Learning vocabulary through generation with translation-ambiguous and semantically-ambiguous words

by

Caitlin Ann Rice

B. A., Grinnell College, 2011

M. S., University of Pittsburgh, 2017

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Caitlin Ann Rice

It was defended on

November 25, 2019

and approved by

Charles A. Perfetti, Distinguished University Professor, Department of Psychology

Tessa Warren, Associate Professor, Departments of Psychology, Linguistics and Communications Sciences and Disorders

David Plaut, Professor, Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University

Thesis Advisor/Dissertation Director: Natasha Tokowicz, Associate Professor, Departments of Psychology and Linguistics

Learning vocabulary through generation with translation-ambiguous and semantically-ambiguous words

Caitlin Rice, PhD

University of Pittsburgh, 2019

Learning a language involves learning both word forms and word meanings, as well as the ways in which these forms and meanings are connected (e.g., Rice & Tokowicz, 2019). Unfortunately for language learners, language is rife with both within-language semantic ambiguity and cross-language translation ambiguity. Ambiguity often leads to difficulty learning and processing new words (e.g., Degani & Tokowicz, 2010). In three experiments, I investigate whether strengthening meaning representations during learning via the generation of semantically-related material (i.e., the generation effect) may mitigate difficulties associated with learning translation-ambiguous and semantically-ambiguous words. In Experiment 1, native English speakers learned word pairs that were translation-ambiguous or unambiguous from German to English using a generation task (write a sentence containing a target word) or a control task (read an experimenter-generated sentence containing a target word). Results revealed that generation was more beneficial for ambiguous than unambiguous words, and furthermore that individual differences in WM and inhibitory control affected ambiguous and unambiguous words in different ways. In Experiment 2, native English speakers learned unknown English ambiguous and unambiguous words using the same generation or control tasks as in Experiment 1. Results revealed a complex interaction of ambiguity, generation, and inhibitory control during free recall. In Experiment 3, native English speakers learned the same words as in Experiment 2 with either generation or a control task, but additionally words were trained with either context sentences or

definitions. Results revealed that ambiguity, generation, and working memory interact during free recall, and furthermore that meanings for words that were trained with definitions and generation were recalled significantly better than words trained with definitions but without generation, or words trained with context sentences with or without generation. Results are examined in light of the semantic settling dynamics account (Armstrong & Plaut, 2016) and the instance-based framework for word learning (Bolger, Balass, Landen, & Perfetti, 2008), and implications for models of translation ambiguity, semantic ambiguity, and generation effects are discussed.

