variables. Analysing data continuously is generally seen as preferable (Altman & Royston, 2006); however, items were taken from Hoffman and Woollams' orthogonal design of high/low diversity and imageability values, therefore lending itself to a categorical design rather than representing the typical variation of diversity an imageability across words. Categorical vs continuous data for these items is plotted for semantic diversity and imageability in Figure 5.2. As our main goal was to replicate Hoffman and Woollams' analyses, we followed their use of categorical variables. Differences between these approaches are considered in the Discussion, and the data was also analysed and reported using continuous values in Appendix 12.

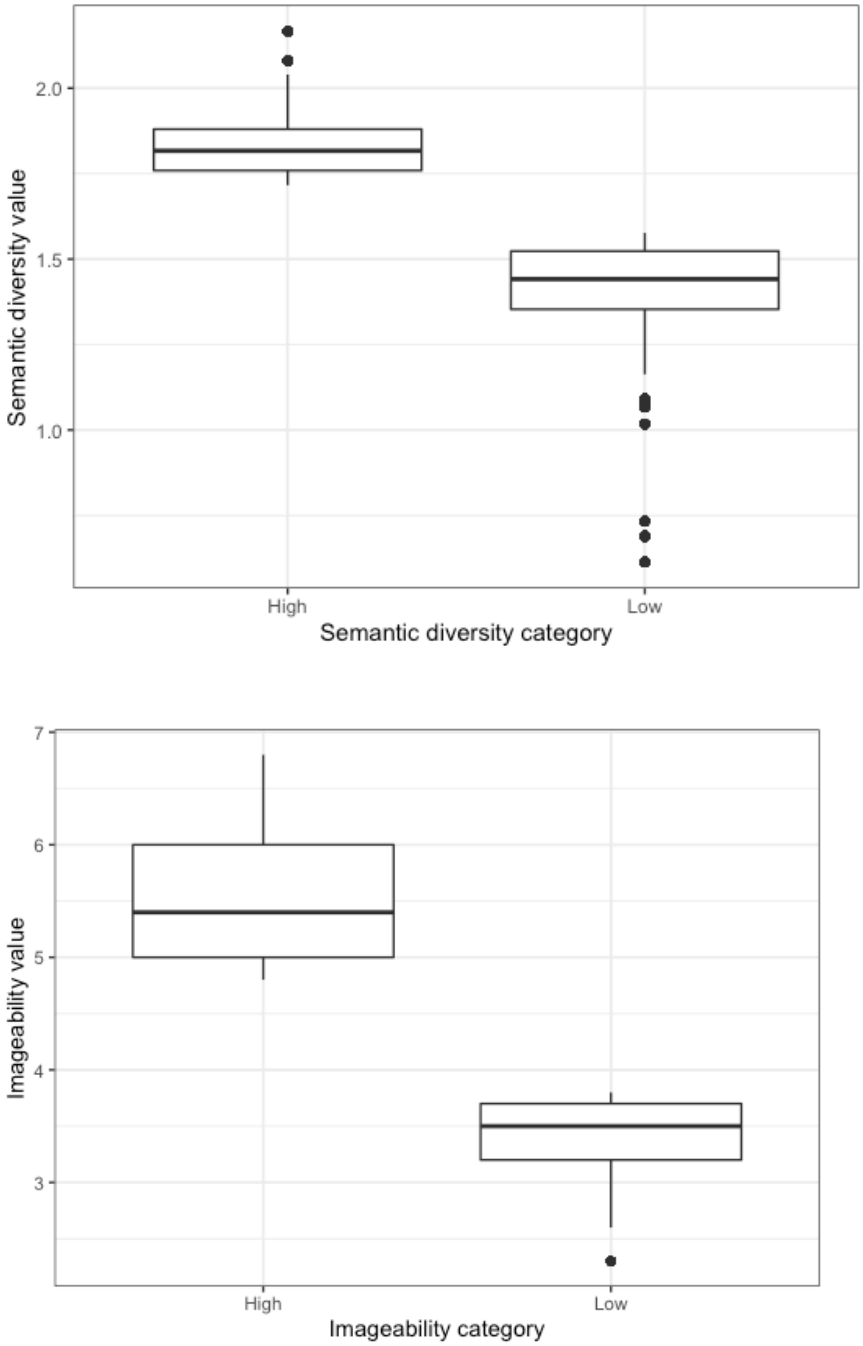


Figure 5.2: Box plots showing the relationship between the high and low categories and the continuous values for semantic diversity (above) and imageability (below).

Next, we checked the distributions of pre-test lexical decision and semantic decision response time data (Figure 5.3). Both conditions were positively skewed (lexical decision, 2.63;

semantic decision, 2.00) and leptokurtic (lexical decision, 18.55; semantic decision, 6.71). To improve the fit, data were log transformed. Mean RTs split by condition can be seen in Figure 5.4.

Table 5.1: Minimum, maximum, mean average and standard deviation values for participants' proportion of accuracy and response time (ms) for pre-test task performance, separated by condition.

	Min	Max	Mean	SD	N participants
Lexical decision					
Accuracy	0	1	0.97	0.18	39
RT	334	3102	656	159	39
Semantic decision					
Accuracy	0	1	0.90	0.30	42
RT	589	4393	1190	354	42

Note: LD = lexical decision; SD = semantic decision.

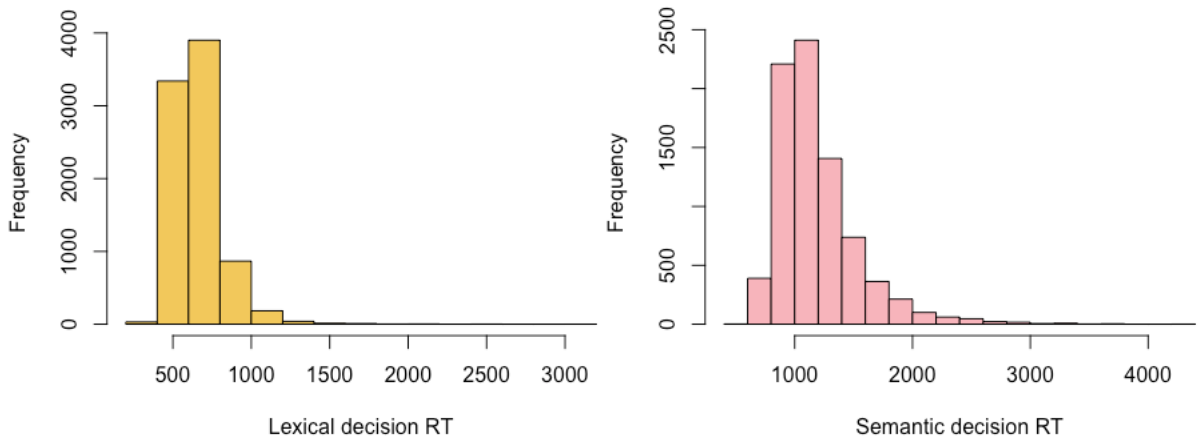


Figure 5.3: Histograms showing the distribution of lexical decision (left) and semantic decision (right) response times.

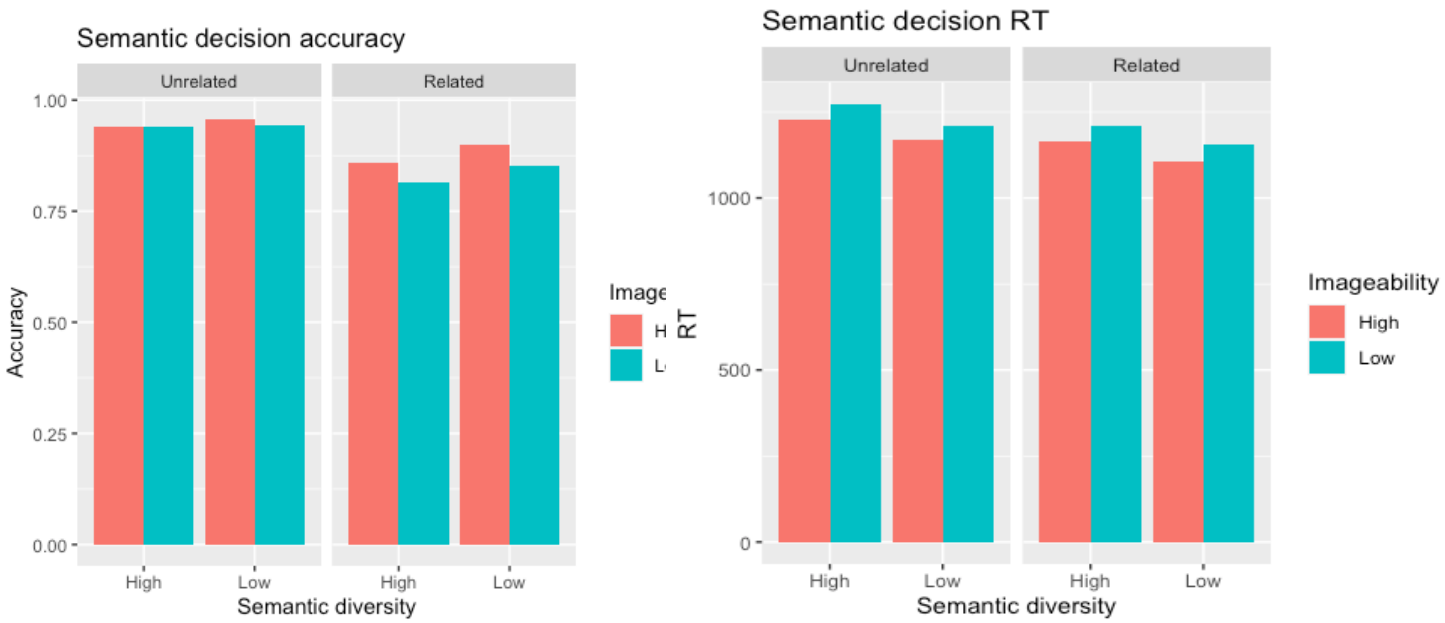
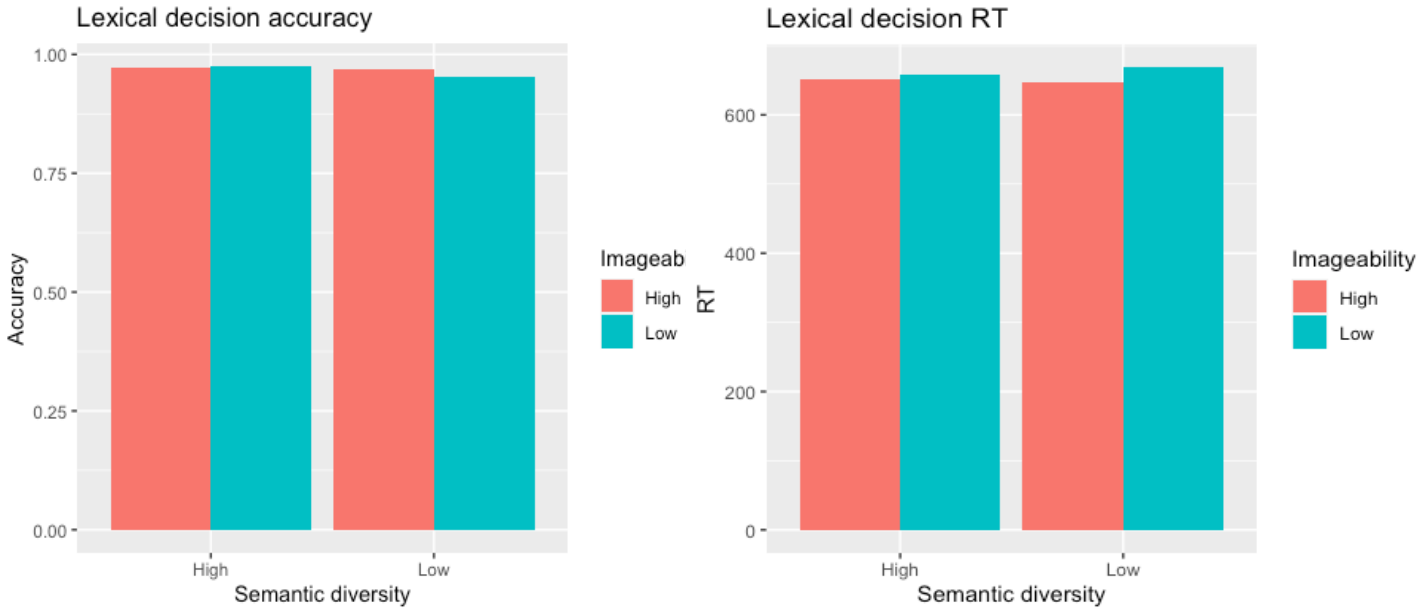


Figure 5.4: Bar plots showing the average performance for words in lexical decision accuracy (top left) and response time (top right) and semantic decision accuracy (bottom left) and response time (bottom right).

Note: Bars are split by semantic diversity and imageability. Semantic decision is split by relatedness,



with semantically unrelated trials on the left and semantically related trials on the right.

Linear mixed-effect models were used to determine whether semantic diversity and imageability predicted pre-test performance, and potential interactions between these variables. For the semantic decision task, we already included relatedness (whether or not the target word was related to its pair) as a predictor. Separate models were tested for lexical decision and semantic decision performance. Tests were run with the package 'lme4', and p-values determined using 'lmerTest'. We included random effects of word and participants. Originally, we included by-participant slopes for semantic diversity and imageability using the model:

Decision performance ~ semantic diversity * imageability + (1| word) + (semantic diversity * imageability | participant)

This model failed to converge. Slopes for imageability were removed, but the model resulted in a singular fit. The model was then simplified again by removing semantic diversity slopes. For lexical decision, this model was:

Lexical decision performance ~ semantic diversity * imageability + (1|word) + (1| participant)

For semantic decision, this model was:

Semantic decision performance ~ semantic diversity * imageability * relatedness + (1|word) + (1| participant)

Semantic diversity and imageability values were dummy coded categorical variables, reflecting high or low variability. Contrast coding was used for low semantic diversity (-0.5) and high semantic diversity (0.5), and for low imageability (-0.5) and high imageability (0.5).

For lexical decision data, semantic diversity did not explain variance in response time, but imageability did, with high imageability words associated with faster decision times. There was no interaction between semantic diversity and imageability. For accuracy, semantic diversity

significantly predicted accuracy, while imageability and the interaction were non-significant (see Table 5.2).

Semantic decision response time was negatively predicted by semantic diversity and positively predicted by imageability, replicating the results of Hoffman and Woollams (see Table 5.3). It was also predicted by relatedness, with participants making decisions faster for word-pairs that were semantically related. None of the tested interactions were significant. For semantic decision accuracy, relatedness was a significant predictor, with unrelated words responded to more accurately than related words. The interaction between imageability and relatedness was also significant, and all other fixed effects non-significant.

Table 5.2: Fixed effects predicting lexical decision response time (above) and accuracy (below) for the model: Lexical decision performance ~ semantic diversity * imageability + (1|word) + (1| participant).

	Estimate	Std Error	df	t	p
Intercept	6.47	0.02	40.98	386.59	<.001
Semantic diversity	-0.01	0.01	225.42	-0.78	.44
Imageability	-0.02	0.01	225.46	-3.04	.03
Interaction	0.02	0.02	225.46	1.50	.13

	Estimate	Std Error	df	t	p
Intercept	0.97	0.01	51.71	162.08	<.001
Semantic diversity	0.01	0.01	229.12	2.16	.03
Imageability	0.01	0.01	229.14	1.22	.23
Interaction	-0.02	0.01	229.13	-1.70	.09

Note: semantic diversity and imageability were entered as binary variables (high or low).

Table 5.3: Fixed effects predicting semantic relatedness decision response time (above) and accuracy (below) for the model: Semantic decision performance ~ semantic diversity * imageability * relatedness + (1|word) + (1| participant).

	Estimate	Std Error	df	t	p
Intercept	7.07	0.02	43.08	304.87	<.001
Semantic diversity	0.04	0.01	442.3	4.62	<.001
Imageability	-0.03	0.01	442.4	-3.63	<.001
Relatedness	-0.06	0.00	7871	-10.89	<.001
SD * I	-0.01	0.01	442.3	-0.41	.66
SD * R	0.00	0.01	7869	0.47	.94
I * R	0.00	0.01	7869	0.43	.82
SD * I * R	0.02	0.02	7869	1.24	.36

	Estimate	Std Error	df	t	p
Intercept	0.95	0.01	70.92	79.85	<.001
Semantic diversity	-0.02	0.01	362.3	-1.29	.20
Imageability	0.00	0.01	362.2	0.29	.77
Relatedness	-0.09	0.01	9034	-15.91	<.001
SD * I	-0.01	0.03	362.3	-0.37	.71
SD * R	-0.02	0.01	9034	-1.46	.14
I * R	0.05	0.01	9033	3.88	<.001
SD * I * R	0.00	0.01	9034	0.11	.92

Note: semantic diversity, imageability, and relatedness were entered as binary variables.

To remain consistent with Hoffman and Woollams (2015), we also analysed results using a 2x2 ANOVA. Here, results did not replicate. Neither semantic diversity or imageability explained variance in lexical decision accuracy or response time. Both diversity and imageability were associated with more accurate semantic decisions, but neither measure explained variance in

semantic decision response time. We also tested linear mixed effect models where semantic diversity and imageability were entered as continuous variables. These results were similar to those reported in this section for lexical decision, but both semantic diversity and imageability both failed to explain variance in semantic decision. Full results from this analysis are reported in Appendix 12.

Summary

When using linear mixed effect models, results were generally consistent with those of Hoffman and Woollams (2015). Words high in semantic diversity were associated with more efficient processing in lexical decision and more effortful processing in semantic decision; while words high in imageability were associated with more efficient processing in both tasks. However, unlike Hoffman and Woollams' findings, the results were not consistent across accuracy and response time.

These results were not stable across different analysis methods: all variables were insignificant predictors using a 2x2 ANOVA, and both diversity and imageability failed to explain significant variance in semantic decision when entered continuously in a linear mixed effect model (see Appendix 12). This adds to the discussion made by Hsiao et al. (2020), who initially failed to find an effect of semantic diversity on lexical decision times. However, in another experiment, Hsiao et al. demonstrated that these effects were present when using the same items and analytical methods as Hoffman and Woollams. Together with the results of this chapter, these findings demonstrate how different statistical methods can yield different results. Hsiao et al. (2020) also replicated Hoffman and Woollams' results using largescale data from the British Lexicon Project. This might therefore also point to an underpowered sample. It could be that the number of words or participants included in our experiment and Hsiao et al.'s initial experiment is not large enough to yield consistent, robust results.

THE EFFECT OF PRE-TEST CONDITION ON LEXICAL DECISION PERFORMANCE AT POST-TEST

To answer this question, the analysis took two steps. First, we used a linear mixed effects model to determine whether there was a difference between pre-test and post-test lexical decision. This part of our analysis was not preregistered but was important as it allowed us to directly compare pre-test and post-test performance. Here, we only included participants who took part in lexical decision for both tasks (N = 39). This allowed us to test whether lexical processing would be more efficient after prior exposure.

Next, we tested a linear mixed effects model to determine whether there was a difference in post-test lexical decision performance between those participants who had previously completed a lexical decision task vs those who had completed a semantic relatedness task.

Descriptive statistics for pre-test and post-test performance from the lexical decision condition are shown in Table 5.4. Raw data had high levels of skew and kurtosis, so log transformations were performed. Distributions for post-test data are shown in Figure 5.5.

The first part of our analysis tested the following model:

LD performance ~ test time * semantic diversity * imageability + (1|word) + (semantic diversity + imageability | participant)

'Test time' referred to whether the lexical decision task was taken pre- or post-test. This model identified a singular fit for both accuracy and response time. Removing the random intercept for words did not resolve this. Slopes for imageability by-participant were removed, but the fit was still singular. We then removed semantic diversity slopes. This gave the following model, reported here:

LD performance ~ test time * semantic diversity * imageability + (1|word) + (1 | participant)

For response time (Table 5.5), the model revealed a main effect of testing time, with participants' responses faster post-test than during pre-test. There were no significant main effects of semantic diversity and imageability, and interactions were also non-significant. For accuracy, no significant effects were observed (see Table 5.6). Response time for pre-test vs post-test lexical decision is shown in Figure 5.6.

Table 5.4: Minimum, maximum, mean average and standard deviation values for lexical decision accuracy (as a proportion) and response times (ms) of participants in the lexical decision condition (N = 39), collapsed across conditions.

All items (N = 40 words)				
Measure	Min	Max	Mean	SD
Pre-test				
Accuracy	0	1	0.97	0.18
RT	334	3102	656	159
Post-test				
Accuracy	0	1	0.96	0.19
RT	280	2241	623	143

Table 5.5: Fixed effects for the model: LD RT ~ test time * semantic diversity * imageability + (1|word) + (1 | participant).

	Estimate	Std Error	df	t	p
Intercept	619.62	11.59	53.99	53.45	<.001
Test time	33.16	4.03	17645.93	8.23	<.001
Semantic diversity	-1.09	6.78	344.48	-0.16	.87

Imageability	7.44	6.78	345.61	1.10	.27
Pre/post*diversity	-3.75	5.71	17645.93	-0.66	.51
Pre/post*imageability	-0.53	5.72	17645.93	-0.10	.93
Diversity*imageability	3.73	9.59	346.07	0.39	.70
Pre/post*diversity* imageability	13.19	8.10	17645.93	1.63	.10

Table 5.6: Fixed effects for the model: LD accuracy ~ test time * semantic diversity * imageability + (1|word) + (1 | participant).

	Estimate	Std Error	df	t	p
Intercept	0.97	0.01	12.76	122.72	<.001
Test time	0.01	0.01	1764	1.18	.24
Semantic diversity	-0.00	0.01	380.1	-0.34	.73
Imageability	-0.00	0.01	1764	-0.24	.81
Pre/post*diversity	0.00	0.01	1764	0.00	.99
Pre/post*imageability	0.00	0.01	1764.93	0.61	.54
Diversity*imageability	-0.00	0.01	380.7	-0.21	.84
Pre/post*diversity* imageability	-0.02	0.01	1764	-1.66	.09

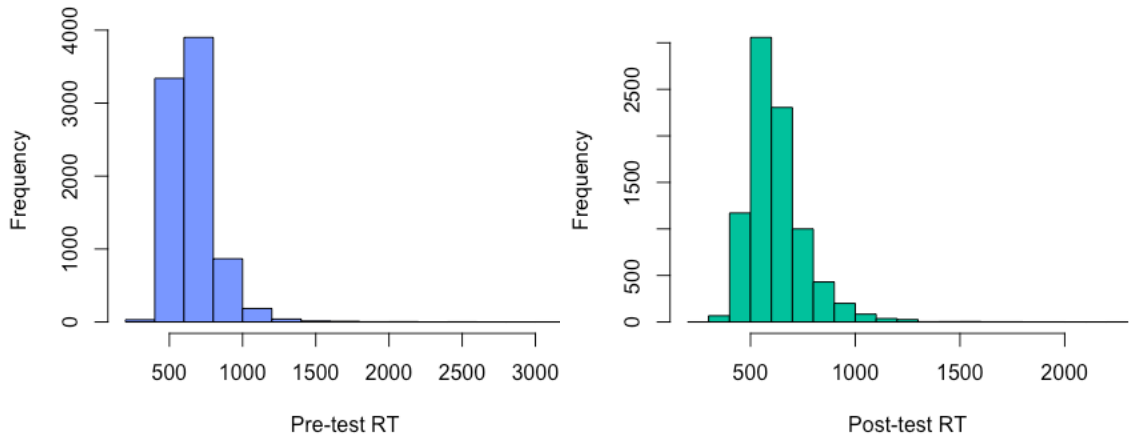


Figure 5.5: Histograms showing the distribution of lexical decision response times for pre-test (left) and post-test (right) tasks.