Table of contents

Preface xv

1.0 Introduction 1

1.1 Modeling semantic ambiguity effects 3

1.2 Theories of L2 vocabulary learning and translation ambiguity 4

1.3 The translation-ambiguity disadvantage 8

1.3.1 Mitigating the translation-ambiguity disadvantage 11

1.4 Theories of generation effects 12

1.5 Overview of experiments 15

1.5.1 Description of proposed research 16

2.0 Experiment 1 19

2.1 Generation effects in L2 vocabulary learning 20

2.2 Generation effects with translation-ambiguous words 22

2.3 Individual differences in L2 vocabulary learning 24

2.4 Experiment overview 26

2.5 Methods 28

2.5.1 Participants 28

2.5.2 Design 28

2.5.3 Stimuli 28

2.5.4 Procedure 30

2.5.4.1 Vocabulary training 31

2.5.4.2 Free recall 32

2.5.4.3 Translation production 33

2.5.4.4 Individual difference tasks 34

2.6 Results 37

2.6.1 Statistical approach 37

2.6.2 Vocabulary training analyses 38

2.6.3 Free recall results 39

2.6.3.1 English free recall accuracy 41

2.6.3.2 German free recall accuracy 45

2.6.4 Translation production results 49

2.6.4.1 Translation production accuracy analyses 50

2.6.4.2 Translation production RT results 55

2.7 Discussion 59

2.7.1 Future directions 63

2.7.2 Conclusions 66

3.0 Experiment 2 67

3.1 Semantic ambiguity effects 68

3.2 Generation effects and word learning 70

3.3 Generation effects with semantically-ambiguous words 73

3.4 Individual differences in ambiguous word learning 74

3.5 Experiment overview 76

3.6 Methods 77

3.6.1 Participants 77

3.6.2 Design 77

3.6.3 Stimuli 78

3.6.4 Procedure 79

3.6.4.1 Familiarity check 80

3.6.4.2 Vocabulary training 81

3.6.4.3 Free recall 81

3.6.4.4 Meaning production 81

3.6.4.5 Individual difference tasks 82

3.7 Results 83

3.7.1 Statistical approach 83

3.7.2 Familiarity check 84

3.7.3 Vocabulary training analyses 85

3.7.4 Free recall results 85

3.7.5 Meaning production results 89

3.8 Discussion 93

3.8.1 Future directions 98

3.8.2 Conclusions 100

4.0 Experiment 3 101

4.1 Word learning from context 102

4.2 Experiment overview 106

4.3 Methods 108

4.3.1 Participants 108

4.3.2 Design 108

4.3.3 Stimuli 108

4.3.4 Procedure 109

4.3.4.1 Familiarity check 110

4.3.4.2 Vocabulary training 110

4.3.5 Free recall 111

4.3.6 Meaning production 111

4.3.7 Forced choice sentence completion 111

4.3.8 Individual difference tasks 112

4.4 Results 113

4.4.1 Statistical approach 113

4.4.2 Familiarity check 113

4.4.3 Vocabulary training analyses 113

4.4.4 Free recall results 114

4.4.5 Meaning production results 120

4.4.6 Forced choice sentence completion results 126

4.5 Discussion 134

4.5.1 Future directions 141

4.5.2 Conclusions 142

5.0 General discussion 143

5.1 Implications for models of semantic ambiguity 144

5.2 Implications for models of generation effects 149

5.3 The role of individual differences 152

5.4 Future directions 154

5.5 Conclusions 156

Appendix A Experiment 1 stimulus characteristics 158

Appendix B Sentence norming for Experiment 1 160

B.1 Methods 160

B.1.1 Participants 160

B.1.2 Procedure 160

Appendix C Language history questionnaire data for Experiment 1 167

Appendix D Sentence norming procedures for Experiments 2 and 3 168

D.1 Methods 168

D.1.1 Participants 168

D.1.2 Procedure 168

Bibliography 175

List of tables

Table 1. Example stimuli for Experiment 1 29

Table 2. Experiment 1 timeline 30

Table 3. Fixed effects estimates for Model 1, English free recall 43

Table 4. Random effects estimates for Model 1, English free recall accuracy 44

Table 5. Fixed effects estimates for Model 2, German free recall 47

Table 6. Random effects estimates for Model 2, German free recall 48

Table 7. Fixed effects estimates for Model 3, translation production accuracy 53

Table 8. Random effects estimates for Model 3, translation production accuracy 54

Table 9. Fixed effects estimates for Model 4, translation production RT 58

Table 10. Random effects estimates from Model 4, translation production RT 59

Table 11. Example stimuli for Experiment 2 79

Table 12. Experiment 2 timeline 80

Table 13. Fixed effects estimates for Model 5, free recall 88

Table 14. Random effects estimates for Model 5, free recall 89

Table 15. Fixed effects estimates for Model 6, meaning production 92

Table 16. Random effects estimates for Model 6, meaning production 93

Table 17. Experiment 3 timeline 110

Table 18. Fixed effects estimates for Model 7, free recall 117

Table 19. Random effects for Model 7, free recall 118

Table 20. Fixed-effects estimates for Model 8, meaning production 122

Table 21. Random effects estimates for Model 8, meaning production 123

Table 22. Fixed effects estimates for Model 9, sentence completion 129

Table 23. Random-effects estimates for Model 9, sentence completion 131

Table 24. Experiment 1 stimulus characteristics 158

Table 25. Experiment 1 stimulus definitions and sentences 161

Table 26. Language history questionnaire data for Experiment 1 167

Table 27. Stimulus and sentence ratings for Experiments 2 and 3 170

List of figures

Figure 1. Predicted semantic activation over time as a function of ambiguity, according to the Semantic Settling Dynamics account (Armstrong, 2012)..... 4

Figure 2. The RHM-RER model of L2 vocabulary learning (Rice & Tokowicz, 2019)..... 6

Figure 3. The RHM-TA for meaning-ambiguous words (Eddington & Tokowicz, 2013)..... 7

Figure 4. Estimated probability of English free recall by word type and Waters total span. 44

Figure 5. Estimated proportion correct for English free recall by training condition, word type, and Simon score..... 45

Figure 6. Predicted probability of correct response for German free recall accuracy by word type and condition..... 48

Figure 7. Estimated probability of German free recall by word type and Simon score. 49

Figure 8. Estimated probability of correct translation production by word type and training condition. 54

Figure 9. Estimated probability of correct translation by word type and Simon score. 55

Figure 10. Estimated proportion correct for free recall by word type, training condition, and Simon score. 89

Figure 11. Estimated probability of free recall by word type, training condition, and Waters set size span. 119

Figure 12. Estimated probability of a correct response by training condition and training material. 124

Figure 13. Estimated probability of correct response by Waters and Simon scores..... 125

Figure 14. Probability of correct response by word type, Waters total span, and Simon score . 126

Figure 15. Estimated probability of correct sentence completion by training condition and training material. 133
Figure 16. Estimated proportion correct by training condition, word type, and training material 133
Figure 17. Estimated proportion correct by word type, Waters total span, and training material 134

Preface

That this dissertation exists at all is a testament to all of the many wonderful friends, family, and mentors in my life, without whom this work would never have been possible. First and foremost, I give my deepest thanks to my advisor Natasha Tokowicz who changed my life when she accepted me into the PLUM lab. Thank you for guiding me through this journey, for providing wisdom, advice, and support, and teaching me what it means to be a researcher and a scientist. I will always be grateful to you.

I have been lucky to have had many wonderful mentors throughout my research journey. In particular, I thank the other members of my dissertation committee: Tessa Warren, Charles Perfetti, and Dave Plaut. Without your guidance, mentorship, and advice this project would not have been possible. I also thank Scott Fraundorf and Blair Armstrong for invaluable mentorship and assistance over the years, and for welcoming me into your labs. I also extend my gratitude to the many mentors who guided and inspired me along the way: Amanda Caldwell-Tarr, Susan Nittrouer, Jen Dobson, Janet Gibson, and Ann Ellis.

Thank you to all of the amazing graduate students in the department for contributing to a positive and intellectually stimulating environment, among them: Gaby Terrazas Duarte, Michelle Colvin, Gina Calloway, Lea Martin, Aleksandra Petkova, Joshua Tremel, Nabila Jamal-Orozco, Xiaoping Fang, Katherine Martin, Alba Tuninetti, Charlie Eddington, Evelyn Milburn, Adeetee Bhide, and many more. And my deepest gratitude goes to all of the many undergraduate research assistants who helped me with data collection and coding along the way, as well as the many participants in these studies – without your hours of work none of this would have been possible.

Thank you to *Language Learning* for providing the funds to complete this work – your support made the completion of this document possible.

My friends have been the most joyful part of my journey through graduate school. To Aleks, Lea, Rebecca, Brendan, and Tyler – thank you for being my first Pittsburgh fam and helping me keep my sanity through the hardest parts of grad school. To all the Future Robots – thank you to each and every one of you for touching my life deeply and for forging friendships that will last for many years. To Sarah and Elliot – thank you for your many years of deep friendship, for all of the Tasty Tuesdays (past and future), and for your advice and love. And to Alexa – thank you for sharing your love and wisdom, and guiding me through so many life changes.

Above all, I give my everlasting love and gratitude to my family. Thank you a million times over to my parents, Susan and Tom, who instilled in me a love of learning that has carried me through all the long years and many struggles of being a student. Your unwavering love and support mean the world to me, and I will forever be grateful for the many sacrifices you made to support me in finding happiness and success in life. To my dear siblings Amy and Tommy, thank you for the laughter and mischief that always remind me how to find childlike joy in life, as well as the love and support that have carried me through hard times. And to my partner Shervin, my deepest thanks for all of your many ways of supporting me in both magical and difficult times. Thank you for opening my eyes to new ways of thinking and being, for always making me laugh, for reminding me to never stop dancing (at least until I reach 100,000), and for never forgetting the emergency chocolates.

1.0 Introduction

Learning a language involves learning both word forms and word meanings, as well as the ways in which these forms and meanings are connected (e.g., Perfetti & Hart, 2002; Rice & Tokowicz, 2019). Unfortunately for language learners, these connections are not always one-to-one – in fact, a large number of words have one-to-many mappings between forms and meanings (e.g., *bat* can mean a baseball bat or the animal bat; Eddington and Tokowicz (2015). Such *semantic ambiguity* increases processing demands on learners, and may interfere with vocabulary learning, a key building block of language (e.g., Beck, McKeown, & Kucan, 2002). Furthermore, semantic ambiguity within a language is one factor that gives rise to one type of cross-language ambiguity, also known as *translation ambiguity* (e.g., the English word drill translates to German as both Bohrer and Übung) (Degani, Prior, Eddington, Arêas da Luz Fontes, & Tokowicz, 2016). Semantic ambiguity has been studied extensively, but translation ambiguity has received comparatively less attention despite the fact that estimates for the prevalence of translation ambiguity are as high as 71%, depending on the language pair in question (e.g., Tokowicz, Rice, & Terrazas-Duarte, 2018). Given this prevalence, how ambiguity affects second language (L2) vocabulary learning is a current topic of interest for psycholinguists, L2 learners, and language instructors. The present study investigated training methods that mitigate difficulties associated with learning translation-ambiguous and semantically-ambiguous words. Improving vocabulary learning outcomes may lead to improved long-term language abilities (e.g., de Groot & van Hell, 2005), which is a valuable undertaking given the rapidly increasing number of people using more than one language to communicate on a daily basis (Commission on Language Learning, 2016).