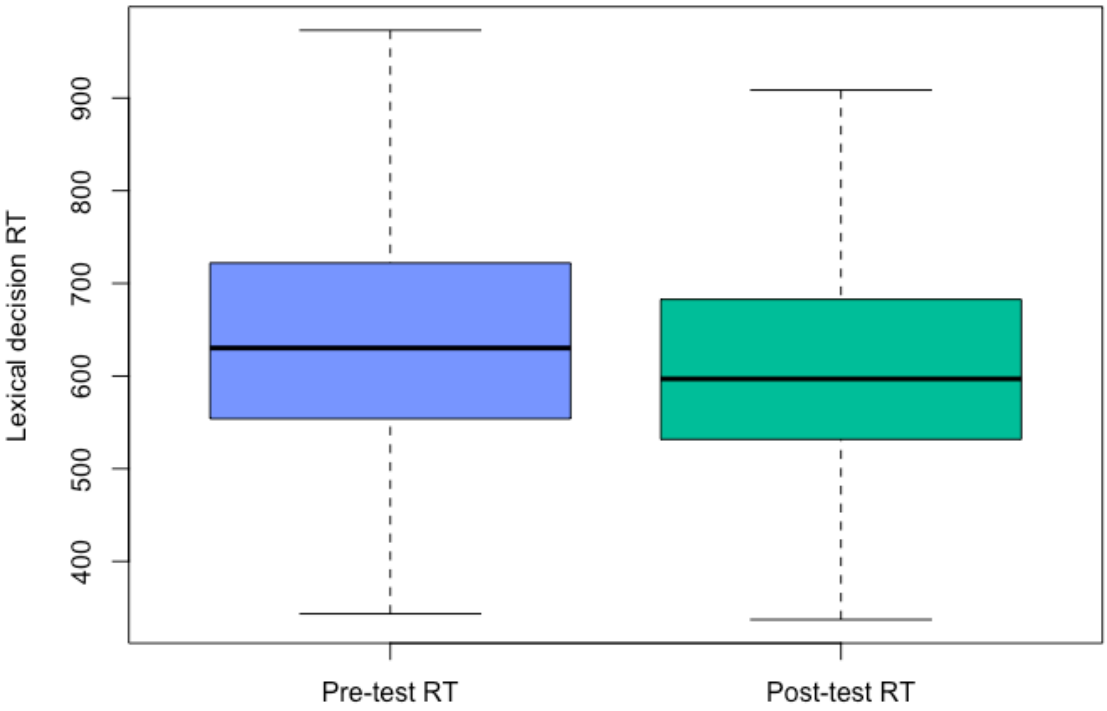


Figure 5.6: Box plots showing the means and distributions of pre-test and post-test RT in the lexical decision condition.

Next, we tested whether post-test performance differed between participants in the lexical decision and the semantic decision pre-test conditions. Descriptive statistics for post-test performance are shown in Table 5.7 for participants who made lexical decisions pre-test, and in Table 5.8 for participants who made semantic decisions pre-test. Distributions are shown in Figure 5.8.

Table 5.7: Descriptive statistics for post-test lexical decision accuracy (proportional; upper section) and response time (ms; lower section) for participants who made lexical decisions pre-test. Minimum, maximum, mean average, standard deviation and number of participants are shown.

Accuracy				
	HIHD	HILD	LIHD	LILD
Min	0	0	0	0
Max	1	1	1	1
Mean	0.97	0.96	0.96	0.96
SD	0.18	0.19	0.19	0.20
N part	39	39	39	39

RT				
	HIHD	HILD	LIHD	LILD
Min	303	280	354	361
Max	2241	1547	1811	1943
Mean	619	618	625	629
SD	137	136	146	144
N	39	39	39	39

Note: HIHD = high imageability, high diversity. HILD = high imageability, low diversity. LIHD = low imageability, high diversity. LILD = low imageability, low diversity.

Table 5.8: Descriptive statistics for post-test lexical decision accuracy (proportional; upper section) and response time (ms; lower section) for participants who made semantic decisions pre-test. Minimum, maximum, mean average, standard deviation and number of participants are shown.

Accuracy				
	HIHD	HILD	LIHD	LILD
Min	0	0	0	0
Max	1	1	1	1
Mean	0.96	0.96	0.96	0.94
SD	0.19	0.19	0.19	0.23
N part	42	42	42	42

RT				
	HIHD	HILD	LIHD	LILD
Min	335	335	305	172
Max	2874	3934	6430	4593
Mean	627	628	637	638
SD	140	155	197	163
N	42	42	42	42

Note: HIHD = high imageability, high diversity. HILD = high imageability, low diversity. LIHD = low imageability, high diversity. LILD = low imageability, low diversity.

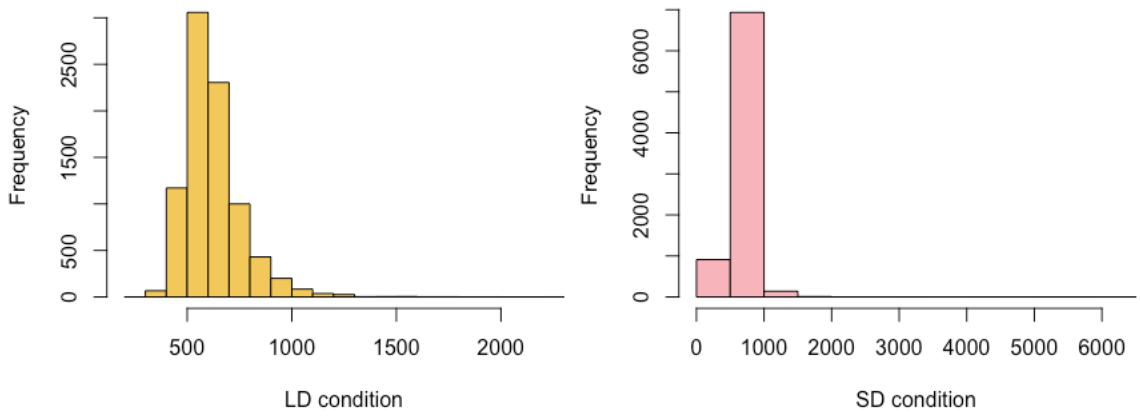


Figure 5.7: Histograms showing the distribution of post-test lexical decision response times for the lexical decision condition (left) and the semantic decision condition (right).

Here, we tested the model:

Post-test performance ~ pre-test condition * semantic diversity * imageability + (1 | word) + (semantic diversity + imageability | Participant)

As with our above analysis, the model failed to converge. This was simplified by removing by-participants slopes for imageability (giving a singular fit) and for semantic diversity. This gave the following model, for which we report our results:

Post-test performance ~ pre-test condition * semantic diversity * imageability + (1 | word) + (1 | Participant)

No significant main effects or interactions were found for this model for either response time (Table 5.9) or accuracy (Table 5.10). These results are shown in Figure 5.8.

Table 5.9: Fixed effects for the model: Post-test RT ~ pre-test condition * semantic diversity * imageability + (1 | word) + (1 | Participant)

	Estimate	Std Error	df	t	p
Intercept	619.7	11.01	103.1	56.30	<.001
Condition	624.3	14.70	88.76	0.43	.67
Semantic diversity	-1.41	5.92	402.3	-0.24	.81
Imageability	7.02	5.93	404.0	1.18	.24
Condition*diversity	0.85	5.72	1791	0.15	.88
Condition* imageability	0.06	5.75	1791	0.01	.99
Diversity*imageability	4.14	8.34	404.8	0.49	.62
Condition*diversity* imageability	-2.53	8.11	1791	-0.31	.76

Table 5.10: Fixed effects for the model: Post-test accuracy ~ pre-test condition * semantic diversity * imageability + (1 | word) + (1 | Participant)

	Estimate	Std Error	df	t	p
Intercept	0.97	0.01	229.9	119.14	<.001
Condition	-0.01	0.01	143.7	-0.71	.48
Semantic diversity	-0.00	0.01	395.1	-0.31	.76
Imageability	-0.00	0.01	396.8	-0.24	.81
Condition*diversity	0.00	0.01	1790	0.55	.58
Condition* imageability	-0.00	0.01	1791	0.64	.52
Diversity*imageability	-0.01	0.01	397.5	-0.20	.84
Condition*diversity* imageability	-0.01	0.01	1791	-1.58	.11

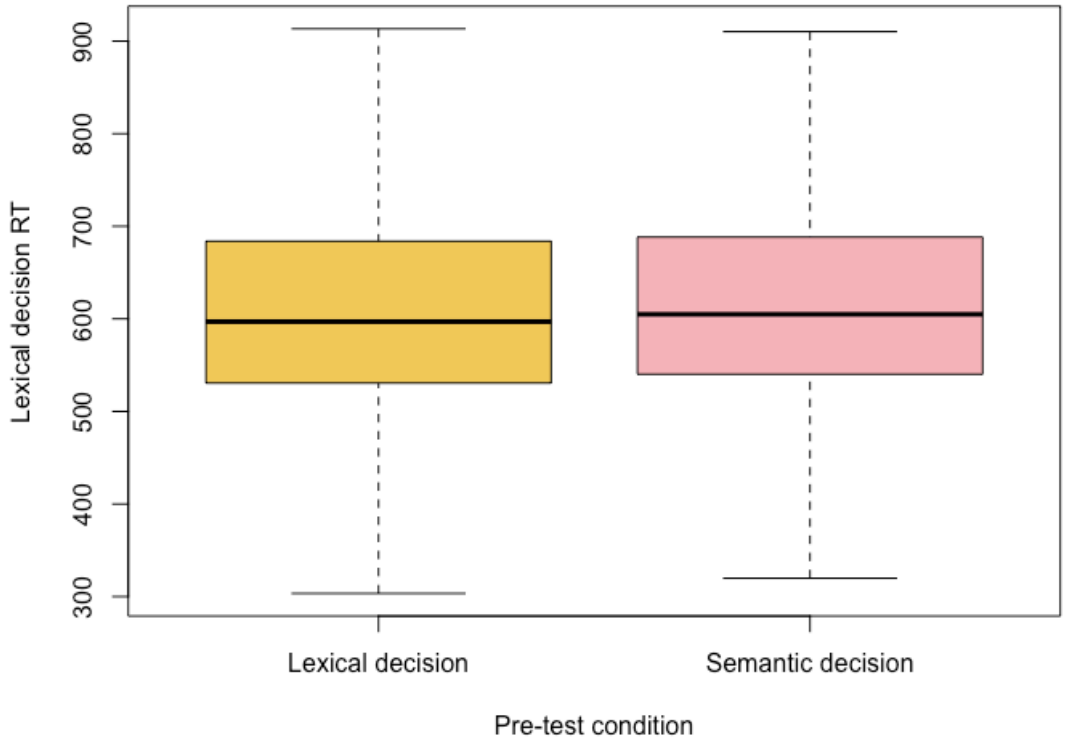


Figure 5.8: Box plots showing the means and distributions of post-test response time for the lexical decision and semantic decision pre-test conditions.

We then repeated the analysis reported in 5.3.1 to test the effects of semantic diversity and imageability on primed (post-test) lexical decision (see Figure 5.9). Conditions were analysed separately. For participants in the lexical decision condition, words high in imageability were associated with faster post-test response times, but imageability did not affect accuracy. Semantic diversity had no effect on either accuracy or response time, and the interaction between diversity and imageability was also non-significant (see Table 5.11). The same pattern of results was seen for participants who made semantic decisions pre-test (see Table 5.12).

Table 5.11: Fixed effects predicting post-test response time (above) and accuracy (below) for participants who completed lexical decisions pre-test. This tested the model: Lexical decision performance ~ semantic diversity * imageability + (1|word) + (1| participant).

	Estimate	Std Error	df	t	p
Intercept	623.93	10.97	39.86	56.86	<.001
Semantic diversity	-2.20	4.28	225.78	-0.52	.61
Imageability	-8.87	4.28	225.84	-2.02	.04
Interaction	5.10	8.56	225.84	0.60	.55

	Estimate	Std Error	df	t	p
Intercept	0.96	0.01	43.58	151.21	<.001
Semantic diversity	0.00	0.01	228.29	0.73	.47
Imageability	0.00	0.01	228.31	0.58	.56
Interaction	-0.00	0.01	228.30	-0.20	.84

Note: semantic diversity and imageability were entered as binary variables (high or low).

Table 5.12: Fixed effects predicting post-test response time (above) and accuracy (below) for participants who completed semantic decisions pre-test. This tested the model: Lexical decision performance ~ semantic diversity * imageability + (1|word) + (1| participant).

	Estimate	Std Error	df	t	p
Intercept	631.76	9.62	43.00	65.68	<.001
Semantic diversity	-1.86	4.50	221.49	-0.41	.68
Imageability	-9.03	4.50	221.55	-2.01	.05
Interaction	-0.80	9.01	221.44	-0.09	.92

	Estimate	Std Error	df	t	p
Intercept	0.96	0.01	57.19	157.85	<.001
Semantic diversity	0.01	0.01	231.99	1.32	.19
Imageability	0.01	0.01	231.98	1.12	.26
Interaction	-0.02	0.01	232.03	-1.60	.11

Note: semantic diversity and imageability were entered as binary variables (high or low).

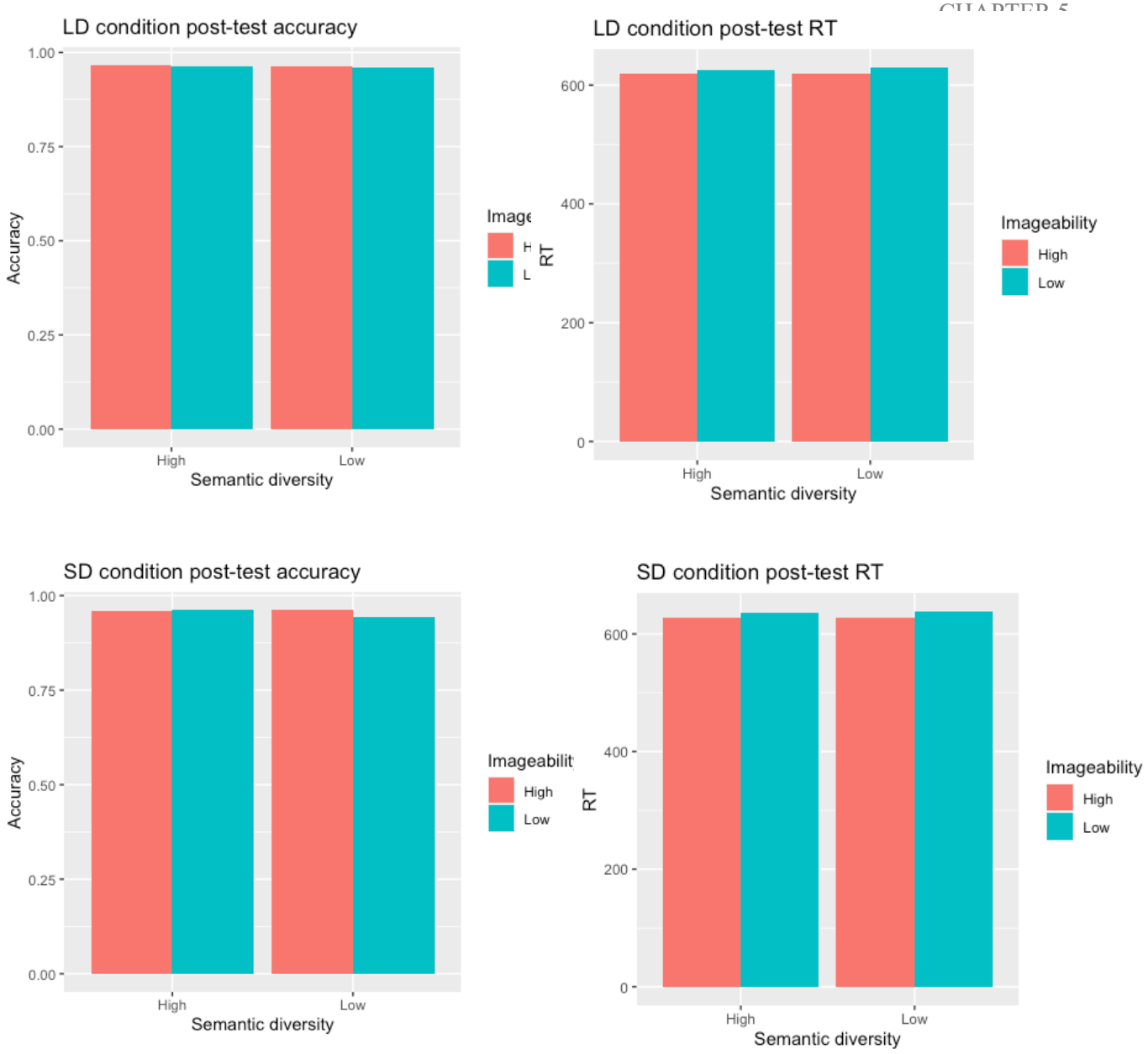


Figure 5.9: Bar plots showing the average post-test lexical decision performance for participants who made lexical decisions pre-test (accuracy top left; RT top right) and participants who made semantic decisions pre-test (accuracy bottom left; RT bottom right).

Note: Bars are split by semantic diversity and imageability.

SUMMARY

For those participants who took part in lexical decision tasks during both pre-test and post-test, lexical decision performance was significantly faster, but not more accurate, at post-test. This

supports our hypothesis that processing becomes faster after previous exposure, and is fitting with theories of lexical organisation in memory (MacDonald, 2013). We did not find any difference in post-test performance between pre-test conditions (lexical decision vs semantic decision). This suggests that both form-based and meaning-based processing had a similar sized effect on lexical subsequent processing.

When participants encountered words for the second time (primed), semantic diversity no longer appeared to affect lexical decision performance. Imageability was still associated with faster responses, regardless of the pre-test condition. This suggests that recent experience can change the extent to which long-term contextual histories influence our current processing.

DISCUSSION

The results from this chapter add to the growing literature reporting differing effects of semantic diversity across tasks. Generally, these findings support those of Hoffman and Woollams (2015) and Hsiao et al. (2020), with high diversity words associated with more efficient lexical processing and less efficient semantic processing; and high imageability words advantaged in both tasks. However, these results did not appear robust. Semantic diversity did not explain significant variance in lexical decision response time or semantic decision accuracy, while imageability did not explain variance in accuracy for either task. Null findings in lexical decision accuracy might reflect a ceiling effect; participants in this condition scored 97% accuracy pre-test.

Hsiao et al. (2020, Experiment 2) initially failed to find an effect of semantic diversity on lexical decision times. Their experiment controlled for word frequency using linear mixed effect models. In a further experiment, Hsiao et al. matched for frequency instead of entering word frequency as a predictor in analysis. With this change, semantic diversity did have a significant

effect. Hsiao et al. concluded that semantic diversity could be confounded by frequency. As the effect was also found using largescale secondary data, Hsiao et al. ventured that this effect was also related to sample size, and was likely small in nature. Our results add to this discussion. In the current experiment, words (taken from Hoffman and Woollams' materials) were matched for word frequency, and frequency was not entered as a predictor. Despite this, diversity did not explain significant variance in lexical decision times. This suggests that the effects of semantic diversity might be smaller than results of previous studies might suggest. We also found non-significant effects of imageability on lexical decision accuracy, and of all variables on semantic decision accuracy. These results may also be due to the small size of these effects.

Hoffman and Woollams (2015) treated their variables categorically. When repeating this experiment in a continuous design, Hsiao et al. (2020) first yielded mixed results: at first, semantic diversity had no effect on lexical decision in adults; but this effect was significant when using the same items as Hoffman and Woollams. Their results were also significant when treating data categorically. The analysis reported in this chapter used linear mixed effect models with categorical predictors. To remain consistent with Hoffman and Woollams, we also analysed results using a 2x2 ANOVA; and, to build upon Hsiao et al.'s discussion, a linear mixed effect model with predictors entered continuously (see Appendix 12).

In our main analysis, the tested model did not include by-participant slopes due to the model first not converging and then producing a singular fit. In hindsight, this could suggest that a continuous approach would be preferable, as using categorical variables can lead to 'overfitting' as the model is often more flexible than it needs to be (Hawkins, 2004). However, our continuous models reported in Appendix 12 also failed to converge. This could be due to overparameterization (i.e., due to the model being too complex for the data). Matuschek et al. (2017) suggested that overparameterization can be improved via both simplification of the model or via collecting a

sufficient sample of data for every item and subject. This suggests that a larger sample might be needed in order to use linear mixed effect models effectively. ANOVA, by comparison, does include slopes, with ANOVA in R removing overparameterization to give a 'just' parameterised model (R Core Team, 2013). This suggests that ANOVA this might be a more appropriate test if treating the data categorically. But ANOVA can also be less powerful than linear mixed effect models, particularly in cases of missing data (Baayan, 2010b).

A better way of analysing the data could be to use continuous variables. The items in this experiment were taken from Hoffman and Woollams, who chose words categorically, thus supporting a categorical analysis. However, continuous variables are generally preferred to categorical as dichotomizing values loses information, which can lead to reduced power and can increase the risk of false positives (Altman & Royston, 2006). Appendix 12 demonstrates that the analytical method is an important consideration. Results were generally insignificant using ANOVA, and both semantic diversity and imageability did not explain variance in semantic decision when entered continuously. It was not possible to include by-participant slopes in any of our analyses, whether categorical or continuous. This suggests that the issue might be one of overparameterization, perhaps related to an inadequate sample. A future, preregistered replication in a larger and more continuous itemset could analyse data continuous and with by-participant slopes. This would reduce likelihood of the model failing to converge and could reduce the problems associated with categorical analysis.

This disparity between these results could also demonstrate the small nature of the effect size: the slight differences in diversity when entered continuous did not appear powerful enough to produce a significant main effect in the linear mixed effect model; and similarly, ANOVA may have not been powerful enough to detect the effects. This might also demonstrate differences in the relationship of long-term experience between lexical decision and semantic decision; the analysis of

Appendix 12 might suggest that the magnitude of diversity and imageability effects is larger for lexical than semantic decision. This could relate to the different processes involved in each task, discussed in the introduction.

Turning back to the disparities between our results and those of Hoffman and Woollams, it appears unlikely that the inconsistency of the results is due to differences in experimental design. The pre-test tasks constitute as a near-replication of Hoffman and Woollams' method. We used a similar sample size (larger than that of Hoffman and Woollams, but smaller than that of Hsiao et al., 2020) and took the itemset directly from Hoffman and Woollams' study. In hindsight, a larger sample size would have been more appropriate. Semantic diversity effects appear to be small (Hsiao et al., 2020) and our results could be underpowered. Future research could use a power calculator or simulation method to determine the appropriate sample size needed to test a small effect.