As mentioned, semantic ambiguity has been the subject of a great deal of past research, which has identified a processing disadvantage for homonyms (words with multiple unrelated meanings; *bat-animal* vs. *bat-baseball*) relative to unambiguous words, and a processing advantage for polysemes (words with multiple related meanings; *paper-newspaper* vs. *paper-academic article*) relative to unambiguous words and homonyms (e.g., Armstrong, Beekhuizen, Rice, Milic, & Stevenson, 2018; Armstrong & Plaut, 2016; Rice, Tokowicz, Fraundorf, & Liburd, 2019; Rodd, Gaskell, & Marslen-Wilson, 2002; Tokowicz & Kroll, 2007).

Translation ambiguity arises when a word in one language maps to multiple words in another language. For example, the English *drill* translates to two German words: *Bohrer* (tool for making holes in objects) and *Übung* (repetitive training method). Just as certain types of semantically-ambiguous words slow first language (L1) processing, translation ambiguity presents a challenge for L2 learners. A number of studies have demonstrated that translation ambiguity leads to slower translation production and recognition for bilingual speakers (e.g., Eddington & Tokowicz, 2013; Laxén & Lavaur, 2010; see review in Tokowicz, 2014; Tokowicz & Kroll, 2007). Of particular interest to the present work, translation ambiguity hinders L2 vocabulary acquisition in adult learners. For example, Degani and Tokowicz (2010) taught native English speakers translation-ambiguous Dutch words, and tested translation production and recognition immediately after testing, and after short and long-term delays. Translation-ambiguous words were produced and recognized less accurately than unambiguous words at all three time points.

This dissertation investigates how translation-ambiguity (Experiment 1) and semantic ambiguity (Experiments 2 and 3) affect L1 and L2 vocabulary learning in adult learners. We next turn to a description of several theoretical models of ambiguity that form the foundation of the present work.

1.1 Modeling semantic ambiguity effects

Why might semantic and translation ambiguity effects arise? One theoretical explanation for within-language semantic ambiguity effects, the *semantic settling dynamics* hypothesis (SSD), proposes that the main determinant of ambiguity effects is how much semantic processing has occurred (e.g., Armstrong, 2012; Armstrong & Plaut, 2016; see Figure 1). According to this model, semantic processing begins when a word is read and excitatory activation gradually increases as multiple potential meanings of the word become activated. Over time this excitatory activation transitions to inhibitory feedback as the reader settles on a single contextually-appropriate meaning. Early excitatory feedback benefits polysemes because their multiple senses share overlapping meaning features, and this cooperative activation points towards the same word without a need for inhibitory feedback. Unambiguous word processing initially proceeds more slowly than polysemous word processing because unambiguous words have a more limited set of meaning features contributing activation. However, unambiguous words, like polysemes, do not need to rely on inhibitory activation to select one meaning over another, and so have an advantage over homonyms. Homonyms suffer from competition due to having multiple unrelated meaning features, and so tend to be recognized more slowly than polysemes and unambiguous words. Although the SSD was developed to explain semantic ambiguity processing, these dynamics may also underlie the learning of translation-ambiguous words. The SSD framework conceptualizes learning as the amount of error in the system—when error rates are higher the system has more to learn. Examining differences in error rates between types of ambiguous words on our dependent measures may therefore be informative. The proposed study investigates this by conducting three parallel experiments, one with translation-ambiguous words (Experiment 1) and the other two with semantically-ambiguous L1 words (Experiments 2 and 3). Finding that error rates are similar for

ambiguous words in L1 and L2, and furthermore that these error rates differ in similar ways depending on ambiguity type may provide evidence that models of semantic ambiguity processing, including the SSD, can be extended to help us understand the learning of ambiguous words. However, because this model was not designed for this purpose, in the General Discussion we describe another model (the instance-based framework; Bolger, Balass, Landen, & Perfetti, 2008) that offers an alternative account of word learning.

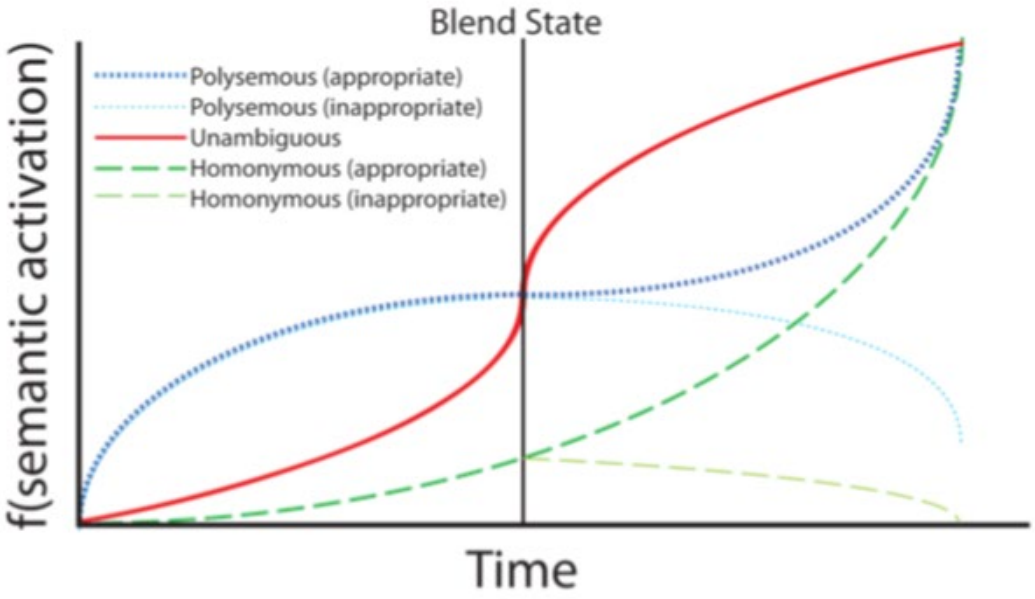

Figure 1. Predicted semantic activation over time as a function of ambiguity, according to the Semantic Settling Dynamics account (Armstrong, 2012)

1.2 Theories of L2 vocabulary learning and translation ambiguity

Despite the tremendous number of studies of L2 vocabulary learning, there are still relatively few theoretical accounts describing the cognitive mechanisms underlying L2 vocabulary

learning, and even fewer that provide a theoretical framework for translation-ambiguous word learning. Of the few theoretical models of L2 vocabulary learning, the Revised Hierarchical Model (RHM; Kroll & Stewart, 1994) and its adaptations (RHM-RER, Rice & Tokowicz, 2019; RHM-TA; Eddington & Tokowicz, 2013) have provided useful frameworks for understanding L2 vocabulary learning and translation ambiguity. Because of the similarity of these models we discuss both in the following section.

The RHM combined two earlier models of L2 word learning: the Word Association and Concept Mediation Models proposed by Potter, So, von Eckardt, and Feldman (1984). The RHM was developed to account for data showing asymmetries between L1-L2 and L2-L1 translation. This model proposed a pattern of connections between L1 and L2 word forms and meanings that varied in strength and direction. The critical proposals of this model are: 1) L1 forms are strongly connected to meaning representations, whereas L2 forms are only weakly connected to meaning representations, at least for beginning learners, and 2) L1 forms are only weakly connected to L2 forms, but L2 forms are strongly connected to L1 forms (Kroll & Stewart, 1994).

Rice and Tokowicz (2019) developed an updated version of the RHM, the RHM-RER. This model was developed to synthesize a multitude of findings from L2 vocabulary learning studies, and furthermore to apply the principles of the RHM directly to vocabulary learning; this contrasts with the original focus of the RHM, which was on production tasks rather than vocabulary learning. The RHM-RER focused on three methods of strengthening connections between L1 and L2 word forms and meanings: repetition, elaboration, and retrieval (see Figure 2). This model predicts that training methods that incorporate both form and meaning representations will be more successful than training methods that only focus only on one of those two representations. Furthermore, the model predicts that training methods that use a combination of

the three strengthening techniques (repetition, elaboration, retrieval) will be more successful than training methods that use one or none of the strengthening mechanisms.