While our design constitutes as a near-replication of Hoffman and Woollams, it is also important to acknowledge the differences between the current experiment and previous work. The semantic judgement task in the current study presented target words with related or unrelated word pairs simultaneously on-screen. This is different to Hoffman and Woollams, who presented word-pairs separately on screen, and Hsiao et al., who presented visual target words accompanied by auditory pairs. Our task might have been more difficult for participants, and this might have resulted in longer response times. Indeed, while Hoffman and Woollams excluded participants who took longer than 2000ms on >5% of trials, we changed this to 4000ms in our analysis as 2000ms was too stringent and resulted in a large loss of data. While it is possible that the difference in task design affected the results, it is unlikely to fully explain the lack of significance in accuracy. Mean accuracy for this task (90.07%) was not at ceiling, and indeed was lower than mean accuracy in lexical decision (96.73%), where significant effects were observed. The results therefore likely reflect the small effect size of semantic diversity and imageability on semantic decision.

This study also tested the effects of lexical processing via lexical or semantic decision on subsequent lexical decision performance. Lexical decision performance was better during task 2 than task 1, suggesting that prior processing did facilitate subsequent processing. These results can be seen to support MacDonald's (2013) Production-Distribution-Comprehension theory, in which words are organised in a way that prioritises words which are most likely to be retrieved. In this theory, words that have recently been retrieved from memory are made more accessible, therefore easing subsequent processing. While post-test lexical decision was faster than pre-test lexical condition, no distinction was seen between pre-test conditions for post-test performance. This suggests that, while the literature reports differences in the mechanisms involved in form-based vs meaning-based processing, these different mechanisms do not result in a discernible difference in subsequent processing. Rather, both types of processing can be seen to make words more accessible for later retrieval. These results also suggest that, so long as words are repeated, there will be a priming effect regardless of the task demands. However, the current study only tested participants' responses to pairs of words. Priming might exert different effects when tasks involve processing words within contexts.

We proposed that semantic decision might lead to a greater boost in performance as this involves deeper processing. In the introduction, we equated this to semantic richness effects on lexical decision, in which semantically rich words activate more meaning-based knowledge and therefore are easier to process. Conversely, making semantic decisions does not appear to facilitate processing to a greater extent than making lexical decisions. This finding highlights the difference between long-term experience indexed by lexical variability and the active experience of encountering a word. The semantic knowledge is associated with a word across long-term experience, thus constituting its richness to facilitate lexical decision (Pexman et al., 2002; Yap et al., 2015). Making a decision using semantic information, however, does not boost processing to a larger extent than making a form-based decision, even in the short-term. This could be because both

lexical and semantic decision involve accessing semantic knowledge. Indeed, semantic activation is proposed to explain the semantic richness effects on lexical decision (Balota, 1990), meaning that the activation of meaning-based knowledge we propose for semantic decision is not unique to this task.

We also suggested that lexical decision might provide more facilitation due to repetition priming. This did not appear to be the case. However, while we observed no differences between pre-test conditions, it is possible that facilitation effects manifested through two separate pathways. Lexical decision might have boosted post-test performance due to repetition priming, and semantic decision might have boosted post-test performance due to semantic activation. Or alternatively, both decision tasks might have facilitated subsequent processing via a blend of form repetition and accessing semantic information. The extent to which different types of processing affect performance through different mechanisms cannot be determined by the results reported here and is a question for future research. It is important to also note methodological differences between conditions in the study's design. While participants in the semantic decision task responded to the same target words, the other words in the two tasks (i.e., related/unrelated word pairs for semantic decision vs non-words for lexical decision) were not the same. This could result in a greater practice effect for participants in the lexical decision condition, which might have confounded the results. To test this, future research could assess the effects of lexical decision and semantic decision using a different post-test measure of lexical processing. This would allow the differences in processing of the two pre-test tasks to be compared without confounding variables.

Finally, we tested whether semantic diversity and imageability would affect primed lexical decision performance. Imageability facilitated lexical decision, with small but significant differences in response time in participants from both pre-test conditions. However, there were no main effects of semantic diversity. This could be because semantic diversity has a small effect size; or could be

due to recent experience, which might temporarily change processing in a way that alters the effects of long-term lexical experience. Linear mixed effect models comparing pre- and post-test performance found no significant interaction between group and semantic diversity. This suggests that, while semantic diversity is no longer significant in primed performance, the effect of semantic diversity is not smaller post-test than pre-test. Therefore, this effect could have 'disappeared' due to performance reaching ceiling. Post-test accuracy was high across participants (96%), and response times were faster and less variable than during pre-test. This could have masked item-level variance associated with semantic diversity effects.

Anderson and Milson (1989) argued that memory is arranged in a way that prioritises information which will most likely be retrieved in the future. This idea is echoed in models of language processing which explain how long-term and recent experience shape lexical behaviours. Experience gathered both across the lifetime and in recent past informs the likelihood of words being retrieved, with words that have contextually diverse histories (Nation, 2017) and words that have been recently produced or comprehended (Macdonald, 2013) being more likely to be retrieved, and therefore easier to process. Due to the small effects of lexical variables and the near-ceiling performance post-test, it is difficult accurately assess how long-term experience interacts with recent lexical encounters. In order to further investigate the role of semantic diversity in primed performance, future research could employ a more difficult post-test task. This would result in more variance in participant responses, and therefore more opportunity to observe change in the effects of lexical variables, if they exist.

In summary, we replicated the lexical advantage and semantic disadvantage associated with semantic diversity in processing tasks. We interpret inconsistencies in the results as arising due to small effect sizes. Both lexical and semantic decision tasks appear to facilitate subsequent lexical processing. These results suggest that form-based and meaning-based processing exert similar

priming effects on subsequent performance, but questions remain as to the mechanisms involved in each of these processing types. Primed lexical decision was predicted by imageability but not by semantic diversity, suggesting that relationship between long-term and recent linguistic experience might differ for different types of semantic information. Further research is needed to determine how recent experience changes the effects of contextual histories on lexical processing. This is important to understanding how past lexical information shapes the efficiency with which words can be retrieved from memory.

CHAPTER 6
CONCLUSIONS, CONTRIBUTION TO THE LITERATURE, AND REFLECTIONS FOR FUTURE RESEARCH

INTRODUCTION

This thesis adds to the literature in numerous ways. First, it advances our understanding of lexical-semantic variables. Second, it provides new norms that future researchers can employ for context and sentence availability. Finally, it addresses the relationship between long-term and recent lexical experience and their influences on lexical processing. In this chapter, the contributions of this thesis to the literature are discussed, alongside reflections and suggestions for future research.

UNDERSTANDING LEXICAL-SEMANTIC SPACE

A primary aim of this thesis was to further our understanding of lexical-semantic space. Semantic diversity has been operationalised in different ways, and Chapter 2 explored the relationships between these different measures and other semantic variables. Slight differences were seen across different measures of semantic diversity, but broadly the measures patterned similarly in their relationships with other variables, and in their fit to lexical decision data. Turning to what predicted variation in semantic diversity at an item level, structural equation modelling tested the fit of latent variables to diversity values. This was based on Hoffman et al.'s (2013) principal components analysis, which identified a two-factor structure in lexical-semantic space, with factor 1 reflecting frequency and occurrence information, and factor 2 reflecting experiential, richness information. Importantly, semantic diversity loaded onto both factors. Chapter 2 therefore used

structural equation modelling to test the extent to which occurrence and richness could explain variance in semantic diversity values. The latent factors provided an acceptable fit to semantic diversity as construed by both Hoffman et al. (2013) and Johns (2021), but the two diversity measures appeared to vary in their relationships with occurrence (as indexed by frequency, age of acquisition, and document count) and richness (as indexed by concreteness, imageability, and age of acquisition). Semantic diversity as calculated using Hoffman et al.'s latent semantic analysis approach was predicted positively by occurrence and negatively by richness, demonstrating that this metric captures both linguistic and conceptual information. By comparison, Johns' measure appeared more closely aligned with linguistic information, as tapped by the latent variable reflecting occurrence.

In Chapter 2, we studied a range of semantic variables informed by the measures included in Hoffman et al.'s (2013) principal components analysis. However, our set of variables diverted from Hoffman et al.'s in order to maximise the number of items in our analysis. Particularly, Hoffman et al.'s itemset was limited by the low number of ratings available for context availability (N = 325). This led us to collect a larger set of norms in Chapter 3 for both context availability and a new variable, sentence availability. By collecting new norms, we then replicated Hoffman et al.'s principal components analysis with a larger itemset. This identified numerous facets of semantic variability, with three factors identified relating to occurrence, richness and affect. This supports the work of Pexman et al. (2002; 2004) who described the multifaceted semantic richness which can be tapped by numerous conceptualisations of meaning-based knowledge.

This thesis lays a foundation for understanding the relationships between different conceptualisations of contextual variation, but there are still questions as to what drives variation between these measures. Chapter 2 explored the relationship between different measures of semantic diversity. From the analysis reporting in Chapter 2, it appears that the size and nature of

the corpus might be more important than other calculation differences and that these might shape diversity statistics to a greater extent than theoretical explanations. However, more work is needed to determine whether this is the case. Due to the numerous differences between measures (including corpus size and sample, breadth of context window, calculation method, and consideration of user-based document count), it is difficult to determine why differences between the measures occur. This in turn makes our application to theories of lexical organisation limited.

In order to test whether these differences explain the differences between diversity measures noted in this thesis, a future study could calculate semantic diversity values using a method that isolates one difference. To take an example, in Chapter 2, we proposed that the breadth of the context window on which diversity is based could account for differences observed between Hoffman et al.'s (2013) and Johns' (2021) measures of semantic diversity. Due to the similarities between Johns (2021) and Johns et al. (2020, which had a narrower contextual window), we suggested that this is likely not the most important factor in driving differences between diversity measures, but it nonetheless might change the nature of what diversity reflects. To test this, diversity could be calculated from the same corpus using the same calculation method, with only the size of the context window from which similarity is measured changed. If broad context captures linguistic information similar to document count (giving a measure of occurrence that considers the spread of documents in which a word occurs), while close-proximity context captures experiential information (e.g., motor, sensory and affect), then we would expect this to be reflected in the correlations of the newly devised measures: the broad-context diversity would correlate strongly with document count and be predicted by occurrence, while the narrow-context diversity would have more moderate correlations with document count and might be more closely tied to concreteness and imageability. If these new measures of semantic diversity did not show this pattern, the results discussed in Chapter 2 might instead reflect differences in the corpora used or the calculation method. Another factor that appears to influence semantic diversity is the inclusion

of individual people's word use, thus giving a proxy for the spread of a word across different users and a reflection of individual contextual histories. Johns' (2021) measure included information about how words were distributed across users, but this individual component was not captured by the other diversity metrics explored in Chapter 2. In Johns' (2021) paper, Johns demonstrated that measures which consider individuals' word use provided a better fit to lexical decision data than measures which do not consider individual word use. It is unclear whether this might explain the differences reported in Chapter 2 regarding semantic diversity's place in lexical-semantic space. Taking this approach to isolate each possible methodological difference would allow us to tease apart the multiple factors that could contribute to differences between diversity measures. This would result in a better understanding of what semantic diversity metrics capture and why.

The larger set of context availability ratings collected in Chapter 3 provided an opportunity to replicate Hoffman et al.'s (2013) principal components analysis with a much larger itemset. Here, we also included sentence availability, age of acquisition and a range of emotion-embodiment variables. Our results replicated those reported by Hoffman et al. (2013). We identified a three-factor structure, with two factors ('occurrence' and 'richness') aligning with Hoffman et al.'s factors 1 and 2. Our additional factor 'affect' was characterised by variables relating to emotional experience. These results further our understanding of the multifaceted semantic richness discussed by Pexman et al. (2007). Semantic variables appear to capture numerous constructs that are, at the item-level, distinct from one another. Both semantic diversity and context availability loaded across 'occurrence' and 'richness'. This differed to Hoffman et al.'s findings, in which semantic diversity was unique in loading onto multiple factors. In the factor analysis reported in Table 3.1, semantic diversity loaded positively onto 'occurrence' and negatively onto 'richness', while context availability loaded positively onto both these factors. This highlights a more complex relationship between context availability and semantic diversity than can be observed in the simple correlations, or indeed than was apparent in Hoffman et al.'s analysis of a more limited itemset. Both diversity and context

availability appear to capture an element of occurrence information; but while diversity relates to abstractness, context availability relates to concreteness. These results demonstrate the complex nature of semantic variables more broadly. In line with Pexman et al. (2002), it appears that there are multiple underlying facets of meaning-based knowledge. Different variables tap these facets in different ways, and it can be difficult to determine the extent to which any two variables capture different types of information.

In this thesis, we have explored lexical-semantic space using correlations, factor analysis, and structural equation modelling at the item level. We have also used linear effect models to test the extent to which the variables discussed explain variance in lexical behaviours. A sensible next step would be to bridge these two areas of study together. Using structural equation modelling, future research could test the extent to which latent semantic factors (i.e., occurrence, richness and affect) explain variance in lexical decision latencies. This would give a more thorough understanding of semantic variability and how it relates to word processing.

TOOLS FOR FUTURE RESEARCH

Chapter 3 provided new ratings for context availability, a variable known to explain variance in lexical decision (Schwanenflugel et al, 1988), and for sentence availability, a new variable that captures the ease with which a word can be placed in a sentence context. These norms have been made openly available through Open Science Framework (https://osf.io/gr73b/), meaning that other researchers can employ them in their analysis and experimental work. These norms give opportunity for research to replicate other effects associated with context availability with a larger itemset. To give some examples, Rice et al. (2019) reported an interaction between context availability and polysemy on participants' lexical decisions for 452 words, with less efficient processing observed for

words with few senses and low context availability. Schwanenflugel and Noyes (1996) reported a negative relationship between reading skill and magnitudes of context availability effects based on 184 words. Zdrazilova and Pexman (2013) found that context availability explained variance in lexical decisions but not in semantic categorisation performance for 366 abstract words.

In the literature, context availability has been observed to facilitate lexical processing, but most studies are limited to a small itemset. While largescale itemsets are available for many variables such as frequency and age of acquisition, the largest freely available set of context availability norms is 325 words (Altarriba et al., 1999). Chapter 3 replicated this processing advantage in an itemset of 2119 words: using data from the BLP and ELP, context availability was found to explain unique variance in lexical decision performance. Chapter 3 also presented norms for sentence availability. This new variable appeared to capture linguistic information in a similar way to semantic diversity, but it also correlated positively with context availability, suggesting that it might also be related to experiential information. These ratings were validated in Chapter 4, where a different group of participants provided higher ratings for words with higher sentence availability values. In Chapter 3, we demonstrated that sentence availability explains significant variance in lexical decision performance after semantic diversity and context availability are controlled for. This demonstrates that, while sentence availability correlates with both diversity and context availability, it is also distinct from both these metrics. As this variable is newly proposed in this thesis, questions remain about its nature. Future research could investigate the effects of sentence availability on other types of lexical processing. In particular, studies may wish to test associations between sentence availability and semantic decision performance. As discussed in Chapters 1 and 6, higher values in semantic diversity and other polysemy measures are associated with slowed semantic decisions (Hoffman & Woollams, 2015; Yap et al., 2011), while high values in other variables including frequency (Collins & Quillian, 1969) and concreteness (Pexman et al., 2007) are associated with more efficient semantic processing. By testing the effects of item-level sentence availability on

people's semantic decision performance, researchers could determine whether this metric patterns similarly to semantic diversity (i.e., associated with slowed response times) or to other linguistic and semantic measures (associated with speeded response times). Chapter 3 proposed that sentence availability captures variance similar to that reflected by semantic diversity, but it could be more akin to other linguistic measures like frequency or could give similar affects to context availability.

This could also aid our understanding of semantic diversity. Semantic diversity's effects on semantic processing have been explained through the polysemy account (Hoffman & Woollams, 2015; see Chapter 2 for discussion), and are congruous with the effects of other variables measuring polysemy, like number of senses (Yap et al., 2011). Though slowed semantic processing is also seen when diversity is calculated using the Semantic Distinctiveness Model (Johns et al., 2016a), this finding does not align neatly with the predictions of the distinctiveness model, which emphasises the effects of rich contextual histories on word processing; therefore, Johns et al.'s semantic diversity might also be related to polysemy. While sentence availability correlates positively with number of senses ($r = .36$), we did not consider this to be a measure of polysemy, per se. Instead, we proposed that this measure would capture linguistic variation based on prior statistically accumulated knowledge. By presenting sentence availability as a new measure of linguistic variation, this allows for investigation of the relationship between contextual variability and polysemy. If high sentence availability patterns similarly to semantic diversity in a semantic decision task (i.e., higher sentence availability being associated with slower semantic decisions) this might suggest that linguistic variation and polysemy are inseparable from each other in existing measures; and perhaps conceptually inseparable, as Hoffman et al.'s (2013) account of diversity suggests. In this case, further research would need to determine whether these facets of semantic knowledge (contextual histories and polysemy) are distinct from one another and can be measured in isolation. However, if high sentence availability is associated with faster semantic decisions, then this would demonstrate a clear distinction between the effects of semantic diversity and sentence availability on semantic

processing. This would suggest that semantic diversity is unique in tapping polysemy, and it might then follow that the ambiguity information captured by diversity metrics is driven by the contextual similarities from which diversity is calculated, rather than the broader conceptualisation of linguistic histories. Further, if semantic diversity and sentence availability pattern differently in their relationships with semantic decision performance, this might suggest that they capture different types of semantic information. This demonstrates how our norms can be employed to further our understanding of lexical information and its relationship to linguistic behaviour.

THE ROLE OF CONTEXT IN LANGUAGE PROCESSING

As well as providing tools for future research, this thesis builds upon work investigating the role of contextual experience in language processing by addressing theoretical issues posed by prior research. Chapter 2 replicated previous work demonstrating that four measures of semantic diversity explain unique variance in lexical decision when other lexical variables are controlled for. This chapter also explored different theoretical explanations of semantic diversity: Hoffman et al.'s (2013) polysemy account and Jones et al.'s (2012) Semantic Distinctiveness Model. While measures based on the Semantic Distinctiveness Model (Johns et al., 2020; Johns, 2021) appeared the best fit to lexical decision latencies, the differences between this and Hoffman et al.'s measure were very small. Such differences could also be driven by differences in the way diversity is calculated, as demonstrated by the disparity between Hoffman et al. and Cevoli et al.'s measures from the use of scaled vs unscaled vectors; or differences in the corpora on which diversity is based. Our results therefore cannot determine whether either theoretical account is preferable, but this chapter does lay the groundwork for future studies to determine what exactly drives differences between semantic diversity measures. Either way, these results confirm that factors associated with a word's context history are associated with lexical decision, beyond word frequency.

This thesis also explored contextual experience through the lens of availability. Chapter 3 demonstrated that contextual and sentence availability measures tap the different types of semantic representation, similar to those described by Vigliocco et al. (2009). Context availability captured experiential information and sentence availability captured linguistic information. Higher levels of both types of availability facilitated lexical decision performance, and were distinct from one another. This demonstrates that there appear to be multiple facets of contextual information. Measures like semantic diversity and sentence availability (though more research is needed into the latter) capture the lexical contexts that words appear in. In the case of semantic diversity, this reflects the contextual history built through years of experience. For sentence availability, we postulate that this might also be driven by information developed through experience of past linguistic contexts. In contrast, context availability is based upon a very different type of context. Rather than the linguistic settings of words, this considers the imagined and more conceptual circumstances attached to words in a similar way to concreteness and imageability. The finding that words experienced in more varied linguistic contexts and more salient experiential contexts are processed more easily in lexical decision has implications for our understanding of lexical processing. The findings here are not altogether new: advantages in lexical decision have long been reported for words high in context availability (Schwanenflugel et al., 1988) and other experiential measures, like concreteness (Barber et al., 2013) and imageability (Balota & Cortese, 2004). However, through the lens of context and sentence availability (two tasks that are very similar in nature yet tap distinct semantic constructs) we can examine the effects of these two types of semantic information.

The role of contextual experience was also explored in Chapters 4 and 5. In Chapter 4, sentence availability was associated with faster reading and rating times. Congruous with research on semantic diversity (e.g., Johns et al, 2016; Johns et al., 2020), this suggests that the linguistic variation leads to more efficient lexical processing. Chapter 5 replicated the results of Hoffman and Woollams (2015), in which imageability facilitated both lexical decision and semantic decision, while

increased semantic diversity was associated with more efficient lexical decisions but less efficient semantic decisions. These results add to the growing literature demonstrating the effects of linguistic vs experiential information on lexical behaviour. Experience of increased linguistic variation (as captured by semantic diversity and sentence availability) gives an advantage to processing when tasks involve processing a word's form (lexical decision), when the word is considered in isolation and without explicit reference to any of its meanings or usages, or when the word is read in a contextually informative sentence. This same variation in experience is associated with a disadvantage for semantic relatedness decision. This could be due to the semantic precision required in this task, as Armstrong and Plaut (2008) argued when reflecting on the relationship between polysemy and lexical processing.

RELATIONSHIP BETWEEN LONG-TERM AND RECENT EXPERIENCE

While semantic diversity and sentence availability capture the linguistic experience accumulated across years of experience, Chapters 4 and 5 investigated the role of short-term, recent experience on subsequent lexical processing. Theories like the Production-Distribution-Comprehension hypothesis (MacDonald, 2013) suggest that recent encounters with language shape subsequent processing; and this idea can also be seen in longer-term accounts of language. For example, the Semantic Distinctiveness Model (Jones et al., 2012) describes the active updating of word knowledge via new contextual experiences, suggesting that each encounter with a word will shape the efficiency with which it can be processed in the future. By employing priming paradigms, we explored the role of recent experience on processing, and asked how this relates to the contextual history that words have already accumulated, as estimated by variables such as semantic diversity and sentence availability.

The priming study described in Chapter 4 was an ambitious attempt to link together the effects of long-term and short-term lexical experience in lexical processing. We failed to find any evidence that recent experience facilitates subsequent lexical processing for either production or comprehension; and a significant interaction found that high availability words were associated with slower reading times when primed. These findings were contrary to our predictions, and (partly due to our experimental complex design) it was difficult to determine what these results reflected. It could be that our null results were due to methodological choices. Alternatively, there could be a genuine small effect of inhibitive priming that was specific to high availability words. As discussed in Chapter 4, our experiment differed from other studies investigating priming effects in production and comprehension. We used an online self-paced reading task. While this paradigm is widely used in behavioural research and is generally seen as reliable (Keating & Jegerski, 2015), its measurements are arguably less precise than comparable measures like eye-tracking (Jegerski, 2018; though eye-tracking faces unique challenges like calibration, see Sharma et al., 2021), and its utility when used in an online study is yet to be established. We also included a sentence availability rating task in which participants were invited to rate how easy or difficult it was to associate a given word with a sentence. This was based on the assumption that participant response time accurately reflected sentence production, and that mental production might reflect similar effects on lexical organisation as producing utterances. In hindsight, a simpler design using robustly tested tasks would have been more appropriate to test our hypotheses.