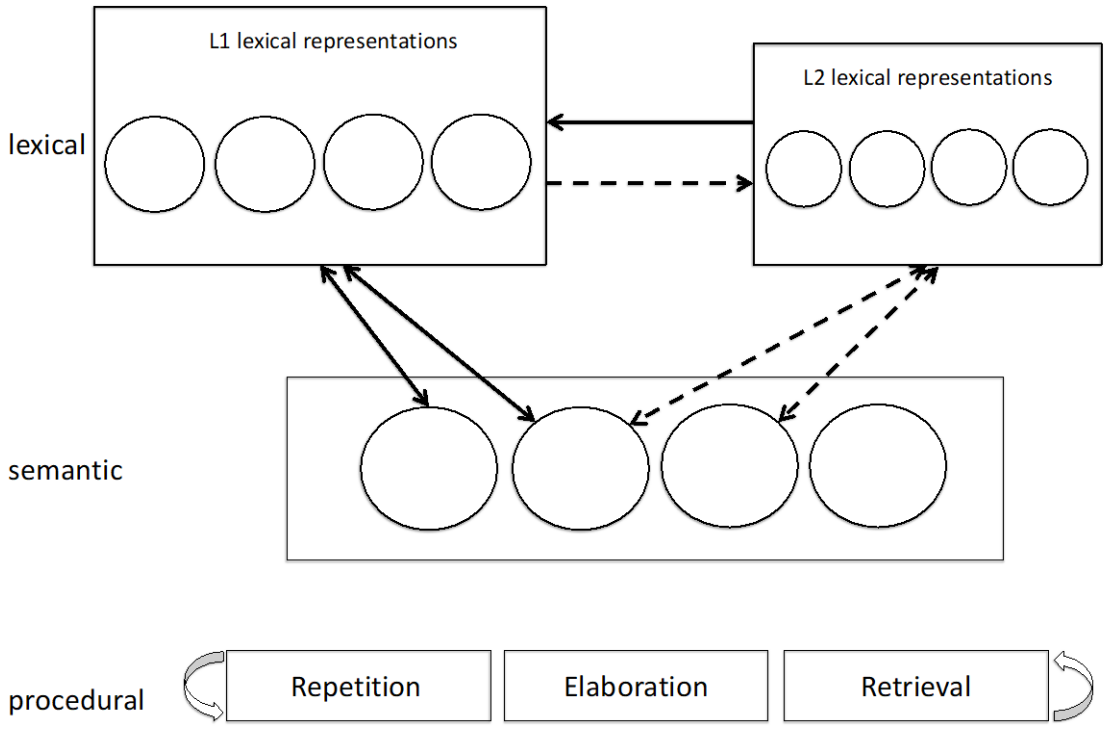

Figure 2. The RHM-RER model of L2 vocabulary learning (Rice & Tokowicz, 2019)

The RHM also formed the basis of the Revised Hierarchical Model of Translation Ambiguity (RHM-TA; Eddington & Tokowicz, 2013). This model made predictions about both form ambiguous (e.g., one English word translated to two different Dutch words that shared a meaning) and meaning ambiguous words (e.g., one English word translated to two different Dutch words that had different meanings), but because the current studies only examine meaning ambiguity, we restrict our discussion of the RHM-TA to its predictions for this type of word. In regard to this type of ambiguous word, the RHM-TA describes the pattern of interconnections between the L1 and L2 word forms and meaning representations. Note that this discussion will describe translation ambiguity in the L1-L2 direction as pictured in Figure 3, but could also be

described in the reverse direction. In this case, there is a single form representation for the L1 word, and this is weakly connected to multiple L2 form representations. Conversely, the L2 form representations are strongly connected to the L1 form representation. The L1 form and meaning representations are strongly bidirectionally connected to each other, but the L2 form and meaning connections are weakly bidirectionally connected to each other. Therefore, this model predicts that the one-to-many connections will slow translation, but that when sufficient context is provided so that one contextually-appropriate meaning can be selected the mapping will function as a one-to-one mapping and translation will become faster.