These reflections led us to design the experiment reported in Chapter 5, which tested the effects of long- and short-term experience through an extension of Hoffman and Woollams' (2015) design. In this experiment, participants first took part in either lexical decision or semantic decision, with performance on this initial task reflecting 'unprimed' behaviour. Participants then encountered the same words in a lexical decision task, meaning that these words were effectively primed by the recent experience afforded by the first task. Primed lexical decisions (task 2) were faster than

unprimed lexical decisions (task 1). This supports Macdonald's (2013) Production-Distribution-Comprehension theory, according to which recent lexical experience increases the ease with which words are subsequently processed. Contrasting this to the null main effect reported in Chapter 4, this suggests that priming effects are dependent on experimental design. It could be that priming effects are task-dependant; or it could be that the measures used in Chapter 4 (sentence rating and reading response time) were not sensitive enough to pick up a main effect of priming if one existed. There might have also been other factors in our design, such as the delay between task 1 and 2, which could have led to the null result.

Another reason for the disparity between the results of Chapters 4 and 5 could be the nature of contextual information. In Chapter 4, words in the reading task were framed within a context of two sentences. Therefore, the particular usage or meaning associated with the word was explicit and transparent. In Chapter 5, however, words were presented with minimum contextual information as they appeared alongside just one other word (in semantic decision), or a non-word (in lexical decision). In the absence of a specific context, participants might draw more heavily on a word's contextual history, as this provides semantic information associated with the word that can be relied upon in the absence of a current contextual framing. How words are processed might differ depending on the amount of contextual information available. This is demonstrated in priming studies of ambiguous words: Simpson (1981) presented participants with homonymous (ambiguous) words, then asked participants to make lexical decisions for words that were related to either the homonym's dominant meaning or its subordinate meaning. The researchers demonstrated that, after priming, lexical decision was more efficient for words related to the dominant meaning than for words related to the subordinate meaning. This effect was changed, however, once primes were presented in a disambiguating context that biased a subordinate meaning (also see Rodd et al., 2013; 2016). This demonstrates that recent contextual experience can change the semantic information that is retrieved during lexical processing. Applied to our results, this argument might

explain why priming via rating led to slower subsequent reading times in Chapter 4, while priming via decision tasks facilitated subsequent lexical decision in Chapter 5. In sentence reading, current contextual information informed how the word should be interpreted, while in decision tasks, performance was more influenced by information accumulated over contextual history. While the words used as stimuli in both Chapters 4 and 5 only had one distinct meaning (i.e., ambiguous words were excluded), they did vary in their contextual histories (as reflected in a range of semantic diversity and sentence availability values, and variations in polysemy too), and this could have affected their processing in a similar way to ambiguity. Future research could test whether there are differences in the pattern of priming effects observed between tasks that provide context vs tasks that present words in isolation (and therefore are potentially more sensitive to the influence of nuanced usage, as indexed by a continuous measure of polysemy like semantic diversity). This would shed light on the disparities reported in Chapters 4 and 5.

The experiment reported in Chapter 5 also tested whether semantic diversity and imageability explained variance in primed lexical decision. During this task, all participants had already encountered target words in either a lexical decision or a semantic decision task, with task 1 effectively priming performance in task 2. Words high in imageability were associated with faster lexical decision response times, whereas there was no significant effect of semantic diversity. This contrasts to unprimed lexical decision performance observed in task 1, where both semantic diversity and imageability were associated with more efficient processing. This difference between the results of primed and unprimed lexical decision suggests that recent experience might change how lexical variables affect processing. A recent encounter with a word might temporarily over-ride the influence from its longer-term contextual history. However, this was not observed in Chapter 4: in this experiment, sentence availability (a marker of long-term contextual histories) predicted both primed and unprimed performance, but there were no main effects of priming. This relationship might therefore be dependent on task or items.

Turning to whether the nature of the task used to provide a recent experience and its effects on subsequent processing as captured by priming, Chapter 5 also investigated whether form-based and meaning-based processing would prime subsequent lexical decision differently. The results from this experiment found no difference in priming as a function of whether participants first experienced words in lexical decision (inviting a reflection on word form) vs. semantic decision (inviting a reflection on word meaning), suggesting that experience making decisions about words will facilitate processing regardless of whether tasks involved form-based or meaning-based processing. However, Chapter 4 demonstrated that recent experience does not always facilitate subsequent processing. In this experiment, no main effects of priming were observed, and an interaction suggested that, for words high in sentence availability, processing was more efficient for unprimed than primed words. Considering the results of both experiments together, it appears that the pattern of priming observed is more complex that simple repetition. More generally, priming effects likely vary across items and tasks, as both the word's contextual history and the current context relevant to task demands will influence how a word is interpreted and processed. In Chapter 5 we argued that both lexical decision and semantic decision tasks could be facilitating subsequent processing in different ways. This view is driven by theory rather than data, and further research is needed to determine how different types of word encounters shapes later lexical processing.

While this thesis used priming to examine the updating of contextual histories, it is unlikely that these effects will be enduring. Priming effects tend to be short-lived, while contextual histories are relatively stable (see Rodd et al., 2016). A different way to study this would be by inducing contextual histories to non-words. This could occur within a word learning paradigm. Learning paradigms have previously demonstrated how linguistic knowledge is accrued across time. For example, Joseph et al. (2014) replicated the 'age of acquisition' effect whereby words that are acquired earlier in life are typically easier to process by presenting non-words in incidental reading across multiple days. The researchers measured participants' eye movements during reading and

demonstrated that words learnt earlier in the experiment were processed more efficiently than words learnt later. Joseph and Nation (2018) used a similar design to manipulate the uniformity of contexts in which words appear. They found that children's reading performance did not differ between uniform and diverse contexts, but diversity interacted with comprehension skill. For children with better comprehension skills, more time was spent reading words in diverse contexts, suggesting that this reading was more effortful and demonstrating that semantic diversity indirectly influences children's reading behaviours. Other researchers have reported benefits for diverse contexts in word learning paradigms in adults (e.g., Bolger et al., 2008; Mak et al., 2020). Such paradigms pose an opportunity to track the 'live' updating of word knowledge. Online measures of reading behaviours like eye-tracking could be used to observe how words are processed as they appear in different linguistic contexts. This could help us to understand how recent experience and accumulated experience interact.

REFLECTIONS ON PREREGISTRATION

Preregistration is the process of stating predictions, experimental designs and analysis plans before collecting or analysing the data. This has been proposed as a way to reduce the replication crisis (Open Science Collaboration, 2015). Preregistration helps to draw the distinction between prediction (a-priori hypotheses) and postdiction (where explanations are given after results are known). This is important as, when hypotheses are developed post-hoc, they might be presented with a false sense of certainty. This can lead to overconfidence in results and reduced reproducibility (Nosek et al., 2018). Preregistering analysis plans can also reduce exploitative methods of exploring the data. For example, 'p-hacking' is a practice in which researchers selectively report p values that are statistically significant. By transparently stating analysis plans ahead of time, this makes clear the

number of tests conducted, therefore allowing the reader to interpret results with the appropriate degree of uncertainty (Nosek et al., 2019).

The experiments described in Chapters 4 and 5 were preregistered via *As Predicted* (https://aspredicted.org). Reflecting on this process, the preregistration gave a good opportunity to scrutinise the method and analysis ahead of data collection. However, some of the planned analyses and previously stated decisions were changed once analysis was underway. An example of this is the exclusion criteria for the experiment reported in Chapter 5. In our preregistration, we stated that data would be excluded following the same procedure as Hoffman and Woollams (2015); that is, participants who took longer than 2000ms on more than 5% of trials would be excluded, and trials more than two standard deviations from the participant mean would be trimmed. We initially followed this, but Hoffman and Woollams' guidance resulted in problems with our data. First, some trials were very short, suggesting that participants were not making a thoughtful decision on these trials. We therefore decided to exclude trials shorter than 100ms, following the literature which advises that responses below this limit do not likely reflect genuine responses (Luce, 1968). A second problem following the initial exclusion criteria was that, while the 2000ms threshold was appropriate for lexical decision data, this resulted in a large loss of data for semantic decision, as participants were generally slower for this task. This could be in part due to a difference between our design and that of Hoffman and Woollams. While they presented the target word and its pair on screen separately, we presented both words simultaneously. Our response time measure had as its onset when participants were presented with both words, while the onset in Hoffman and Woollams' experiment was set to the appearance of the second word in the pair. After inspecting the data, we retained the 2000ms criterion for lexical decision but changed the threshold to 4000ms for semantic decision. This example demonstrated that preregistered plans do not always prepare for the reality of your data. While deciding how to adjust our exclusion criteria, there were numerous decisions we could have made, and it was difficult to balance between implementing

rigorous exclusion criteria in order to have clean, normally distributed data vs retaining as much data as possible in order to naturally reflect performance. Either one of these options has its limitations, and we are not alone in debating where to draw the line in excluding data (for review, see Baayen & Milin, 2010). In hindsight, some of these decisions could have been made before data collection in our preregistration. This would have allowed for more objective decision making, as we could have determined the exclusion criteria before viewing our data. This realisation also points to the importance of collecting a large enough amount of pilot data. All of the experiments reported in this thesis were piloted in order to ensure that the task would run smoothly, and to give a basis from which to build scripts for data analysis ahead of data collection. However, collecting a larger amount of pilot data could have better informed our analytical decisions, such as the decisions regarding exclusion thresholds described above.

Another example of deviation from preregistration is the analysis reported in Chapter 5. We originally preregistered our analysis to use linear mixed effect models with predictor variables entered continuously. Despite the preregistration plans, we faced indecision when deciding which analytical technique was most appropriate to report in the main body of the chapter. An ANOVA had the appeal of remaining consistent with Hoffman and Woollams, who used a 2x2 ANOVA to explore the effects of semantic diversity and imageability on processing tasks. Further, Hsiao et al. (2020) initially failed to replicate the Hoffman and Woollams finding of a facilitatory effect of semantic diversity on lexical decision response times using linear mixed effect models, but they did replicate the results when using analysis of variance and entering variables categorically, as per Hoffman and Woollams' own observations. Following Hoffman and Woollams' original analysis therefore seemed a legitimate approach as it allowed a clear comparison to previous literature. On the other hand, ANOVA might not be the most suitable method for analysing the results of Chapter 5. Whelon (2008) argued that comparisons of means might not be powerful enough to detect small differences in response time data. Rather, Whelon recommended analysing central tendencies to explore

differences between conditions. Barr et al. (2013) also discussed the limitations of ANOVA, and instead advocated for the use of linear mixed-effect models. Using a powerful analysis might be particularly important for Chapter 5, as we expected small effect sizes. Ultimately, we reported linear mixed effect models with categorical predictors, following Hsiao et al.'s (2020) successful replication of Hoffman and Woollams' results. This approach should have been preregistered. For transparency and to remain consistent with the preregistration, we also reported analysis using linear mixed effects models where predictors were entered first continuously; and analysis of variance in Appendix 12.

In hindsight, the preregistrations made for these experiments were not good enough as they did not prevent me making post-hoc changes to data analysis. However, Nosek et al. (2019) presented preregistration as a process that is improved with practice. My first attempts at preregistration have been a learning experience and provide a basis upon which to improve. While we diverted from our original plans, these diversions are made transparent through published preregistration, which is an important aspect of open science practice.

REFLECTIONS ON ANALYTICAL METHODS

Throughout the thesis, linear mixed effect models were used to analyse the effects of item-level variables on lexical processing. Linear mixed effect models are typically thought to be more robust tests than ANOVA as they are more powerful (Barr et al., 2013) and are better at accounting for missing data (Baayan, 2010b). This is relevant for the analysis presented in this thesis, where cut-off thresholds for lexical decisions and reading times resulted in some lost data. When specifying mixed effect models, random slopes are an important factor in reducing the risk of false positives, and thus make results more likely to reflect true effects (Barr et al., 2013). However, problems can

arise when models are overparameterized and fail to converge (Matuschek et al, 2017). By comparison, ANOVA does include slopes.

In this thesis, many of the models failed to converge when random slopes were included; or identified a singular fit. Barr et al. (2013) discussed why models fail to converge. They outlined small samples of both item and participant, and model misspecifications, where mixed effects models may have either left out important variables or included irrelevant variables. The convergence issues in my analysis, therefore, might be due to underpowered samples of words and of participants. This has been discussed as a limitation of my empirical work in Chapters 4 and 5. For the secondary data analysis, however, it is more likely that the number of observations is sufficient. This could, then, point to a case of model misspecification, whereby the model does not account for everything it should.

As a result of convergence issues, a large part of the analysis reported included random intercepts instead of slopes. However, this is not an ideal solution. Schielzeth and Forstmeier (2009) argued that, while replacing random slopes with random intercepts is a common approach, this can lead to overconfident estimates and biased significance levels.

Barr et al. (2013) made suggestions for fixing convergence issues. These included analytical approaches, such as simplifying the model by removing interactions, centring and sum-coding predictor variables, and increasing the number of iterations in the estimation procedure. Schielzeth and Forstmeier (2009) suggested plotting data to identify how much variation exists between individuals, as this can help to determine the extent to which including random intercepts would be appropriate. Cases of model misspecification can also be approached theoretically, by questioning whether the hypothesis proposed is the most appropriate one, and whether the model is the best test of the hypothesis. Replications of the work presented in this thesis should take steps to avoid issues of convergence, in order to allow for slopes in analysis. Appropriate sample sizes should be

identified by running sufficient pilot studies or simulations. Future studies should plan how data will be handled to improve likelihood of convergence following the guidance outlined by authors such as Barr et al. (2013). If convergence issues still occur, this might suggest a problem of model misspecification. In this case, the hypotheses and models should be revisited, as these might not be appropriate in explaining the data.

REFLECTIONS ON ONLINE TESTING

All of the experimental data reported in this thesis were collected online, using Gorilla to create experiments, and Prolific to recruit participants. Online studies allow researchers to reach a larger sample of participants (Adjerid & Kelley, 2018) whilst saving researchers time and costs associated with participant testing and reimbursement (Lefever et al., 2004; Palan & Schitter, 2018). Evidence suggests that participants recruited via online participant pools score similarly on tasks compared with community-based samples (Goodman et al., 2013). Online participant pools also allow for a variable and diverse sample. In-person studies typically use undergraduate students, which are likely not representative of the broader population (Henrich et al., 2010). This is particularly important for language studies, as there is a known relationship between socio-economic status and language measures like vocabulary (Hua & Wang, 2021). By comparison, online testing allows for a larger sample which can be selected based on a rich variable set (Adjerid & Kelley, 2018). Casler et al. (2013) found that participants identified via online recruitment sites were more socio-economically and ethnically diverse than participants recruited via social media posts or undergraduate student samples. However, online tests also have their limitations. Large proportions of data can be lost due to technical difficulties (Granello & Wheaton, 2004) or due to participants not completing the tasks fully (Cantrell & Lapinacci, 2007). The validity of online data has also been

questioned, with Newman et al. (2021) citing inattentiveness and dishonest responses as two ways that data quality might be compromised.

Across this thesis, there were different motivations for using online data collection. Chapter 3 employed an online experimental design and recruitment method to collate ratings for words from a large sample of participants. Previous norming studies have also used online participant pools (e.g., Kuperman et al., 2012). This allowed us to gather data cost-effectively with participants from a sample broader than an undergraduate population. The decision to collect data online in Chapters 4 and 5 was in large part due to the COVID-19 pandemic, which made collecting data in person impractical. Online data collection therefore allowed us to minimise the impact social distancing regulations had on conducting research. However, this also brought with it methodological constraints. For example, testing online motivated us to collect all data within one session, therefore limiting the scope of how we could investigate the updating of contextual histories. This led us towards priming paradigms to test the short-term effects of recent experience rather than measuring how contextual histories are accumulated across time. In hindsight, word learning paradigms might have allowed us to test how linguistic information is accumulated more directly. This would give opportunity to study the 'updating' of word representations as described by Jones et al.'s (2012) Semantic Distinctiveness Model. Such paradigms are not impossible to conduct online, and the online testing of this thesis has provided a valuable experience that would allow me to consider this in the future.

Chapter 4 used a self-paced reading task to test reading efficiency for individual words as they appeared, one at a time. Van der Schoot et al. (2012) found that self-paced reading behaviour patterned similarly to data collected from eye fixations during more naturalistic sentence reading. Eye-tracking allows for the dynamic study of reading behaviours in real time and provides detailed data which can be used to reflect cognitive processes (Joseph et al., 2013). That self-paced reading

data patterns similarly to eye fixations therefore suggests that this is a valid reflection of processing efficiency. However, van der Schoot et al. (2012) measured whole reading times at the sentence level. At the word level, self-paced reading has its limitations. As Jegerski (2013) discussed, when words are presented on the screen individually, this is less like natural reading, in which individuals do not always process information in a linear fashion. While eye-tracking techniques can capture the regressions participants make between words (Jarodzka & Brand-Gruwel, 2017), this information is not measured by self-paced reading times. There are also different ways of presenting words on screen. Our study presented each word centred on-screen. However, Jegerski (2013) argued that presenting words in a moving window is preferable as this better reflects natural reading. In a moving window design, a sentence is positioned on the screen with all but one word masked. As participants press a key, this unmasks the subsequent word while masking the prior. The effect is that participants read words one at a time whilst the structure of the task is kept as similar as possible to natural reading. In hindsight, this would have been a better design to employ in Chapter 4. Future research could directly compare online and in-lab self-paced reading performance, and could compare this to eye-tracking. This would allow for evaluation of the extent to which online self-paced reading tasks reliably reflect reading behaviour. A hope for the future is that eye movements might be measured online, facilitating the largescale collection of data for suitably powered experiments.

Further, this thesis mostly measured lexical behaviours through lexical and semantic decision performance. These decision tasks provide a useful measure of basic lexical processing, and largescale participant data are made available by projects such as the BLP (Keuleers et al., 2012) and ELP (Balota et al., 2007). This allowed us to test hypotheses using largescale secondary data, reducing costs associated with data collection. However, it is important to extend beyond these types of task. As discussed in Chapter 5, processing single words might be very different to reading words in contexts. This is particularly relevant when studying contextual histories, as metrics like

semantic diversity also capture polysemy; and when words have multiple meanings or senses, the current context is important to interpreting that word's meaning (also see Rodd et al., 2016). A starting point could be using secondary data that have considered other types of lexical performance. For example, megadata are available for word naming latencies (e.g., Balota et al., 2007; Cortese et al., 2018).

While this thesis has made use of online testing, its results warrant investigation through other, more precise measures. Extending beyond online testing, and beyond lexical decision tasks, would allow for a more thorough exploration of contextual histories and their relationship with lexical processing. This would be particularly interesting for sentence availability, as this is a new metric which has not yet been tested outside the context of this thesis.

FINAL THOUGHTS

Semantic knowledge is a complex and multi-dimensional phenomenon that accumulates via linguistic contexts and lived experience. Lexical variables are useful tools to capture by proxy these semantic constructs and appear to capture the history of linguistic or experiential knowledge attached to the word. These variables are often closely related. In order to have meaningful lexical measures, it is important to both consider the theoretical conceptualisations of these variables, and the way these variables are calculated.

As well as long-term experience as indexed by lexical variables, recent experience is also important to how words are processed. From the priming studies reported in Chapters 4 and 5, the effects of recent experience appear to depend on both the lexical properties of the word and the task in hand. While this thesis contributes to our understanding of contextual histories and their influence on lexical processing, it has also presented areas of research that need further

investigation. Experiments are needed to establish the relationship between sentence availability and processing beyond lexical decision.

Preregistered replications of experiments are necessary to determine how and when recent experience can update the effects of longer-term knowledge (e.g., a word's semantic diversity) on lexical processing. This will allow for testing of the interpretations and ideas set out in this thesis. This would therefore improve our understanding of how words are learned and retrieved from memory during the course of processing, and what the consequences of these processing encounters are when words are experienced in the future.

BIBLIOGRAPHY

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

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*, 899–917. https://doi.org/10.1037/amp0000190

Adorni, R., & Proverbio, A. M. (2012). The neural manifestation of the word concreteness effect: An electrical neuroimaging study. *Neuropsychologia*, *50*(5), 880-891.

Allen, M., & Badecker, W. (2002). Inflectional regularity: Probing the nature of lexical representation in a cross-modal priming task. *Journal of Memory and Language, 46*(4), 705-722.

Altarriba, J., Bauer, L. M., & Benvenuto, C. (1999). Concreteness, context availability, and imageability ratings and word associations for abstract, concrete, and emotion words. *Behavior Research Methods, Instruments, & Computers*, *31*(4), 578-602.

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *Bmj*, *332*(7549), 1080.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703.

Ando, R. K. (2000). Latent semantic space: Iterative scaling improves precision of inter-document similarity measurement. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 216-223).

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2018). Gorillas in our Midst: Gorilla. sc, a new web-based Experiment Builder. *bioRxiv*, 438242.

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 30, No. 30).

Baayen, R. H. (2010). Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon*, *5*(3), 436-461.

Baayen, R. H. (2010). languageR: Data sets and functions with" Analyzing Linguistic Data: A practical introduction to statistics". *http://cran. r-project. org/package= languageR*.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12-28.

Baayen, R. H., Milin, P., Đurđević, D. F., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, *118*(3), 438.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). The CELEX lexical database (cd-rom).

Bainbridge, J. V., Lewandowsky, S., & Kirsner, K. (1993). Context effects in repetition priming are sense effects. *Memory & Cognition*, *21*(5), 619-626.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Balota, D. A., Ferraro, F. R., & Connor, L. T. (1991). On the early influence of meaning in word recognition: A review of the literature. *The Psychology of Word Meanings*, 187-222.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(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), 255-278.

Barber, H. A., Otten, L. J., Kousta, S. T., & Vigliocco, G. (2013). Concreteness in word processing: ERP and behavioral effects in a lexical decision task. *Brain and Language*, *125*(1), 47-53.

Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R package version, 2*(1), 74. https://doi.org/10.18637/jss.v067.i01

Bates, E., Burani, C., D'Amico, S., & Barca, L. (2001). Word reading and picture naming in Italian. *Memory & Cognition*, *29*(7), 986-999.

Beekhuizen, B., Armstrong, B. C., & Stevenson, S. (2021). Probing lexical ambiguity: Word vectors encode number and relatedness of senses. *Cognitive Science*, *45*(5), e12943.

Betts, H. N., Gilbert, R. A., Cai, Z. G., Okedara, Z. B., & Rodd, J. M. (2018). Retuning of lexical-semantic representations: Repetition and spacing effects in word-meaning priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(7), 1130.

BNC Consortium. (2007). British national corpus. *Oxford Text Archive Core Collection*.

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition, 104*(3), 437-458.

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*(2), 122-159.

Bolker, B. M. (2015). Linear and generalized linear mixed models. *Ecological Statistics: Contemporary Theory and Application*, 309-333.

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2001). Age-of acquisition and word frequency in the lexical decision task: Further evidence from the French language. *Current Psychology of Cognition*, *20*(6), 401-444.

Branigan, H. P., Pickering, M. J., McLean, J. F., & Stewart, A. J. (2006). The role of local and global syntactic structure in language production: Evidence from syntactic priming. *Language and Cognitive Processes, 21*(7-8), 974-1010.

Brezina, V., & Timperley, M. (2017, September). How large is the BNC? a proposal for standardised tokenization and word counting. In *Proceedings of Corpus Linguistics*.

Brysbaert, M. (1996). Word frequency affects naming latency in Dutch when age of acquisition is controlled. *European Journal of Cognitive Psychology*, *8*(2), 185-194.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in German. *Experimental psychology*, *58*(5), 412.

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45-50.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467-479.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904-911.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*(2), 272-277.

Butler, B., & Hains, S. (1979). Individual differences in word recognition latency. *Memory & Cognition*, *7*(2), 68-76.

Butler, L. T., & Berry, D. C. (2004). Understanding the relationship between repetition priming and mere exposure. *British Journal of Psychology*, *95*(4), 467-487.

Caldwell-Harris, C. L. (2021). Frequency effects in reading are powerful–But is contextual diversity the more important variable?. *Language and Linguistics Compass*, *15*(12), e12444.

Cantrell, M. A., & Lupinacci, P. (2007). Methodological issues in online data collection. *Journal of Advanced Nursing*, *60*(5), 544-549.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*(6), 2156-2160.

Cervetti, G. N., Wright, T. S., & Hwang, H. (2016). Conceptual coherence, comprehension, and vocabulary acquisition: A knowledge effect?. *Reading and Writing*, *29*(4), 761-779.

Cevoli, B., Watkins, C., & Rastle, K. (2021). What is semantic diversity and why does it facilitate visual word recognition?. *Behavior Research Methods*, *53*(1), 247-263.

Chapman, C. A., & Martin, R. C. (2021). Effects of semantic diversity and word frequency on single word processing. *Journal of Experimental Psychology: General*.

Christensen, R. H. (2019). Ordinal-Regression Models for Ordinal Data. R package version 2019.4-25. *Retrieved May, 1*, 2019.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal Learning and Verbal Behavior*, *8*(2), 240-247.

Colombo, L., & Burani, C. (2002). The influence of age of acquisition, root frequency, and context availability in processing nouns and verbs. *Brain and Language*, *81*(1-3), 398-411.

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 384-387.

Cortese, M. J., Yates, M., Schock, J., & Vilks, L. (2018). Examining word processing via a megastudy of conditional reading aloud. *Quarterly Journal of Experimental Psychology*, *71*(11), 2295-2313.

Dawson, N., Hsiao, Y., Banerji, N., Tan, A. W. M., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research, 1*(1), 9-53.

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research, 23*(1), 9-36.

Fieder, N., Wartenburger, I., & Abdel Rahman, R. (2019). A close call: Interference from semantic neighbourhood density and similarity in language production. *Memory & Cognition*, *47*(1), 145-168.

Forbach, G. B., Stanners, R. F., & Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory & Cognition*, *2*(2), 337-339.

Gimenes, M., & New, B. (2016). Worldlex: Twitter and blog word frequencies for 66 languages. *Behavior Research Methods*, *48*(3), 963-972.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224.

Gottlob, L. R., Goldinger, S. D., Stone, G. O., & Van Orden, G. C. (1999). Reading homographs: orthographic, phonologic, and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(2), 561.

Grainger, J. (1990). Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, *29*(2), 228-244.

Granello, D. H., & Wheaton, J. E. (2004). Online data collection: Strategies for research. *Journal of Counselling & Development*, *82*(4), 387-393.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14-23.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1-12.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2-3), 61-83

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259-273. https://doi.org/10.1016/j.jml.2010.06.002

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition?. *Psychological Science*, *20*(6), 729-739.

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1331.

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.

Hoffman, P., Meteyard, L., & Patterson, K. (2014). Broadly speaking: vocabulary in semantic dementia shifts towards general, semantically diverse words. *Cortex*, *55*, 30-42.

Hoffman, P., Ralph, M. L., & Rogers, T. (2021). Semantic diversity is best measured with unscaled vectors: Reply to Cevoli, Watkins and Rastle (2020).

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.

Hopman, E. W., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science, 29*(6), 961-971.

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*, 114-126.

Hsiao, Y., Bird, M., Norris, H., Pagán, A., & Nation, K. (2020). The influence of item-level contextual history on lexical and semantic judgments by children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(12), 2367.

Hua, M., & Wang, L. (2021). Effects of socioeconomic status on children's vocabulary brain development. *Aggression and Violent Behavior*, 101670.

Hudson, P. T., & Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language*, *24*(1), 46-58.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*(3), 306.

James, L. E., & Burke, D. M. (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1378.

Jarodzka, H., & Brand-Gruwel, S. (2017). Tracking the reading eye: Towards a model of real-world reading. *Journal of Computer Assisted Learning*, *33*(3), 193-201.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*(2), 278-305.

Jegerski, J. (2013). Self-paced reading. In *Research Methods in Second Language Psycholinguistics* (pp. 36-65). Routledge.

Jegerski, J. (2018). Sentence Processing in Spanish as a Heritage Language: A Self-Paced Reading Study of Relative Clause Attachment. *Language Learning*, *68*(3), 598-634.

Johns, B. T. (2021). Disentangling contextual diversity: Communicative need as a lexical organizer. *Psychological Review*, *128*(3), 525.

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *23*(4), 1214-1220.

Johns, B. T., Dye, M., & Jones, M. N. (2020). Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, *73*(6), 841-855.

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.

Johns, B. T., & Jones, M. N. (2022). Content matters: Measures of contextual diversity must consider semantic content. *Journal of Memory and Language*, *123*. https://doi.org/10.1016/j.jml.2021.104313

Johns, B. T., Sheppard, C. L., Jones, M. N., & Taler, V. (2016). The role of semantic diversity in word recognition across aging and bilingualism. *Frontiers in Psychology*, *7*, 703.

Johnston, R. A., & Barry, C. (2006). Age of acquisition and lexical processing. *Visual Cognition*, *13*(7-8), 789-845.

Jones, M. N., Johns, B. T., & Recchia, G. (2012). The role of semantic diversity in lexical organization. *Canadian Journal of Experimental Psychology, 66*(2), 115.

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, *166*, 190-211.

Joseph, H. S., Nation, K., & Liversedge, S. P. (2013). Using eye movements to investigate word frequency effects in children's sentence reading. *School Psychology Review*, *42*(2), 207-222.

Joseph, H., Wonnacott, E., Forbes, P., & Nation, K. (2014). Becoming a written word: Eye movements reveal order of acquisition effects following incidental exposure to new words during silent reading. *Cognition*, *133*(1), 238-248.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*(4), 474-516.

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, *37*(1), 1-32.

Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(4), 601.

Kensinger, E. A., & Corkin, S. (2003). Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words?. *Memory & Cognition*, *31*(8), 1169-1180.

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

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.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978-990. https://doi.org/10.3758/s13428-012-0210-4

Kuperman, V., & Van Dyke, J. A. (2013). Reassessing word frequency as a determinant of word recognition for skilled and unskilled readers. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(3), 802.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Package 'lmertest'. *Journal of Statistical Software, 82*(13), 1-26. https://doi.org /10.18637/jss.v082.i13

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

Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Lefever, S., Dal, M., & Matthíasdóttir, Á. (2007). Online data collection in academic research: advantages and limitations. *British Journal of Educational Technology*, *38*(4), 574-582.

Lester, N., Feldman, L., & del Prado Martín, F. M. (2017). *You can take a noun out of syntax...: Syntactic similarity effects in lexical priming.* Paper presented at the CogSci.

Loebell, H., & Bock, K. (2003). Structural priming across languages. *Linguistics, 41*(5), 791-824.

Lohnas, L. J., Polyn, S. M., & Kahana, M. J. (2011). Contextual variability in free recall. *Journal of Memory and Language*, *64*(3), 249-255.

Lovelace, E. A. (1988). On using norms for low-frequency words. *Bulletin of the Psychonomic Society*, *26*(5), 410-412.

Lowder, M. W., & Gordon, P. C. (2017). Print exposure modulates the effects of repetition priming during sentence reading. *Psychonomic Bulletin & Review*, *24*(6), 1935-1942.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization* (No. 8). Oxford University Press on Demand.

Luke, S. G., & Christianson, K. (2013). SPaM: A combined self-paced reading and masked-priming paradigm. *Behavior Research Methods, 45*(1), 143-150.

Luzzatti, C., Raggi, R., Zonca, G., Pistarini, C., Contardi, A., & Pinna, G. D. (2002). Verb–noun double dissociation in aphasic lexical impairments: The role of word frequency and imageability. *Brain and Language*, *81*(1-3), 432-444.

MacDonald, M. C. (2013). How Language Production Shapes Language Form and Comprehension. *Frontiers in Psychology, 4*, 226.

MacLeod, C. M. (1989). Word context during initial exposure influences degree of priming in word fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(3), 398.

Mak, M. H., Hsiao, Y., & Nation, K. (2021). Anchoring and contextual variation in the early stages of incidental word learning during reading. *Journal of Memory and Language*, *118*, 104203.

Mak, M. H., & Twitchell, H. (2020). Evidence for preferential attachment: Words that are more well connected in semantic networks are better at acquiring new links in paired-associate learning. *Psychonomic Bulletin & Review*, *27*(5), 1059-1069.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*, 25-45.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305-315.

Mayr, S., & Buchner, A. (2007). Negative priming as a memory phenomenon. *Zeitschrift für Psychologie/Journal of Psychology, 215*(1), 35-51.

McCarron, S. (2022). *Author recognition test.* Unpublished manuscript. University of Oxford.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech*, *44*(3), 295-322.

McKone, E., & Dennis, C. (2000). Short-term implicit memory: Visual, auditory, and cross-modality priming. *Psychonomic Bulletin & Review*, *7*(2), 341-346.

McNamara, T. P., & Healy, A. F. (1988). Semantic, phonological, and mediated priming in reading and lexical decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 398.

Melinger, A., & Koenig, J. P. (2007). Part-of-speech persistence: The influence of part-of-speech information on lexical processes. *Journal of Memory and Language*, *56*(4), 472-489.

Mitchell, D. C., Kieras, D. E., & Just, M. A. (1984). New methods in reading comprehension research. *Chap. An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading, pp. 69–89.* Hillsdale, New Jersey, USA: Laurence Erlbaum Associates.

Moffat, M., Siakaluk, P. D., Sidhu, D. M., & Pexman, P. M. (2015). Situated conceptualization and semantic processing: effects of emotional experience and context availability in semantic categorization and naming tasks. *Psychonomic Bulletin & Review, 22*(2), 408-419. https://doi.org/10.3758/s13423-014-0696-0

Monsell, S. (2012). The nature and locus of word frequency effects in reading. In *Basic Processes in Reading* (pp. 156-205). Routledge.

Monsell, S., Doyle, M. C., & Haggard, P. N. (1989). Effects of frequency on visual word recognition tasks: Where are they?. *Journal of Experimental Psychology: General*, *118*(1), 43.

Muraki, E. J., Abdalla, S., Brysbaert, M., & Pexman, P. M. (2022). Concreteness ratings for 62,000 English multiword expressions. *Behavior Research Methods*, 1-10.

Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill. *npj Science of Learning*, *2*(1), 1-4.

Newman, A., Bavik, Y. L., Mount, M., & Shao, B. (2021). Data collection via online platforms: Challenges and recommendations for future research. *Applied Psychology*, *70*(3), 1380-1402.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., ... & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815-818.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600-2606.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673-690.

Oltra-Massuet, I., Sharpe, V., Neophytou, K., & Marantz, A. (2017). Syntactic priming as a test of argument structure: A self-paced reading experiment. *Frontiers in Psychology, 8*, 1311.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Özdemir, R., Roelofs, A., & Levelt, W. J. (2007). Perceptual uniqueness point effects in monitoring internal speech. *Cognition*, *105*(2), 457-465.

Pagán, A., Bird, M., Hsiao, Y., & Nation, K. (2020). Both semantic diversity and frequency influence children's sentence reading. *Scientific Studies of Reading*, *24*(4), 356-364.

Pagán, A., & Nation, K. (2019). Learning words via reading: Contextual diversity, spacing, and retrieval effects in adults. *Cognitive Science*, *43*(1), e12705.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17*(C), 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Parkin, A. J. (1982). Phonological recoding in lexical decision: Effects of spelling-to-sound regularity depend on how regularity is defined. *Memory & Cognition*, *10*(1), 43-53.

Perea, M., Soares, A. P., & Comesaña, M. (2013). Contextual diversity is a main determinant of word identification times in young readers. *Journal of Experimental Child Psychology*, *116*(1), 37-44.

Pexman, P. M. (2020). How does meaning come to mind? Four broad principles of semantic processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 74*(4), 275. https://doi.org/10.1037/cep0000235

Pexman, P. M., Hargreaves, I. S., Edwards, J. D., Henry, L. C., & Goodyear, B. G. (2007). The neural consequences of semantic richness. *Psychological Science*, *18*(5), 401-406.

Pexman, P. M., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161-167.

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.

Pexman, P. M., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale, 53*(4), 323–334.

Pexman, P. M., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*(3), 542-549.

Pexman, P. M., Muraki, E., Sidhu, D. M., Siakaluk, P. D., & Yap, M. J. (2019). Quantifying sensorimotor experience: Body–object interaction ratings for more than 9,000 English words. *Behavior Research Methods*, *51*(2), 453-466.

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition*, *28*(4), 657-666.

Plummer, P., Perea, M., & Rayner, K. (2014). The influence of contextual diversity on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 275.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ratcliff, R., & McKoon, G. (1988). A retrieval theory of priming in memory. *Psychological Review*, *95*(3), 385.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences*, *26*(6), 462-483.

Rayner, K., & Raney, G. E. (1996). Eye movement control in reading and visual search: Effects of word frequency. *Psychonomic Bulletin & Review*, *3*(2), 245-248.

Revelle, W. R. (2017). psych: Procedures for personality and psychological research. Northwestern University, Evanston, Illinois. R package version 2.1.9, https://CRAN.R-project.org/package=psych.

Rice, C. A., Tokowicz, N., Fraundorf, S. H., & Liburd, T. L. (2019). The complex interactions of context availability, polysemy, word frequency, and orthographic variables during lexical processing. *Memory & Cognition*, *47*(7), 1297-1313.

Rodd, J. M. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science*, *15*(2), 411-427.

Rodd, J. M., Cai, Z. G., Betts, H. N., Hanby, B., Hutchinson, C., & Adler, A. (2016). The impact of recent and long-term experience on access to word meanings: Evidence from large-scale internet-based experiments. *Journal of Memory and Language, 87*, 16-37.

Rodd, J. M., Cutrin, B. L., Kirsch, H., Millar, A., & Davis, M. H. (2013). Long-term priming of the meanings of ambiguous words. *Journal of Memory and Language*, *68*(2), 180-198.

Rodd, J., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245-266.

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.

Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, *57*(3), 348-379.

Romani, C., Mcalpine, S., & Martin, R. C. (2008). Concreteness effects in different tasks: Implications for models of short-term memory. *Quarterly Journal of Experimental Psychology*, *61*(2), 292-323.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906-914.

Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PloS One*, *12*(6), e0179004.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487-494.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *The Journal of the Acoustical Society of America*, *35*(2), 200-206.

Scarborough, D. L., Cortese, C., & Scarborough, H. S. (1977). Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 1.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral ecology*, *20*(2), 416-420.

Schuberth, R. E., & Eimas, P. D. (1977). Effects of context on the classification of words and nonwords. *Journal of Experimental Psychology: Human Perception and Performance*, *3*(1), 27.

Schwanenflugel, P. J., Akin, C., & Luh, W.-M. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition, 20*(1), 96-104. https://doi.org/10.3758/BF03208259

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499-520.

Schwanenflugel, P. J., & Noyes, C. R. (1996). Context availability and the development of word reading skill. *Journal of Literacy Research, 28*(1), 35-54. https://doi.org/10.1080/10862969609547909

Sharma, S., Kim, H., Harris, H., Haberstroh, A., Wright, H. H., & Rothermich, K. (2021). Eye tracking measures for studying language comprehension deficits in aphasia: A systematic search and scoping review. *Journal of Speech, Language, and Hearing Research*, *64*(3), 1008-1022.

Siakaluk, P. D., Pexman, P. M., Aguilera, L., Owen, W. J., & Sears, C. R. (2008). Evidence for the activation of sensorimotor information during visual word recognition: The body–object interaction effect. *Cognition*, *106*(1), 433-443.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, *20*(1), 120-136.

Simpson, G. B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, *96*(2).

Stadthagen-Gonzalez, H., & Davis, C. J. (2006). The Bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, *38*(4), 598-605. https://doi.org/10.3758/BF03193891

Stark, C. E., & McClelland, J. L. (2000). Repetition priming of words, pseudowords, and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(4), 945.

Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 760.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory.

Strain, E., & Herdman, C. M. (1999). Imageability effects in word naming: an individual differences analysis. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *53*(4), 347.

Taylor, E., Nation, K., & Hsiao, Y. (2022). Context availability and sentence availability ratings for 3,000 English words and their association with lexical processing. *Journal of Cognition*, *5*(1).

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*, *52*(6), 2372-2382.

Tillotson, S. M., Siakaluk, P. D., & Pexman, P. M. (2008). Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075-1078.

Van der Schoot, M., Reijntjes, A., & van Lieshout, E. C. (2012). How do children deal with inconsistencies in text? An eye fixation and self-paced reading study in good and poor reading comprehenders. *Reading and Writing*, *25*(7), 1665-1690.

Van Hell, J., & De Groot, A. M. (1998). Disentangling context availability and concreteness in lexical decision and word translation. *The Quarterly Journal of Experimental Psychology: Section A, 51*(1), 41-63. https://doi.org/10.1080/713755752

Van Heuven, W. J., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176-1190.

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190.

Vergara-Martínez, M., Comesaña, M., & Perea, M. (2017). The ERP signature of the contextual diversity effect in visual word recognition. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(3), 461-474.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219-247.

Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(5), 1256–1271.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191-1207. https://doi.org/10.3758/s13428-012-0314-x

Waszak, F., Hommel, B., & Allport, A. (2005). Interaction of task readiness and automatic retrieval in task switching: Negative priming and competitor priming. *Memory & Cognition, 33*(4), 595-610.

Whelan, R. (2008). Effective analysis of reaction time data. *The Psychological Record*, *58*(3), 475-482.

Winhuisen, L., Thiel, A., Schumacher, B., Kessler, J., Rudolf, J., Haupt, W. F., & Heiss, W. D. (2005). Role of the contralateral inferior frontal gyrus in recovery of language function in poststroke aphasia: a combined repetitive transcranial magnetic stimulation and positron emission tomography study. *Stroke*, *36*(8), 1759-1763.

Yap, M. J., Lim, G. Y., & Pexman, P. M. (2015). Semantic richness effects in lexical decision: The role of feedback. *Memory & Cognition*, *43*(8), 1148-1167. https://doi.org/10.3758/s13421-015-0536-0

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review*, *18*(4), 742-750.

Zdrazilova, L., & Pexman, P. M. (2013). Grasping the invisible: Semantic processing of abstract words. *Psychonomic Bulletin & Review*, *20*(6), 1312-1318.

Zeno, S., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The Educator's Word Frequency Guide*. Touchstone Applied Science Associates.

APPENDIX 1

PREDICTING LEXICAL DECISION WITH DIFFERENT SEMANTIC DIVERSITY MEASURES

The graphs in this appendix relate to the results reported in Chapter 1, testing the following model:

Lexical decision ~ Frequency + AoA + Concreteness + SemD + (1|Word) + (1|Participant)

Tables 1 and 2 use the semantic diversity as calculated by Hoffman et al. (2013). Tables 3 and 4 show the semantic diversity measure calculated by Cevoli et al. (2021). Tables 5 and 6 show the measure of Johns et al. (2020), tables 7 and 8 show the measure of Johns (2021), and Tables 9 and 10 show the model fitted with document count from Hoffman et al. (2013) instead of semantic diversity.

Hoffman's semantic diversity

Table 1: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of BLP lexical decision with Hoffman's measure.

	Estimate	Std Error	Z value	P value
Intercept	1.11	0.27	4.13	<.001
Frequency	0.84	0.03	25.29	<.001
AoA	-0.18	0.01	-18.75	<.001
Concreteness	0.05	0.02	2.23	<.05
SemD (Hoffman)	0.46	0.07	6.62	<.001

	Estimate	Std Error	Df	T value	P value
Intercept	664.49	13.22	61.23	50.27	<.001
Frequency	-31.39	0.84	7489.90	-37.21	<.001
AoA	6.35	0.27	7582.94	23.75	<.001
Concreteness	-3.15	0.62	7454.85	-5.11	<.001
SemD (Hoffman)	-9.00	2.03	7737.87	-4.44	<.001

Table 2: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of ELP lexical decision with Hoffman's measure.

	Estimate	Std Error	Z value	P value
Intercept	1.75	0.35	4.94	<.001
Frequency	0.50	0.04	11.57	<.001
AoA	-0.20	0.01	-14.32	<.001
Concreteness	0.15	0.03	4.58	<.001
SemD (Hoffman)	0.28	0.11	2.65	<.01

	Estimate	Std Error	Df	T value	P value
Intercept	758.50	18.86	2730.40	40.23	<.001
Frequency	-33.00	2.22	2454.47	-14.87	<.001
AoA	10.25	0.73	2448.19	14.01	<.001
Concreteness	-6.57	1.67	2448.00	-3.94	<.001
SemD (Hoffman)	-8.25	5.76	2450.71	-1.43	.15

Cevoli's semantic diversity

Table 3: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of BLP lexical decision with Cevoli's measure.

	Estimate	Std Error	Z value	P value
Intercept	1.00	0.25	4.02	<.001
Frequency	0.96	0.03	31.07	<.001
AoA	-0.20	0.01	-21.91	<.001
Concreteness	-0.02	0.02	-0.96	.34
SemD (Cevoli)	0.39	0.06	6.72	<.001

	Estimate	Std Error	Df	T value	P value
Intercept	678.71	13.14	56.37	51.66	<.001
Frequency	-34.58	0.82	7806.10	-42.18	<.001
AoA	6.50	0.26	7892.56	24.88	<.001
Concreteness	-1.78	0.56	7702.08	-3.18	<.01
SemD (Cevoli)	-10.77	1.69	7857.46	-6.39	<.001

Table 4: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of ELP lexical decision with Cevoli's measure.

	Estimate	Std Error	Z value	P value
Intercept	1.51	0.33	4.54	<.001
Frequency	0.63	0.05	13.51	<.001
AoA	-0.21	0.01	-14.77	<.001
Concreteness	0.07	0.03	2.43	<.05
SemD (Cevoli)	0.26	0.09	2.78	<.01

	Estimate	Std Error	Df	T value	P value
Intercept	791.80	17.98	2677.50	44.04	<.001
Frequency	-37.55	2.40	2376.60	-15.67	<.001
AoA	10.35	0.74	2373.38	13.93	<.001
Concreteness	-4.35	1.56	2371.73	-2.79	<.01
SemD (Cevoli)	-18.42	5.06	2381.15	-3.64	<.001

Johns' (2020) vSDM

Table 5: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of BLP lexical decision with Johns' (2020) measure.