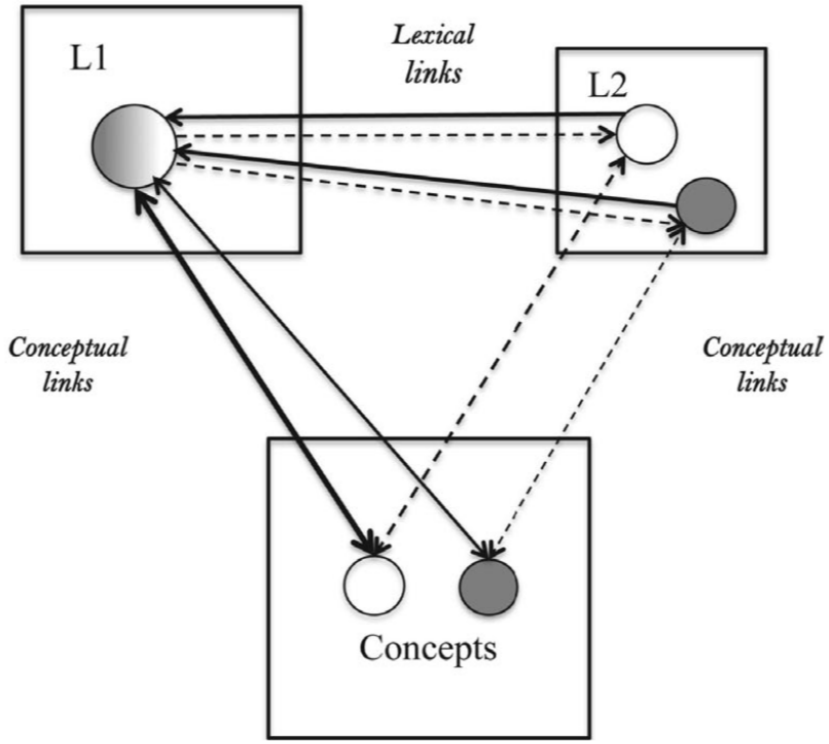

Figure 3. The RHM-TA for meaning-ambiguous words (Eddington & Tokowicz, 2013)

1.3 The translation-ambiguity disadvantage

With the theoretical framework of the RHM in mind, we now turn to a discussion of the translation-ambiguity disadvantage and then to a discussion of several instructional methods that may mitigate this disadvantage. The translation-ambiguity disadvantage describes the finding that words that have multiple translations across languages are generally more difficult to learn than translation-unambiguous words (Tokowicz, 2014), and a number of studies have demonstrated that translation ambiguity leads to slower translation production and recognition for bilingual speakers (Eddington & Tokowicz, 2013; Laxén & Lavaur, 2010; Tokowicz & Kroll, 2007).

Early studies of the translation-ambiguity disadvantage examined how ambiguous words were recognized and processed. For example, Tokowicz and Kroll (2007) reported that Spanish-English bilinguals were slower to produce translations for words with multiple meanings than for words with one meaning. Similarly, Laxén and Lavaur (2010) reported that bilinguals are slower and less able to recognize the correct translation of a translation-ambiguous word than a translation-unambiguous word.

Degani and Tokowicz (2010) were the first to investigate how translation ambiguity affects novel word learning. In this study, Degani and Tokowicz taught native English speakers Dutch-English word pairs that were translation-ambiguous and translation-unambiguous. Translation ambiguous words fell into one of two types: form-ambiguous translation pairs and meaning-ambiguous translation pairs. They examined performance on translation recognition and L2-L1 translation production tasks immediately after testing, a short delay, and a longer delay. They reported that translation-ambiguous words were translated more slowly and less accurately on both immediate and delayed translation production tests. Furthermore, the ambiguity disadvantage was more pronounced for form-ambiguous than meaning-ambiguous words, possibly because learning

to map two translations to a single meaning (i.e., form ambiguity) is a harder task than learning to map one translation to multiple meanings (i.e., meaning ambiguity).

As highlighted by Degani and Tokowicz (2010), there are multiple ways in which translation ambiguity arises. For example, one way in which translation ambiguity occurs is when words do not have direct translation equivalents in another language (for example, the German word *Schadenfreude* does not translate to any one word in English, but rather encompasses several English translations such as *gloating*, *spitefulness*, and *glee*). Additionally, translation ambiguity can be due to either lexical or semantic ambiguity, or even a combination of lexical and semantic ambiguity. Lexical ambiguity (synonymy or near-synonymy) within a language accounts for the majority of cases of translation ambiguity (e.g., Tseng, Chang, & Tokowicz, 2014; Tokowicz, Kroll, de Groot, & van Hell, 2002). For example, the English word *shy* translates to *schüchtern* and *scheu* in German, both of which represent the same meaning. However, translation ambiguity can also arise from semantic ambiguity within a language. In this case, a word that has multiple meanings in one language, such as the English word *bark*, also has multiple translations across languages to encompass these meanings (i.e., in German the tree bark meaning translates to *Baumrinde*, and the bark of a dog meaning translates to *Bellen*). It is also possible for both lexical and semantic ambiguity in one language to give rise to translation ambiguity, such as how the English word *bank* can denote either the river bank meaning (which translates to both *Ufer* and *Böschung* in German) or the financial institution meaning (which translates to both *Kasse* and *Bank* in German). To further complicate matters, these sources of translation ambiguity can occur in both directions (from L1-L2 and L2-L1).