	Estimate	Std Error	Z value	P value
Intercept	-0.01	0.21	-0.06	.95
Frequency	0.87	0.03	27.06	<.001
AoA	-0.15	0.01	-19.98	<.001
Concreteness	-0.02	0.02	-1.12	.26
SD (Johns, 2020)	0.20	0.02	11.58	<.001

	Estimate	Std Error	Df	T value	P value
Intercept	738.46	13.24	50.81	55.77	<.001
Frequency	-26.91	0.95	9412.21	-28.44	<.001
AoA	4.70	0.24	9509.21	19.31	<.001
Concreteness	-2.02	0.52	9249.77	-3.91	<.001
SD (Johns, 2020)	-10.60	0.55	9537.39	-19.17	<.001

Table 6: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of ELP lexical decision with Johns' (2020) measure.

	Estimate	Std Error	Z value	P value
Intercept	-0.30	0.37	-0.82	.41
Frequency	0.21	0.05	4.30	<.001
AoA	-0.17	0.01	-12.03	<.001
Concreteness	0.12	0.03	4.23	<.001
SD (Johns, 2020)	0.38	0.03	11.65	<.001

	Estimate	Std Error	Df	T value	P value
Intercept	885.29	20.09	2695.47	44.07	<.001
Frequency	-17.10	2.44	2454.59	-7.01	<.001
AoA	8.22	0.73	2446.22	11.23	<.001
Concreteness	-5.55	1.49	2446.78	-3.73	<.001
SD (Johns, 2020)	-20.74	1.74	2447.56	-11.89	<.001

Johns' (2021) UCD-SD-PR

Table 7: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of BLP lexical decision with Johns' (2021) measure.

	Estimate	Std Error	Z value	P value
Intercept	0.97	0.20	4.77	<.001
Frequency	1.09	0.03	33.93	<.001
AoA	-0.16	0.01	-20.53	<.001
Concreteness	-0.03	0.02	-1.53	.13
SD (Johns, 2021)	0.07	0.03	2.11	<.05

	Estimate	Std Error	Df	T value	P value
Intercept	679.76	13.08	48.36	51.98	<.001
Frequency	-35.05	1.03	9774.86	-34.00	<.001
AoA	5.01	0.25	9490.74	19.89	<.001
Concreteness	-2.61	0.56	9309.61	-4.68	<.001
SD (Johns, 2021)	-6.13	0.93	9116.26	-6.62	<.001

Table 8: Fixed effects of variables for predicting accuracy (upper section) and reaction time (lower section) of ELP lexical decision with Johns' (2021) measure.

	Estimate	Std Error	Z value	P value
Intercept	3.26	0.32	10.23	<.001
Frequency	0.03	0.06	0.55	.59
AoA	-0.16	0.01	-11.46	<.001
Concreteness	0.22	0.03	7.07	<.001
SD (Johns, 2021)	0.60	0.06	10.25	<.001

	Estimate	Std Error	Df	T value	P value
Intercept	681.84	17.22	2785.69	39.60	<.001
Frequency	-4.38	3.21	2456.11	-1.36	.17
AoA	7.93	0.74	2450.13	10.73	<.001
Concreteness	-11.22	1.56	2449.68	-7.19	<.001
SD (Johns, 2021)	-34.27	2.91	2453.71	-11.77	<.001

Document count (taken from Hoffman et al., 2013)

Table 9: Fixed effects of variables for predicting accuracy (upper section) and response time (lower section) of BLP lexical decision with document count.

	Estimate	Std Error	z	p
Intercept	1.29	0.24	5.27	<.001
Frequency	1.14	0.04	32.31	<.001
AoA	-0.19	0.01	-19.75	<.001
Concreteness	-0.07	0.02	-3.47	<.001
Doc count	-0.32	0.02	-16.42	<.001

	Estimate	Std Error	Df	t	p
Intercept	676.09	12.85	54.72	52.60	<.001
Frequency	-41.45	0.91	7489.91	-45.50	<.001
AoA	6.27	0.26	7558.58	23.96	<.001
Concreteness	-0.08	0.55	7417.97	-0.14	.89
Doc count	8.36	0.48	7145.49	17.51	<.001

Table 10: Fixed effects of variables for predicting accuracy (upper section) and response time (lower section) of ELP lexical decision with document count.

	Estimate	Std Error	z	p
Intercept	1.23	0.32	3.89	<.001
Frequency	0.83	0.05	17.40	<.001
AoA	-0.19	0.01	-13.93	<.001
Concreteness	0.04	0.03	1.39	.16
Doc count	-0.38	0.03	-11.09	<.001

	Estimate	Std Error	Df	t	p
Intercept	805.40	17.21	2767.16	46.79	<.001
Frequency	-51.00	2.46	2435.76	-20.73	<.001
AoA	9.35	0.72	2431.71	13.01	<.001
Concreteness	-1.74	1.53	2429.93	-1.13	0.258
Doc count	18.98	1.68	2444.05	11.32	<.001

We also used linear regressions to calculate the proportion of variance explained in BLP lexical decision reaction time. This allowed an additional comparison of model fits by looking at the R-squared values. Here, the model remained the same but did not include random effects of item or participant. The regressions patterned similarly to the mixed-effect models. A comparison of R-squared values is shown in Table 11, including the base model (without semantic diversity) and a model including document count instead of semantic diversity.

Small differences were seen between the diversity measures; however, as all models explained a small proportion of variance, it is difficult to determine how meaningful these differences are. Of the semantic diversity measures, Johns et al. (2020) best fit BLP data, with this model explaining 7.853% of variance in response times, while Johns (2021) best fit ELP data, with the model explaining 3.335% of variance in response times. Hoffman and Cevoli's measures failed to supersede the fit of the base model for BLP data and the model including document count for both BLP and ELP data.

Table 11: Proportion of the variance in lexical decision explained by the model: Lexical decision ~ frequency + age of acquisition + concreteness + semantic diversity or document count.

Measure	BLP rt	ELP rt
Base model	7.562	3.090
Document count	6.494	3.258
Hoffman	6.239	3.093
Cevoli	6.520	3.229
Johns (2020)	7.853	3.333
Johns (2021)	7.619	3.335

Note: Base model structure was as follows: Lexical decision ~ frequency + age of acquisition + concreteness. Document count and each of the semantic diversity measures were tested on top of this model, separately. Proportions were based on adjusted R-squared values.

APPENDIX 2

Instructions for Context Availability and Sentence Availability Norms

Context Availability

Slide 1

You will be presented with a series of individual words. Your task is to rate each word for how easily or difficultly it can be associated with a context or circumstance in which it might appear.

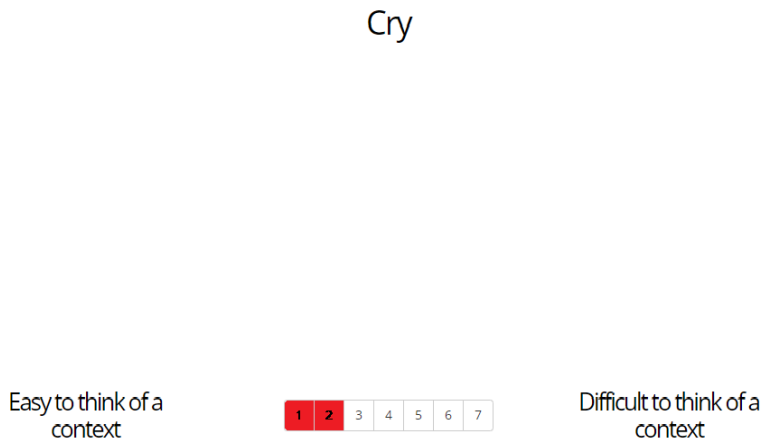
If you can easily place a word in a context/circumstance, rate this towards the lower end of the spectrum. If it is difficult to place the word in a context/circumstance, rate this towards the upper end of the spectrum.

Slide 2

For example, if you saw the word 'cry' and immediately thought of a baby starting to cry in a crib, you might rate this word a 1 or 2.

Let's see this on the next slide...

Slide 3

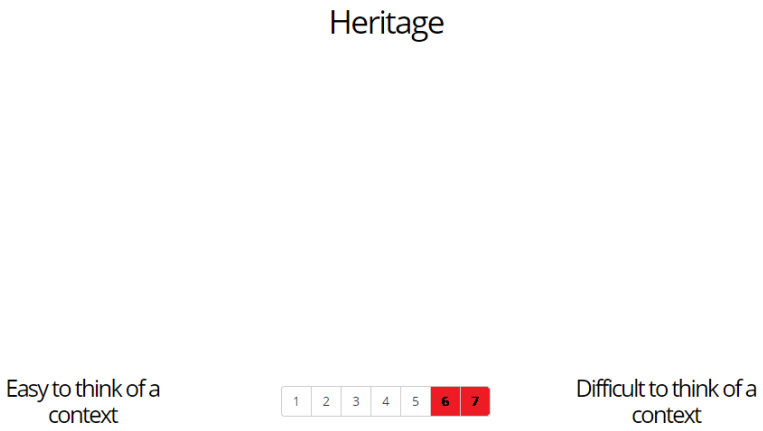


Slide 4

Or, if you saw the word ‘heritage’ and took several moments to come up with a context in which ‘heritage’ might appear, you might rate this word a 6 or 7.

Let's see this on the next slide...

Slide 5



Slide 6

You may also see some words that you don't know. If you see a word that you don't know, please select 'I don't know this word' to move onto the next item.

Sentence Availability

Slide 1

You will be presented with a series of individual words. Your task is to rate each word for how easily or difficultly it can be associated with a sentence in which it might appear.

If you can easily place a word in a sentence rate this towards the lower end of the spectrum. If it is difficult to place the word in a sentence, rate this towards the upper end of the spectrum.

Slide 2

For example, if you saw the word ‘cry’ and immediately thought of the sentence "the baby is starting to cry in the crib", you might rate this word a 1 or 2.

Let's see this on the next slide...

Slide 3

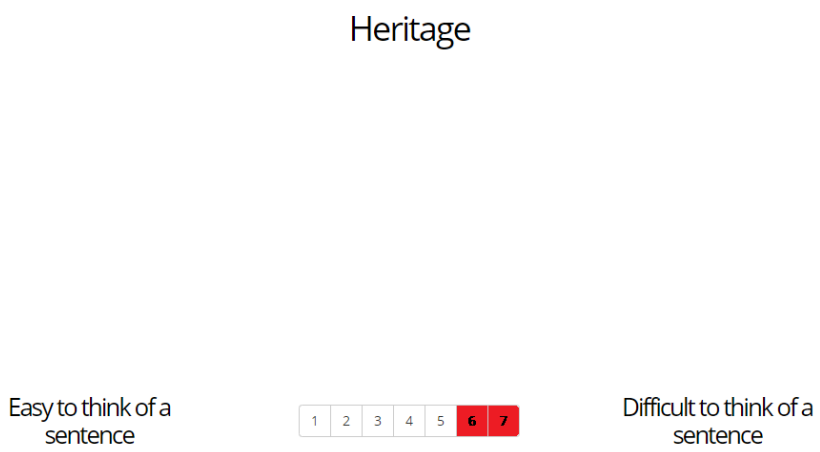


Slide 4

Or, if you saw the word ‘heritage’ and took several moments to come up with a sentence in which ‘heritage’ might appear, you might rate this word a 6 or 7.

Let's see this on the next slide...

Slide 5



Slide 6

You may also see some words that you don't know, or some words that are made up completely. If you see a word that you don't know, please select 'I don't know this word' to move onto the next item.

APPENDIX 3

A COMPARISON OF WORDS APPEARING IN BOTH THE CURRENT AND ALTARRIBA ET AL.'S (1999) NORMS

It was possible to directly compare 117 words that appeared in both our set of norms and the norms collected by Altarriba et al. (1999). Our measure had a slightly higher minimum (4.00) and mean (5.71) rating, but there were not large differences between the two measures of context availability (see Table 1).

These 117 words were then compared to Altarriba's full item-set of 325 words (Table 2). The overlapping words had slightly higher context availability ratings than the average in Altarriba's full item-set (based on Altarriba's ratings), but other than this, no large differences between these 117 words and the full set of Altarriba's norms were observed. A Pearson test gave a correlation of .645 between these values. This is lower than expected to validate the new norms, and a t-test confirmed that these values differed significantly, $t(180.38)=2.846$, $p<.05$.

Table 1: Range and mean values for Altarriba et al.'s context availability norms and our context availability norms.

	Min	Max	Mean
Altarriba CA	3.30	7.00	5.71
New CA	4.00	7.00	6.01
CA observations	15	21	18.1
SA	4.75	7.00	6.35
SA observations	15	20	17.88

Note: These values refer to only the words which appeared in both our item-set and Altarriba et al.'s item-set (N=117, with 116 items with valid ratings for sentence availability). Number of observations in our experiment are also shown.

Table 2: Descriptive statistics for the 117 words that occurred in both the current item-set and Altarriba et al.'s item-set, alongside statistics for the full item-set of 325 words collected by Altarriba et al. (1999).

	CA value	Concreteness	Imageability	SemD	Frequency
Overlapping words (N=117)	5.710	4.839	4.388	1.730	4.633
Altarriba itemset (N=325)	5.068	4.301	4.121	1.748	4.432

Note: Context availability, concreteness and imageability values are those reported by Altarriba et al.

APPENDIX 4

ANALYSIS OF CONTEXT AND SENTENCE AVAILABILITY NORMS, EXCLUDING FUNCTION WORDS

This analysis supplements our norms for context availability (CA) and sentence availability (SA). In our norms analysis, function words were rated highly for sentence availability. Associating a word with a sentence seems easy for words that are contextually flexible but not semantically rich. For example, 'or' (6.85), 'an' (6.94) and 'at' (6.40) were all rated highly by participants in our study. Here, we investigated whether function words pattern differently to non-function words.

Our analysis had two parts. First, we excluded function words and repeated the analysis reported in our norms paper. Second, we looked at the relationships between lexical variables in function words only.

Excluding function words

A list of 127 function words (shown at the end of this document) were excluded from analysis. These were taken from the stop words list commonly used in linguistic processing research. Of these words, 78 had information for CA and 83 had information for SA.

Results patterned similarly to those reported in our paper. Both availability measures were normally distributed, with a slight skew towards high ratings (Figure 1).

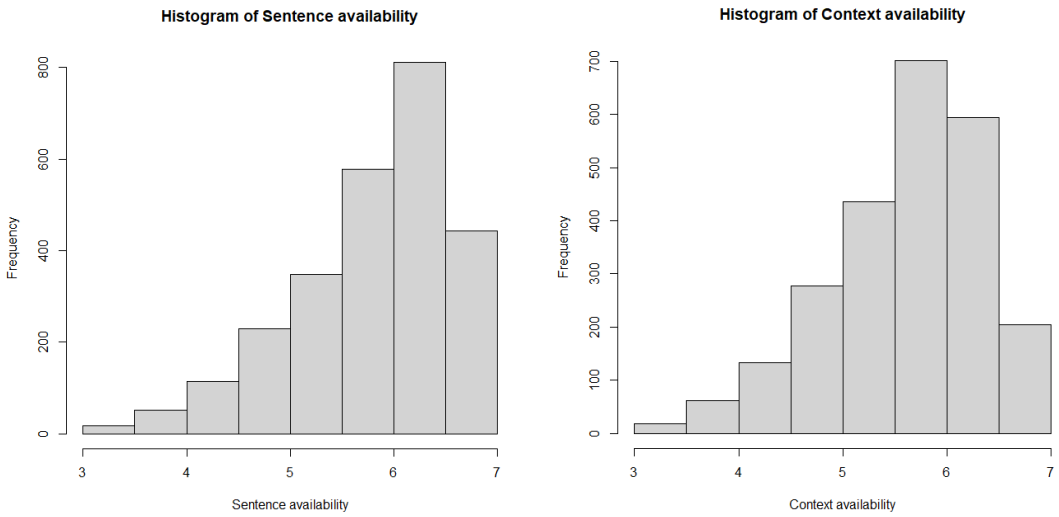


Figure 1: Histograms showing the distribution of CA (5.67 , 0.73) and SA (5.82 , 0.75).

Availability norms are plotted against each other in Figure 2, with examples labelled to illustrate the relationship between CA and SA. Figure 3 shows Pearson correlations between CA, SA and a range of other variables.

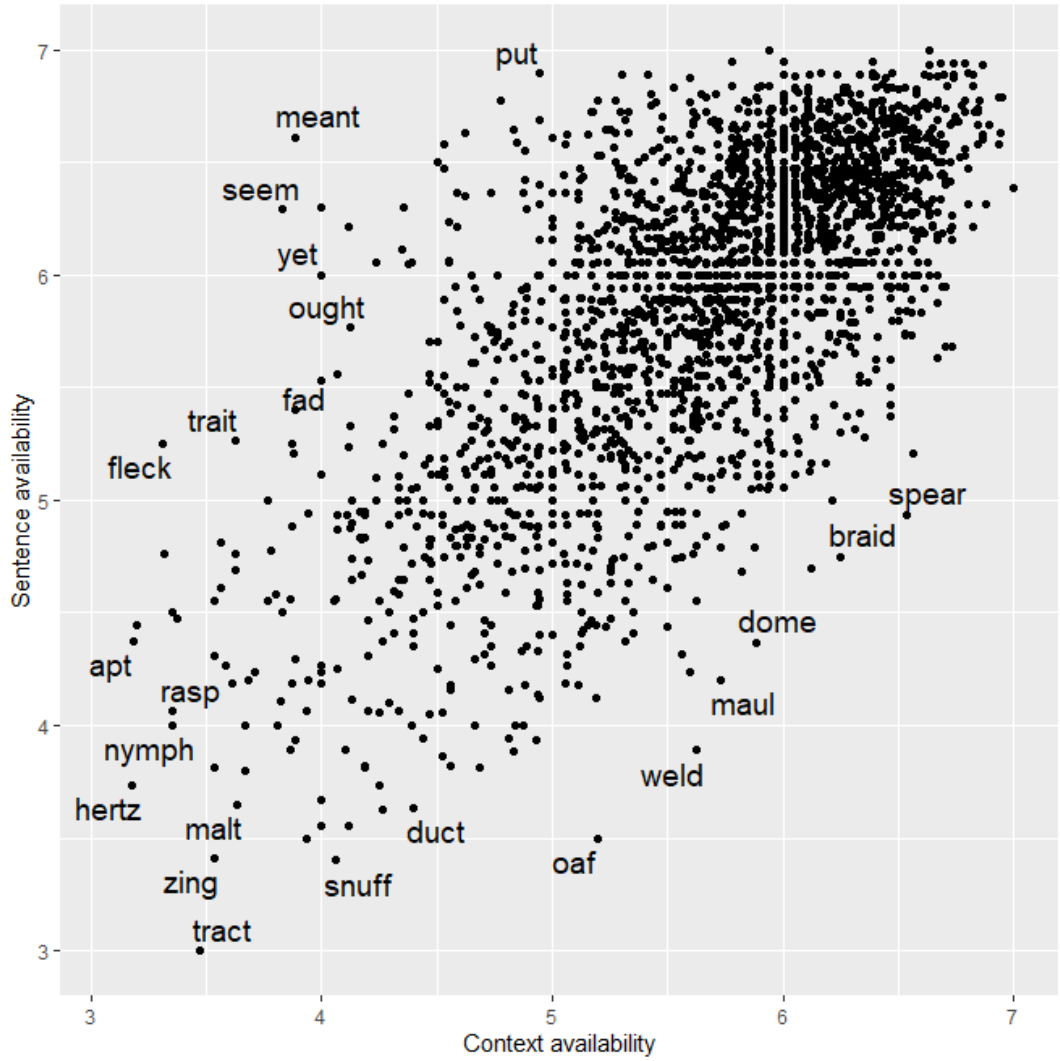


Figure 2: Scatterplot showing the relationship between context availability and sentence availability, with examples labelled.

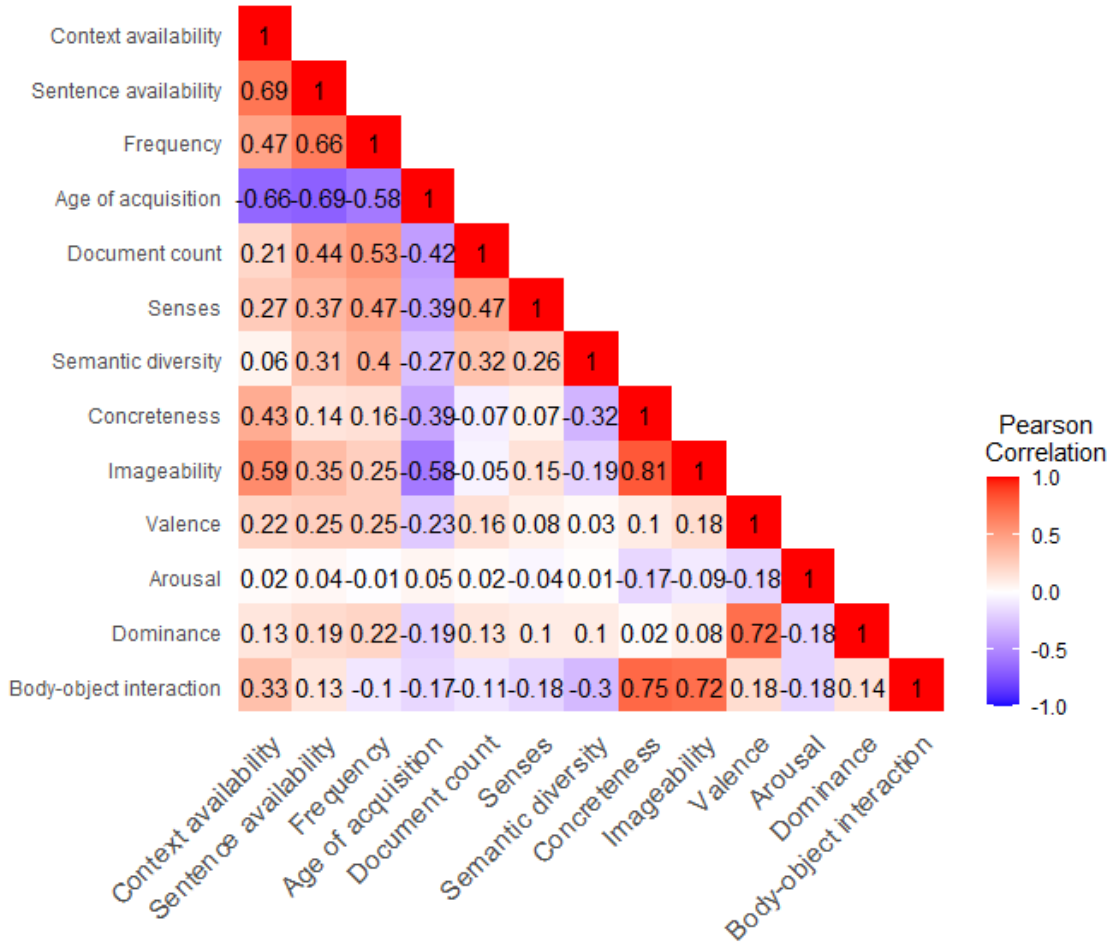


Figure 3: Pearson correlations for CA, SA and other lexical variables, based on N available for each pairwise correlation.

As in our reported analysis, three factors were identified: occurrence, richness and affect. The loadings patterned similarly to our main analysis, with context availability loading onto occurrence and richness, and sentence availability loading onto just occurrence. Loadings are shown in Table 1.

Table 1: Factor loadings for an exploratory factor analysis using oblique rotation.

	Occurrence	Richness	Affect
CA	0.58	0.41	
SA	0.80		
Zipf Freq	0.813		
AoA	-0.75		
Doc Count	0.62		
Senses	0.58		
SemD	0.47	-0.42	
Concreteness		0.89	
Imageability		0.881	
Valence			0.81
Arousal			
Dominance			0.87
BOI		0.84	

	Occurrence	Richness	Affect
Occurrence		.20	.23
Richness	.20		.12
Affect	.23	.12	

Note: Words included had values for all variables. Loadings less than 0.3 not included; lower part of table shows correlation between factors.

Next, we fitted these norms to data from the British Lexicon Project (BLP, Keuleers et al., 2012; Table 2) and English Lexicon Project (ELP, Balota et al., 2007; Table 3). Both CA and SA were significant predictors of lexical decision accuracy and reaction time from both datasets. As in our full itemset, our norms were associated with higher accuracy and shorter RTs.

Table 2: Fixed effects predicting lexical decision accuracy (upper section) and RT (lower section in BLP data.

	Estimate	SE	z	p
Intercept	2.86	0.15	19.59	<.001
Freq	0.56	0.07	8.10	<.001
AoA	0.04	0.07	0.53	.60
SemD	0.15	0.05	3.07	<.01
Imageability	0.23	0.05	4.42	<.001
CA	0.22	0.06	4.05	<.001
SA	0.33	0.06	5.62	<.001

	Estimate	SE	df	t	p
Intercept	588.89	11.33	37.76	51.99	<.001
Freq	-21.31	1.74	1987.71	-12.27	<.001
AoA	5.48	1.74	1975.67	3.16	<.01
SemD	-3.44	1.33	2026.95	-2.59	<.01
Imageability	-4.27	1.28	1970.88	-3.34	<.001
CA	-10.19	1.49	2009.63	-6.82	<.001
SA	-10.26	1.61	2025.05	-6.37	<.001

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + context availability + sentence availability + (1|participant) + (1|word).

Table 3: Fixed effects predicting lexical decision accuracy (upper section) and RT (lower section in ELP data.

	Estimate	SE	z	p
Intercept	-2.27	0.45	-4.99	<.001
Freq	0.52	0.05	10.48	<.001
AoA	-0.06	0.02	-3.69	<.001
SemD	0.66	0.11	5.99	<.001
Imageability	0.26	0.03	9.83	<.001
CA	0.09	0.06	1.49	.14
SA	0.22	0.06	3.74	<.001

	Estimate	SE	df	t	p
Intercept	973.76	25.40	2165.38	38.34	<.001
Freq	-33.45	2.55	2036.30	-13.10	<.001
AoA	3.42	0.87	2034.08	3.94	<.001
SemD	-29.08	5.94	2032.80	-4.89	<.001
Imageability	-11.31	1.36	2033.47	-8.32	<.001
CA	-4.40	3.18	2031.36	-1.41	.16
SA	-14.15	3.27	2035.79	-4.33	<.001

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + context availability + sentence availability + (1|participant) + (1|word).

The similarities between these results and those in the full dataset suggest that our reported results are not being driven by function words. However, as function words only made up a small proportion of our itemset, it is unclear from the above analysis whether these words would pattern differently to the rest of our results. We therefore explored the correlations between variables for the function words.

Function words only

As some variables had very little overlap with this itemset (for example, arousal, dominance and valence values were only available for 6 items), we did not conduct factor analysis or regression analysis on this itemset.

In function words, CA and SA were not significantly correlated, $r = .15$, $p = .197$. Function words were significantly higher in SA ($M = 6.51$) than CA ($M = 5.15$), $p < .001$. The correlations demonstrated that CA was still closely related to concreteness ($r = .70$), imageability ($r = .65$) and body-object interaction ($r = .57$). It was also negatively correlated with semantic diversity ($r = -.38$), a finding which we did not observe in the full itemset. As with non-function words, SA was closely related to frequency ($r = .56$), but here it was uncorrelated with semantic diversity ($r = .03$).

This helps us to further understand the relationship between our norms and contextual variation. In 'meaningful' words, SA is related to variables capturing frequency, ambiguity and textual diversity information, while CA seems more closely related to semantic richness. For function words, however, SA is most strongly correlated with dominance, frequency, and document count, and is uncorrelated with semantic diversity and number of senses.

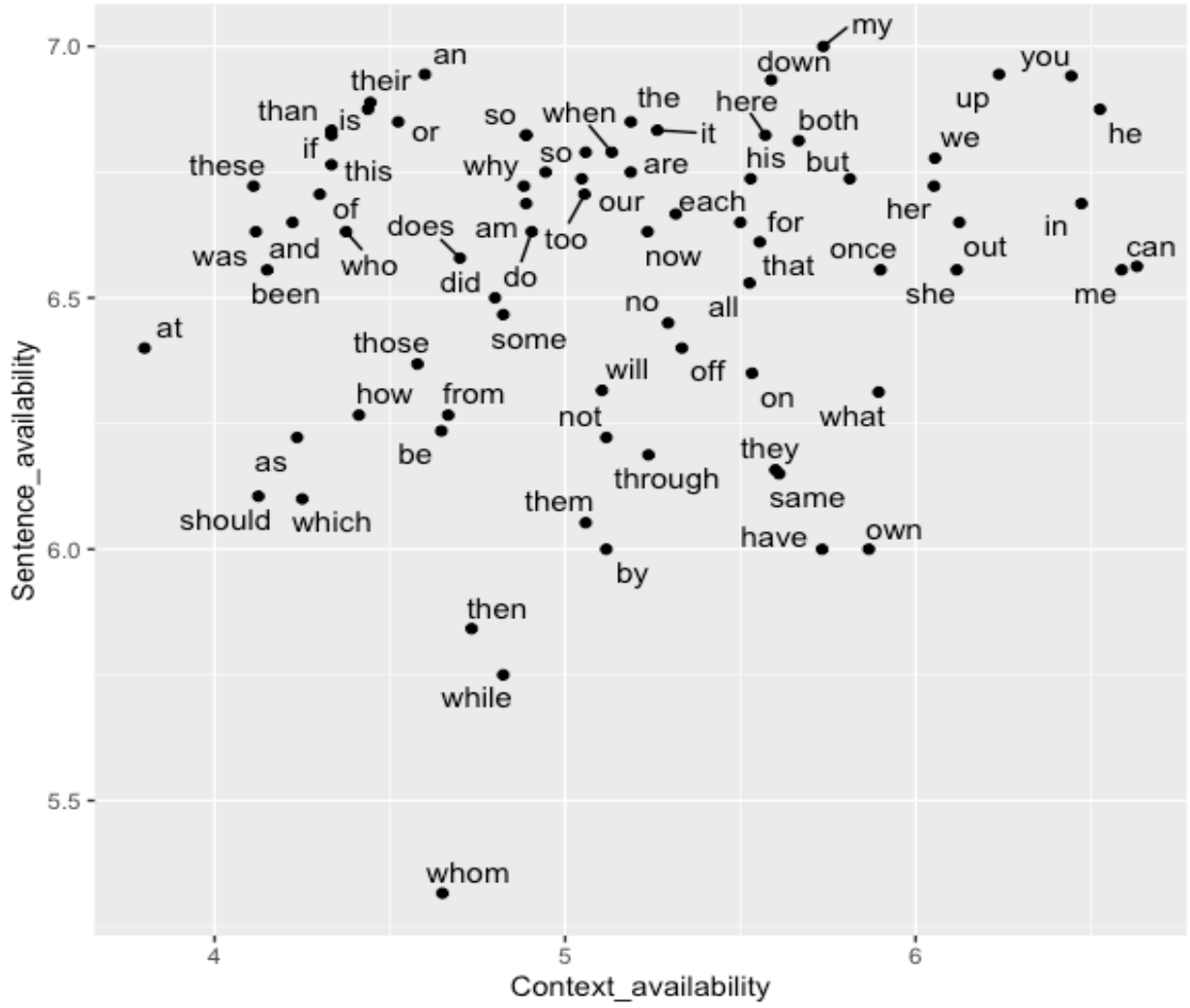


Figure 4: Scatterplot showing the relationship between context availability and sentence availability in function words, with examples labelled.

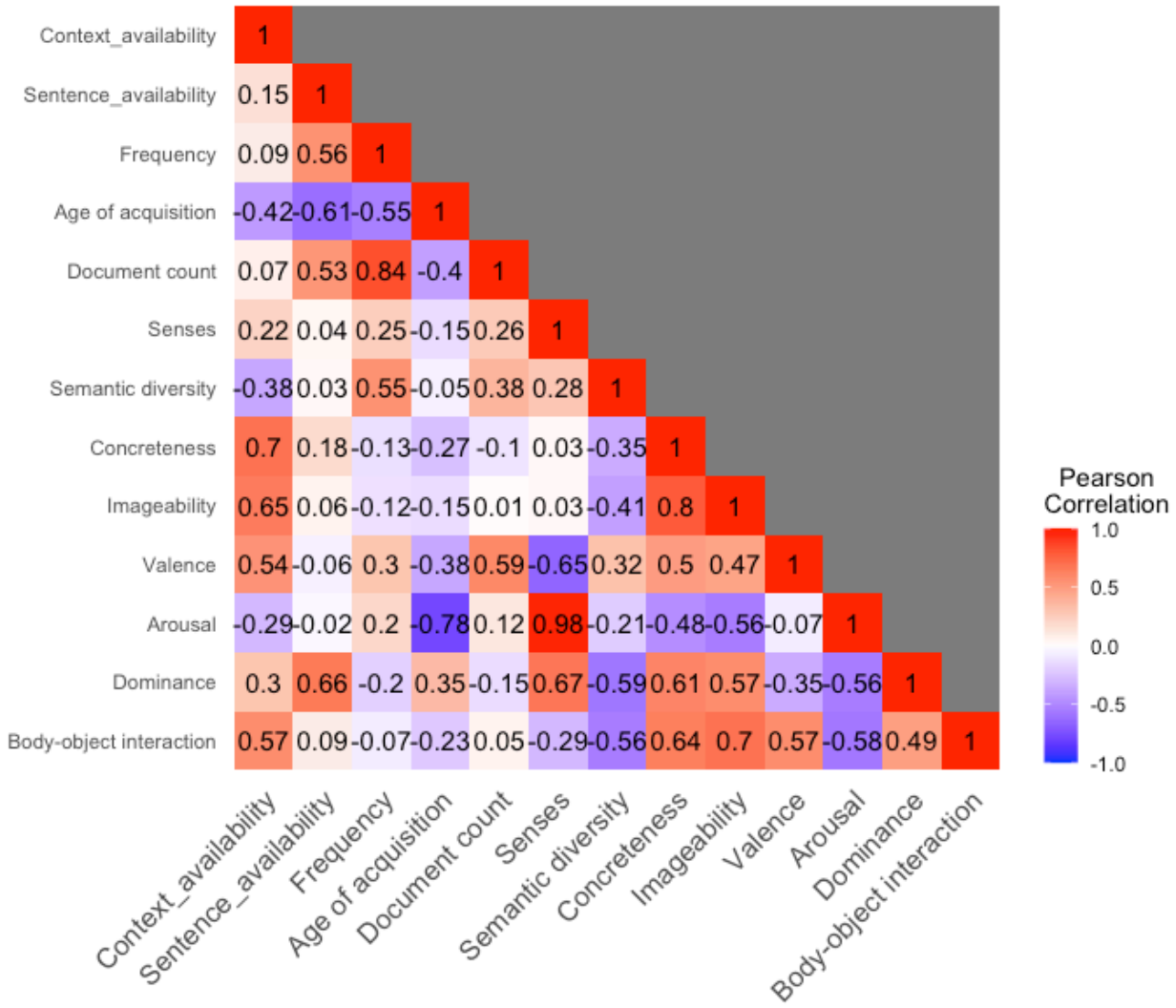


Figure 5: Pearson correlations for CA, SA and other lexical variables in function words, based on N available for each pairwise correlation.

Note: context availability (N = 78), sentence availability (N = 83), frequency (N = 127), age of acquisition (N = 108), document count (N = 123), number of senses (N = 105), semantic diversity (N = 124), concreteness (N = 122), imageability (N = 86), arousal, valence and dominance (N = 6), body-object interaction (N = 90).

Below is a list of the function words excluded from analysis:

I, me, my, myself, we, our, ours, ourselves, you, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, done, should, now.

APPENDIX 5

PREDICTING AVAILABILITY FROM OTHER SEMANTIC VARIABLES

We used regressions to determine which lexical variables explain variance context availability (CA) and sentence availability (SA). This analysis supplements our norms for CA and SA, and provides a more thorough investigation of availability in lexical-semantic space.

To predict our norms, we used the variables included in our main analysis. To avoid multicollinearity, concreteness and body-object interaction were excluded due to their strong correlations with imageability (r = .82 and .73, respectively), and dominance was excluded due to its strong correlation with valence (r = .72). Our model was therefore as follows:

SA ~ frequency + semantic diversity + imageability + document count + number of senses + age of acquisition + valence + arousal

Frequency, imageability, document count, age of acquisition and arousal all explained unique variance in sentence availability (see Table 1). Semantic diversity, number of senses, and valence were non-significant predictors. Our initial model accounted for 57.1% of variance in SA, $F(8, 1689) = 283.8$, $p < .001$.

We then excluded non-significant predictors and repeated the analysis (Table 2). This model explained 56.6% of variance, $F(5, 2202) = 577.3$, $p < .001$.

Table 1: Main effects for predicting sentence availability. Model: SA ~ frequency + semantic diversity + imageability + document count + number of senses + age of acquisition + valence + arousal.

	Estimate	SE	t	p
Intercept	4.77	0.17	28.00	<.001
Frequency	0.40	0.03	15.32	<.001
Semantic diversity	0.03	0.05	0.62	.54
Imageability	0.05	0.01	4.27	<.001
DC	-0.00	0.00	-4.55	<.001
Senses	-0.00	0.00	-0.06	.96
AoA	-0.12	0.01	-16.99	<.001
Valence	-0.01	0.01	-0.91	.36
Arousal	0.05	0.01	4.01	<.001

Table 2: Main effects for predicting sentence availability. Model: SA ~ frequency + imageability + document count + age of acquisition + arousal.

	Estimate	SE	t	p
Intercept	4.57	0.13	35.36	<.001
Frequency	0.43	0.02	20.79	<.001
Imageability	0.05	0.01	5.00	<.001
DC	-0.00	0.00	-5.26	<.001
AoA	-0.11	0.01	-18.71	<.001
Arousal	0.06	0.01	5.53	<.001

We repeated the same procedure for context availability. Here, number of senses and valence did not significantly contribute to the model. The initial model (Table 3) explained 56.1% of variance in CA, $F(8, 1593) = 256.3$, $p < .001$. When non-significant predictors were excluded (Table 4), the model explained 55.5% of variance, $F(6, 1973) = 412.2$, $p < .001$.

Table 3: Main effects for predicting context availability. Model: CA ~ frequency + semantic diversity + imageability + document count + number of senses + age of acquisition + valence + arousal.

	Estimate	SE	t	p
Intercept	4.44	0.17	25.54	<.001
Frequency	0.32	0.03	11.79	<.001
Semantic diversity	-0.16	0.05	-3.07	<.01
Imageability	0.18	0.01	15.26	<.001
DC	-0.00	0.00	-6.58	<.001
Senses	0.00	0.00	0.55	.56
AoA	-0.10	0.01	-13.82	<.001
Valence	-0.01	0.01	-0.56	.58
Arousal	0.07	0.01	5.31	<.001

Table 4: Main effects for predicting context availability. Model: CA ~ frequency + semantic diversity + imageability + document count + age of acquisition + arousal.

	Estimate	SE	t	p
Intercept	4.38	0.15	28.76	<.001
Frequency	0.31	0.02	12.98	<.001
Semantic diversity	-0.11	0.05	-2.46	<.05
Imageability	0.18	0.01	17.13	<.001
DC	-0.00	0.00	-7.36	<.001
AoA	-0.10	0.01	-15.71	<.001
Arousal	0.07	0.01	5.90	<.001

Sentence availability was predicted by frequency, imageability, document count, age of acquisition and arousal. By comparison, context availability was predicted by the same variables plus semantic diversity.

Semantic diversity did not explain unique variance in sentence availability. This suggests that, while SA is linked to contextual variation, it is distinct from corpus-based measures of diversity. The fact that SA is correlated with but not explained by semantic diversity could be explained by similarities in their underlying mechanisms. Particularly, both of these variables correlate strongly with frequency ($r = .66$ for SA; $r = .42$ for semantic diversity).

Together, the variables tested here explained just over half of the variance in CA and SA. The remaining variance in availability could be explained by lexical variables not included in this analysis, or could be unique to CA and SA.

We also ran our initial models again, but included our availability norms as predictors (i.e., CA was used to predict SA; and SA to predict CA). Context availability explained significant variance in sentence availability, and this model accounted for 63.8% of variance. For CA, SA was a significant predictor, with the model accounting for 61.4% of variance. This supports our findings that CA and SA are closely related yet distinct variables.

Finally, we explored how much variance in CA and SA could be accounted for by frequency. We ran our initial models again, this time excluding frequency. In SA, adding frequency to the model explained approximately 6% more variance, with the model excluding frequency accounting for 51.21% of variance. SA was positively predicted by semantic diversity, imageability, document count, valence and arousal, and negatively predicted by age of acquisition. In CA, frequency accounted for about 4% of variance. Imageability, valence and arousal were positive predictors, and age of acquisition was a negative predictor. According to ANOVA comparisons, including frequency significantly improved model fit for both predicting SA, $F(-1, 1690) = 234.84$, $p < .001$, and CA, $F(-1, 1594) = 139.1$, $p < .001$.

APPENDIX 6

PRE-REGISTRATION FOR CHAPTER 4

'SENTENCE AVAILABILITY AND CROSS-MODAL PRIMING'
(ASPREDICTED #62725)

Author(s)
Ellen Taylor (University of Oxford) - ellen.taylor@psy.ox.ac.uk
Kate Nation (University of Oxford) - kate.nation@psy.ox.ac.uk
Yaling Hsiao (University of Oxford) - yaling.hsiao@psy.ox.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
According to Production-Distribution-Comprehension accounts of language (for example, MacDonald, 2013), we learn through both producing and comprehending language. The ease of accessing words in the lexicon is dependent on both item-level properties (e.g. the length or frequency of a word) as well as an individual's recent experience with a word. These two factors are not mutually exclusive: lexical variables often capture the summed experience a word accumulates across a lifetime.

Here, we propose that lexical variables are susceptible to change via recent experience. We aim to test this using a cross-modal priming task, focusing on a production-based rating task of 'sentence availability' (a semantic measure reflecting how easily words can be associated with a sentence context), and a comprehension-based sentence reading task. This will also allow us to test potential differences between production and comprehension based priming, which have been disputed in previous studies (Bock, Dell, Chang & Onishi, 2007, found that these were equally weighed in syntactic priming; Hopman & MacDonald, 2018, found that production is more beneficial for learning of artificial language than comprehension).

Target words and filler words will be exposed to participants either through sentence reading or availability rating (the priming stage). Participants will then complete a distractor task, in which they will see a string of digits one-by-one on screen and then type them into a textbox. Then, participants will encounter the same word list again in either sentence reading or availability rating, whichever they have not yet completed (the outcome stage).

(i) There will be a cross-modal effect of priming such that words will be read more quickly in self-paced reading if they have previously been encountered in the production rating task, compared to other participants reading the same words who have not encountered them in rating.

(ii) There will be a cross-modal effect of priming such that words will be rated more highly and more quickly in the production rating task if they have previously been encountered in sentence reading, compared to other participants rating the same words who have not

encountered them in reading.

(iii) Words higher in established sentence availability (according to previously collected norms) will be read faster in the reading task than words of low sentence availability. They will also be responded to more quickly in the production rating task.

3) Describe the key dependent variable(s) specifying how they will be measured.
(i) Sentence reading task: reading time of individual words from self-paced sentence reading. This will be measured for both the target word and the spill-over word (capturing potential variance carried over from the target position). This time will be measured from word onset to participant pressing the 'space' key for the next word to appear.

(ii) Sentence availability task:
(a) Rating score: sentence availability rating for each target word on a Likert scale ranging from 1 (difficult to think of a sentence) to 7 (easy to think of a sentence).
(b) Rating time: time taken to make each rating, from the onset of trial to when the rating decision is entered.

4) How many and which conditions will participants be assigned to?
Four conditions in a 2 (order: reading first vs. rating first) x2 (sentence availability: high vs. low) design.

Order: This is a between-participants variable. Half of participants will partake in self-paced reading first, then a distractor task, and finally rating. The other half will partake in rating first, then distractor task, and finally self-paced reading.

Sentence availability: This is a within-participant variable. Target words will consist of 20 high availability and 20 low availability words, based on norms collected in a previous study. Word lists have been matched for Zipf frequency and imageability. A separate experiment was conducted to match self-paced reading stimuli for predictability in 40 participants. There were no statistical differences found between high and low SA word lists.

Each participant will respond to 20 target words. Filler-items will also be included (10 filler words for the rating task; 15 filler sentences for the comprehension task).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
Linear mixed-effect models will be used to investigate main effects of priming (task when presented second vs when presented first), availability (high vs low words), and an interaction between these two, using reaction time data. Fixed effects will be included for the interaction between sentence availability condition and priming condition. Random effects will be included for item and participant. As priming order was observed between groups, this will be included as a fixed factor in the random effect for both participant and item. As reactions to sentence availability might differ between subjects, this will be included as a fixed factor for participants.

(i) For self-paced reading, single-word response time will be the dependent variable. Based on pilot analysis, reading RT will be log transformed. The model will then be tested both for the target word and for the spill-over word. This model will be as follows:

Target reading RT ~ priming order * sentence availability + (1+priming order | item) + (1 + priming order + sentence availability | participant)

Spill-over reading RT ~ priming order * sentence availability + (1+priming order | item) + (1 + priming order + sentence availability | participant)

(ii) For the rating production task, we will explore both response time and Likert Scale response. For response time, the model will be as follows:

Rating RT ~ priming order * sentence availability + (1+priming order | item) + (1 + priming order + sentence availability | participant)

For the Likert Scale response (ordinal data):

Likert rating ~ priming order * sentence availability + (1+priming order | item) + (1 + priming order + sentence availability | participant)

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Alongside the 20 target sentences, there will be 15 filler sentences. 10 of these filler sentences will be followed by a forced-choice comprehension question. Participants who fail to answer at least 80% of comprehension questions correctly in the self-reading task will be excluded.

As an attention check, participants who fail 50% or more trials in the distractor task will be excluded from the analysis.
Participants who give the same answer for 90% or more of ratings in the rating task will be excluded.

Based on a pilot experiment, there will be a cut-off threshold for reaction time data at 2,500ms per word for self-paced reading, and at 8,000ms per trial for Likert rating. Outliers beyond 2.5 SDs of the mean across the dataset will be removed from both reading and rating RT data.

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
The experiment was piloted with 80 participants (20 per word list/condition).
Power analysis was then performed using 'simR'. This suggested a total sample size of 200 participants (50 per list/condition).

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
Availability ratings will be compared with norms collected in a previous study which used the same procedure as the rating task.

Digit recall performance will be collected as a distractor task, and may be investigated as a measure of working memory.

As the 'primed' response will always be measured during the last task participants encounter, we will conduct analysis to ensure that any potential differences between response times are due to priming and not simply order effects (i.e. participants naturally speeding their responses throughout the experiment). To assess differences between priming effects and order effects, we will compare response times for filler items. These items will only be encountered once per participant; and the two tasks (reading/rating) will be analysed separately. RTs will be compared between participants when the task is in Position 1 (the first task participants encounter) and when the task is in Position 3 (the last task participants encounter). For the reading task, we will analyse reading time of filler words in filler sentences. For the rating task, we will analyse rating response time for filler words. We expect to find no significant difference between response time for these items. If no significant difference is found here, this allows us to assume that any effects found in target words are not due to order of presentation.

https://aspredicted.org/di6kp.pdf

APPENDIX 7

TESTING SELF-PACED READING STIMULI FOR PREDICTABILITY

This study concerns the self-paced reading stimuli used in Chapter 4. The purpose of this experiment was to ensure that both stimuli lists did not differ in their predictability. Predictability of words influences how we process language (for review, see Van Petten & Luka, 2012). To ensure that this did not confound participant reading times, we tested our self-paced reading stimuli for predictability using a sentence continuation experiment.

40 participants (different to those in the main priming study) were presented with the prior and target sentences included in the self-paced reading study. Each participant was randomly assigned to one of two groups, ensuring that they did not respond to the same sentence frame twice.

The sentences were cut off before the target word. Participants were asked to type the next word that they would expect to see next. For example:

The comet shot through the sky. It was a great

In this example, the target word was 'awe'. We computed whether or not the participants' responses matched the target word.

A preliminary experiment found no difference between groups when participants gave one response for each sentence frame, $t(788.54) = -0.393$, $p = .695$. However, one word ('drought') was expected by 30.77% of participants, which was substantially more predictability than other words.

Based on this finding, the phrasing of sentences was altered to have a more even predictability across words. We then tested the sentence continuation experiment again in 40 different participants, recruited via Prolific. Here, each participant was asked to give three possible words that could each continue the sentence. This allowed us to maximise our data.

Table 1 shows the percentage of responses for each context that matched the target word (i.e. where participants correctly predicted the target word). The majority of words (24 out of 40) received a predictability norm of 0 according to this study. This is substantially more predictable than the other words. We then investigated whether there were differences in predictability between the high and low availability conditions.

A t-test found no difference between the two groups, $t(2136) = 1.454$, $p = .146$. A linear mixed effect model was then performed, including random effects of word and participant. The effect of condition on predictability were negligible (see Table 2). Therefore, despite variability in the predictability of our sentences, we can assume that the self-paced reading sentences were matched for predictability.

Table 1: Percentage of responses where words were correctly predicted, based on 120 responses from 40 participants.

Word	SA Condition	Predictability (%)
song	High	11.67
crumb	High	11.11
strand	Low	10.53
blouse	High	8.93
boot	High	5.56
brink	Low	3.70
bail	Low	2.0
script	High	1.96

groom	High	1.92
dome	Low	1.92
ant	High	1.81
drought	Low	1.79
harp	Low	1.74
yield	Low	1.72
mink	Low	1.67
waltz	Low	1.64
awe	High	0.00
brand	High	0.00
draft	High	0.00
fate	High	0.00
fund	High	0.00
loaf	High	0.00
mate	High	0.00
maze	High	0.00
mile	High	0.00
niece	High	0.00
root	High	0.00
trap	High	0.00
wreck	High	0.00
bulge	Low	0.00
clan	Low	0.00
clause	Low	0.00
lynx	Low	0.00
mast	Low	0.00
midst	Low	0.00

mule	Low	0.00
ranch	Low	0.00
shaft	Low	0.00
spear	Low	0.00
stance	Low	0.00

Table 2: Fixed effects of condition on predictability.

	Estimate	Std Error	z	p
Intercept	-5.00	0.60	-8.36	<.001
SA Condition	-0.25	0.68	-0.37	0.71

Note: this tested the model, prediction ~ SA condition + (1|participant) + (1|word).

APPENDIX 8

ANALYSIS OF FILLER ITEMS

In the design of Chapter 4's experiment, priming was reflected in a change in participants' performance in the rating-prime condition and the reading-prime condition (see Method for more details). Due to the design of our experiment, 'primed' performance was always measured during the final task participants completed. Extraneous variables could lead to either decreased (distractors, fatigue) or increased (practice effects) performance across tasks. That is, priming is always confounded by order in the design of this experiment. To explore this, we analysed responses to the filler items produced in either the first task or the second task. Filler items differed between reading and rating tasks. This meant that all participants encountered the same filler items, but only once. Filler items were therefore always unprimed. If we observed a difference in filler item performance between the reading-prime and rating-prime groups, this would suggest that priming effects on reading may be confounded by task order.

All filler words were only encountered once by participants, meaning that they were always unprimed regardless of which group participants were in. In our analysis, performance was compared for the same words between groups: the reading-prime group encountered fillers during their first task, and the rating-prime group encountered fillers during their second task, but all filler words were unprimed. These were only seen by participants in one task – in a sense then, they can be considered as unprimed.

In the reading task, filler words were embedded in sentences, and were presented in the fifth position of their sentence to remain consistent with the target words presented during self-paced reading. In the rating task, filler words were centred on screen. In each trial, participants were asked to rate how easily they could think of a sentence for the word. This meant that filler words were not discernible to participants from target words. The only methodological difference was that

target words were the same for both reading and rating tasks, whereas filler items differed for each task.

We also controlled filler words for lexical variables. In the reading task, filler words were not significantly different from target items in frequency, $t(10.231) = -0.97$, $p = .354$, or imageability, $t(11.624) = -0.86$, $p = .41$. In the rating, filler words were also matched to target words for Zipf frequency, $t(7.0991) = -0.33$, $p = .75$, or imageability, $t(7.565) = -0.67$, $p = .52$.

Analysis considered performance on the 20 filler words (10 per task). We compared reading times, rating times and rating scores for these words when they were encountered in the first task vs the second task.

T-tests found no difference between log reading times (M) of filler words when self-paced reading was presented as the first task (6.18) vs the third task (6.21), $t(1706) = 1.21$, $p = .23$, but participants reading sentences during their first task (6.160) were significantly faster at reading spill-over filler words than participants who encountered self-paced reading as their third task (6.22), $t(1697.7) = 2.75$, $p = <.01$. A small difference was found between positions for response time, $t(1872.5) = -2.24$, $p = .03$, and Likert scores, $t(1865.7) = 2.49$, $p = .01$. Here, words rated during the first task (7.87 log RT, 5.83 Likert score) were responded to faster and rated as easier to associate with sentences than words that were read first (7.92 log RT, 5.65 Likert score).

APPENDIX 9

LINEAR MIXED EFFECT MODEL STRUCTURES

In Chapter 4, linear mixed effect models were used to investigate the effects of and interaction between sentence availability and priming on reading times, Likert ratings, and rating response times. This appendix reports the simplified models that were tested and reported in Chapter 4.

As priming order was observed between groups, this was included as a fixed factor in the random effect for both participant and item. As reactions to sentence availability might differ between subjects, this was included as a fixed factor for participants. Therefore, the following model was initially tested:

> Priming order * sentence availability + (1+priming order | item) + (1 + priming order + sentence availability | participant)

For each outcome variable, this model was tested and then simplified until the model converged. If the model failed to converge, or converged with singular fit, the fixed factor explaining the least amount of variance was removed. This gave the following models, which correspond to the results reported in this study:

For reading time of the target word, spill-over word, and Likert rating score was, SA Condition * Priming Condition + (1 |participant).

For reaction time of the rating task, SA Condition * Priming Condition + (1+ SA|participant).

APPENDIX 10

EXPLORATORY ANALYSES FOR CHAPTER 4

Do participants' ratings correlate with previous availability norms?

First, we tested Pearson correlation between participants' ratings in the current experiment and sentence availability norms. While significant, the correlation was much lower than expected (r = .349). We then tested correlations by condition (see Table 1). Correlations were significant in all conditions. This relationship patterned similarly between primed and unprimed conditions, and between high and low availability words.

Table 1: Pearson correlations between participants' ratings and sentence availability norms, split by experimental condition (upper section) and interaction condition (lower section).

Condition	r	p
Unprimed words (N = 40)	.30	<.001
Primed words (N = 40)	.35	<.001
High availability (N = 20)	.13	<.001
Low availability (N = 20)	.17	<.001

Condition	r	p
Unprimed words, low SA	.18	<.001
Unprimed words, high SA	.15	<.001
Primed words, low SA	.16	<.001
Primed words, high SA	.10	<.01

All conditions correlated positively with our previous norms, validating our values collected in Chapter 3. However, correlations were lower than expected as the task remained the same. This could be due to the range of items included in our experiments. Chapter 3 provided participants with a broad range of words, while participants in the current study encountered only a few words, which were tightly controlled. Participants might have made their ratings relative to the other words presented, therefore explaining the disparity in our results.

Is our wordlist representative of the lexicon?

As the correlations were lower than expected, we explored whether the items used in our study could reproduce the results found in our norms study. Our wordlists were tightly controlled, allowing for a distinction in sentence availability while matched for frequency and imageability. This small and restricted sample of words might not be a natural representation of the broader lexicon. To test whether these words pattern similarly to a larger sample, we attempted to replicate our result in facilitating lexical decision.

In Taylor, Nation and Hsiao's previous study collecting 3,000 norms for context and sentence availability, sentence availability explained unique variance in lexical decision performance once frequency, age of acquisition, imageability and semantic diversity were controlled for. This result was true for data from both the British Lexicon Project (BLP; Keuleers, Lacey, Rastle, & Brysbaert, 2012) and the English Lexicon Project (ELP; Balota et al., 2007).

Zipf frequency was taken from Van Heuven et al. (2014), age of acquisition from Kuperman et al. (2012), imageability from Cortese and Fugett (2004), semantic diversity from Hoffman et al. (2013), and sentence availability from my previous norming study. Consistent with the previous

study, linear mixed effect models were fitted using lme4 in R and random effects were included for word and participant. All variables were centred and scaled.

Within our 40 target words, sentence availability failed to explain unique variance. In fact, the only variables which contributed significantly to the model were age of acquisition (for accuracy) and frequency (for response time, see Table 2). When fitted to ELP data, a similar pattern was found for RT, with no variables explaining unique variance in the data. The model for explaining ELP accuracy data failed to converge, but a linear regression model demonstrated that only frequency was a significant predictor (see Table 3).

Table 2: Fixed effects predicting lexical decision accuracy (upper section) and RT (lower section) in BLP data (N = 40 words).

	Estimate	Std Error	z	p
Intercept	3.83	0.43	8.96	.08
Frequency	-0.01	0.39	-0.02	.99
AoA	-0.97	0.43	-2.24	.03
SemD	-0.34	0.25	-1.33	.18
Imageability	-0.52	0.37	-1.40	.16
SA	0.47	0.41	1.16	.25

	Estimate	Std Error	t	p
Intercept	562.46	13.62	41.31	<.001
Frequency	-19.34	8.41	-2.30	.03
AoA	4.72	10.82	0.44	.67
SemD	-3.41	7.04	-0.48	.63

Imageability	-19.88	10.49	-1.89	.07
SA	-0.30	9.31	-0.03	.97

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + sentence availability + (1|participant) + (1|word).

Table 3: Coefficients predicting lexical decision accuracy (upper section, using linear regression) and RT (lower section, using linear mixed effects) in ELP data (N= 40 words).

	Estimate	Std Error	z	p
Intercept	0.94	0.01	147.04	<.001
Frequency	0.02	0.01	2.18	.03
AoA	-0.01	0.01	-1.29	.20
SemD	-0.00	0.01	-0.04	.97
Imageability	0.01	0.01	1.01	.31
SA	-0.01	0.01	-0.54	.59

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + sentence availability.

	Estimate	Std Error	t	p
Intercept	662.32	11.30	58.63	<.001
Frequency	-24.62	14.24	-1.73	.09
AoA	29.95	18.03	1.66	.11
SemD	-7.03	11.82	-0.59	.56
Imageability	-6.41	17.56	-0.37	.72
SA	2.11	15.61	0.14	.89

Note. Model structure: LD ~ Freq + AoA + SemD + imageability + sentence availability + (1|participant) + (1|word).

As we could not replicate our results from Chapter 3, this suggests that our wordlist might not be representative of the wider sample used for our CA and SA norms. This could be because we matched words for frequency and imageability while manipulating sentence availability. However, note that the N in this analysis is smaller for both items and participants, and so results might not be comparable.

APPENDIX 11

PRE-REGISTRATION FOR CHAPTER 5

'*Investigating the role of recent lexical experience and semantic variability on lexical processing*'
(AsPredicted #88684)

Author(s)
Ellen Taylor (University of Oxford) - ellen.taylor@psy.ox.ac.uk
Yaling Hsiao (University of Oxford) - yaling.hsiao@psy.ox.ac.uk
Kate Nation (University of Oxford) - kate.nation@psy.ox.ac.uk

1) Have any data been collected for this study already?
No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?
1. How will semantic diversity and imageability affect lexical processing?

This hypothesis concerns the first task that participants complete. Here, we aim to replicate the results of Hoffman and Woollams (2015). Participants will be randomly assigned to one of two conditions. Each group will take part in either lexical decision or semantic relatedness decision tasks for words that vary in semantic diversity and imageability.
In line with Hoffman and Woollams's (2015) findings, we predict that high semantic diversity words will be associated with faster RTs for lexical decision and slower RTs for semantic relatedness judgements.
We predict that imageability will facilitate both tasks. High imageability words will be associated with faster lexical decision and semantic relatedness decision.

2. Will lexical decision and semantic decision facilitate subsequent lexical processing?

All participants will take part in two tasks:
1) Pre-test: lexical decision OR semantic relatedness decision
2) Post-test: lexical decision

The same items will be presented for pre-test and post-test.

During the pre-test, participants will be randomly assigned to one of two tasks: lexical decision or semantic relatedness decision. We predict that participants' lexical processing will be facilitated after recent lexical experience in both conditions. Lexical decision will be faster post-test than baseline (taken from pre-test performance for the lexical decision condition, compared either between- or within-subjects depending on the condition).

Semantic relatedness decision involves a more active reflection on meaning, while lexical decision involves reflecting on form, which is less cognitively demanding. The effects of this processing on subsequent lexical experience might therefore be different. We predict that

semantic decision will facilitate subsequent lexical decision to a greater extent than lexical decision. For high semantic diversity words, which have high levels of polysemy and are associated with slower semantic decision times, the effect of recent experience might be more pronounced. Therefore, for high diversity words, post-test performance will be faster for the semantic decision condition than the lexical decision condition.

3) Describe the key dependent variable(s) specifying how they will be measured.
We have two key dependent variables:

1) Performance pre-test (either lexical decision or semantic relatedness decision). This will be used to assess the effects of imageability and semantic diversity on lexical processing. We will measure participants' accuracy and response time.

2) Performance in post-test (lexical decision). This will be used to determine the effects of previous lexical experience on word processing. We will measure participants' accuracy and response time.

4) How many and which conditions will participants be assigned to?
Participants will be randomly assigned to one of two conditions. Depending on the condition, participants will be given a different pre-test task: lexical decision or semantic decision. The post-test task will be the same for both conditions.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.
1. How will semantic diversity and imageability affect lexical processing?

Linear mixed effect models will be used to assess whether semantic diversity and imageability affect lexical processing. Both lexical variables will be entered into the same model. This model will be tested separately for the effects on the two outcome variables (lexical decision and semantic relatedness decision). We will control for random effects of item and participant. The model will therefore be as follows:

Pre-test RT and accuracy ~ semantic diversity * imageability + (1| word) + (semantic diversity * imageability | participant)

This model will be used to predict pre-test performance for both conditions, separately. We will therefore test the fit of this model for:
1. Lexical decision RT
2. Lexical decision accuracy
3. Semantic relatedness decision RT
4. Semantic relatedness decision accuracy

2. Will lexical decision and semantic decision facilitate subsequent lexical processing?

Lexical decision accuracy and response time will be analysed within subjects for participants in Condition 1 (lexical decision) and between subjects for participants in Condition 2 (semantic relatedness decision). Lexical decision performance of participants in Condition 1 in pre-test will act as baseline.
Linear mixed effect models will be used to determine whether the type of task participants partake in during pre-test (lexical decision/semantic relatedness decision) predicts subsequent

lexical decision response time. Random effects of word and participant will be included. We will also assess the effects of semantic diversity and imageability to investigate whether effects differ across the range of lexical properties. The model will therefore be as follows:

Post-test performance ~ Condition * semantic diversity * imageability (1|word) + (semantic diversity * imageability |participant)

This model will be used to predict post-test RT and accuracy separately.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.
Based on Hoffman and Woollam's (2015) exclusion criteria, we will exclude participants who take longer than 2,000ms to respond on more than 5% of trials. We will then trim each participants' data by moving RTs that fall more than 2 standard deviations from the mean.

7) How many observations will be collected or what will determine sample size?
No need to justify decision, but be precise about <u>exactly</u> how the number will be determined.
Hoffman and Woollams (2015) included 49 participants in their experiments investigating lexical decision (N = 24) and semantic relatedness decision (N = 25). As we will include between subject analysis, we will sample 50 participants per condition, per wordlist.

8) Anything else you would like to pre-register?
(e.g., secondary analyses, variables collected for exploratory purposes, unusual analyses planned?)
We will measure print exposure using the Author Recognition Test. This will be measured for all participants between the pre-test and post-test tasks. Exploratory analysis will investigate the relationship between participants' print exposure and lexical processing. We predict that lexical decision and semantic decision performance will be faster for participants with higher ART scores. We also predict that the strength of the priming effect and ART score will be negatively correlated, such that participants with less prior print exposure will receive more of a facilitation 'boost' from recent lexical experience. Here, we will use linear mixed effect modelling. We include semantic diversity and imageability to assess whether these variables interact with ART score.
Pre-test performance ~ ART score * imageability * semantic diversity + (ART score | word) + (semantic diversity + imageability | participant)
Post-test performance ~ Condition * ART score * imageability * semantic diversity + (ART score | word) + (semantic diversity + imageability | participant)

These models will be used to predict post-test RT and accuracy separately.

https://aspredicted.org/ud9ym.pdf

APPENDIX 12

HOW DO SEMANTIC DIVERSITY AND IMAGEABILITY AFFECT INITIAL LEXICAL PROCESSING?: SUPPLEMENTARY ANALYSIS FOR CHAPTER 5

Following Hoffman and Woollams (2015), a 2x2 ANOVA was used to test the effects of lexical variables on performance. For lexical decision, there was no effect of semantic diversity, $F(1, 152) = 3.79$, $p=.05$) or response time, $F(1, 152) = 0.05$, $p=.82$). Imageability also did not have an effect on accuracy, $F(1, 152) = 1.24$, $p=.27$, or response time, $F(1, 152) = 1.79$, $p=.18$. No interaction was found between semantic diversity and imageability for either accuracy, $F(1, 152) = 2.33$, $p=.13$, or response time, $F(1, 152) = 0.39$, $p=.54$.

For semantic decision, words high in semantic diversity were responded to more accurately, $F(1, 160) = 4.384$, $p=.04$, but diversity did not explain variance in response time, $F(1, 160) = 3.53$, $p=.06$. High imageability words were associated with better accuracy, $F(1, 160) = 4.62$, $p=0.3$, but imageability did not predict response time, $F(1, 160) = 2.05$, $p=.15$. There was also a significant main effect of relatedness, with related trials being responded to faster, $F(1,160) = 4.50$, $p=.04$, than unrelated trials; but this was not significant for accuracy data, $F(1,160) = 0.06$, $p=.82$. Relatedness significantly interacted with semantic diversity in accuracy data. Here, higher accuracy was seen for low diversity words in trials which were semantically related; but there were no differences in accuracy between items in unrelated trials. All other interactions were non-significant.

Next, I tested the same linear-mixed effect model as reported in the results section of Chapter 5, but I used continuous variables rather than categoric. Here, instead of dummy coding semantic diversity and imageability as binary variables which were either high or low, these variables were entered continuously. Semantic diversity values were taken from Hoffman et al. (2013), and imageability values taken from Cortese and Fugett (2004). As variables were continuous, contrast coding was not used. Relatedness remained a binary variable.

Accuracy and reaction time performance were tested separately. Models were as follows. For lexical decision:

Lexical decision performance ~ semantic diversity * imageability + (1|word) + (1| participant)

For semantic decision:

Semantic decision ~ semantic diversity * imageability * relatedness + (1|word) + (1| participant)

When entered continuously, imageability significantly predicted lexical decision response time, with high imageability words responded to faster. Both semantic diversity and imageability were associated with more accurate responses in lexical decision, and there was a significant interaction between these variables (Table 1). For semantic decision, neither of these variables, nor relatedness, significantly predicted participants' performance (Table 2).

Table 1: Fixed effects predicting lexical decision response time (above) and accuracy (below) for the model: Lexical decision performance ~ semantic diversity * imageability + (1|word) + (1| participant).

	Estimate	Std Error	df	t	p
Intercept	6.71	0.09	244.21	71.46	<.001
Semantic diversity	-0.11	0.06	230.23	-2.02	.04
Imageability	-0.05	0.02	228.75	-2.39	.01
Interaction	0.02	0.01	228.11	1.70	.09

	Estimate	Std Error	df	t	p

Intercept	0.75	0.07	232.52	10.44	<.001
Semantic diversity	0.12	0.04	229.97	2.67	<.01
Imageability	0.04	0.02	229.27	2.37	.02
Interaction	-0.02	0.01	229.17	-2.05	.04

Table 2: Fixed effects predicting semantic decision response time (above) and accuracy (below) for the model: Semantic decision performance ~ semantic diversity * imageability * relatedness + (1|word) + (1| participant).

	Estimate	Std Error	df	t	p
Intercept	7.04	0.11	468.5	65.33	<.001
Semantic diversity	0.64	0.06	440.1	0.99	.33
Imageability	-0.02	0.02	438.3	-0.82	.41
Relatedness	-0.10	0.12	7860	-0.89	.38
S * I	-0.00	0.01	439.8	-0.16	.88
S * R	0.02	0.07	7857	0.34	.74
I * R	-0.01	0.03	7852	0.20	.85
S * I * R	0.00	0.02	7853	-0.12	.91

	Estimate	Std Error	df	t	p
Intercept	0. 95	0.16	364.4	5.84	<.001
Semantic diversity	-0.01	0.10	363.1	-0.23	.91
Imageability	0.01	0.04	360.7	0.36	.72
Relatedness	-0.14	0.14	9091	-0.94	.35
S * I	-0.01	0.02	361.9	-0.25	.80

S * R	-0.02	0.09	9078	-0.26	.80
I * R	0.03	0.03	9075	0.80	.42
S * I * R	-0.00	0.02	9066	-0.24	.81

These results generally fit the pattern of those reported with lexical variables entered categorically. Changes in significance between these two sets of results suggest that these effects may be small.