These variations in the source of translation ambiguity have important implications for language learning and processing. Tokowicz et al. (2002) collected number-of-translation norms

for Dutch-English translation pairs (around 25% of which were translation ambiguous), and reported that the number of translations was associated with the semantic similarity of the translations, such that semantic similarity was lower for words that had more translations. This suggested that translation ambiguity is both relatively common, and also that there is variability in how closely the multiple translations of translation ambiguous words capture the meaning of the source word. Given that the number of translations that a word has is known to affect how quickly words are learned and processed (e.g., Kroll & Tokowicz, 2001; Tokowicz & Kroll, 2007), it is also important to understand factors that are associated with the number of translations.

Tseng et al. (2014) collected English-Mandarin number-of-translation norms using English words from the Dutch-English translation pairs from Tokowicz et al. (2002). They reported that 67% of word pairs were translation-ambiguous, as compared to only around 25% of the words in the Dutch-English norms. Interestingly, there were significant correlations between the English-Mandarin number of translations, the number of translations in English-Dutch norms (Tokowicz et al., 2002), as well as the number of translations for a set of English-German norms collected by Eddington & Tokowicz, 2013). Tseng et al. interpreted this as evidence that the English words shared some aspect that made them more likely to be translation-ambiguous, even when translated across different language pairs.

As described above, Degani and Tokowicz (2010) demonstrated that translation ambiguity that arises from lexical ambiguity (which they referred to as form-ambiguity) has different effects on translation recognition and production than translation-ambiguity that arises from semantic ambiguity. A goal of the present study is to investigate the effects of *semantic* ambiguity in both L1 and L2, and therefore the translation-ambiguous word pairs we use in Experiment 1 are exclusively words that are translation-ambiguous due to L1 semantic ambiguity. These words were

selected to best answer the present questions, but we remind the reader that there are many other sources of translation-ambiguity beyond this specific type.

1.3.1 Mitigating the translation-ambiguity disadvantage

One line of research following from the findings of Degani and Tokowicz (2010) has focused on instructional manipulations that encourage learners to establish appropriate form-meaning mappings from initial encounters with a word. For example, Degani, Tseng, and Tokowicz (2014) taught native English speakers Dutch-English word pairs that were form- and meaning-ambiguous and unambiguous. The multiple translations of the ambiguous words were either taught in the same training session, or taught in separate sessions. They replicated the translation-ambiguity disadvantage for both immediate and delayed testing, but found that this disadvantage could be offset somewhat by teaching both translations in the same session as opposed to separate sessions.

In contrast to studies that seek to strengthen form-meaning connections, a second line of inquiry into mitigating the translation-ambiguity disadvantage, proposed here for the first time, is strengthening the meaning representations of ambiguous words. Past research has demonstrated that strengthening meaning representations is critical to learning L2 vocabulary (e.g., Coomber, Ramstad, & Sheets, 1986; Craik & Lockhart, 1972; Rice & Tokowicz, 2019), but this has not yet been explored with translation-ambiguous words. Because ambiguous words are characterized by multiple meaning representations, increasing semantic activation during learning may have varying effects depending on the semantic relatedness of the meanings. For instance, simultaneously activating the unrelated meanings of homonyms may lead to competition and make learning more difficult, whereas simultaneously activating the related meanings of polysemes may

lead to cooperation and facilitate learning. If true, this would be helpful information for psycholinguists and educators, so that semantic training methods could be used carefully with semantically-ambiguous words. One method of strengthening semantic representations is through generation, which will be discussed in the next section.

1.4 Theories of generation effects

The *generation effect* describes the finding that learner-generated items are remembered better than read items (Slamecka & Graf, 1978). Generation has been shown to encourage activation of a greater number of meaning features than reading (Hirschman & Bjork, 1988), and this additional semantic activation during learning may facilitate later retrieval (McElroy & Slamecka, 1982). The benefits of generation have been reported with a wide variety of materials, including word lists (e.g., Slamecka & Graf, 1978), nonsense words (McElroy & Slamecka, 1982; Nairne & Widner, Jr., 1987), numbers (e.g., Gardiner & Rowley, 1984), and mathematical equations (e.g., McNamara & Healy, 2000) (for a meta-analysis of this effect, see: Bertsch, Pesta, Wiscott, & McDaniel, 2007). In the current study, we investigate sentence generation, a method of generation used by Eddington, Martin, and Tokowicz (2012) and Tokowicz and Jarbo (2009). This method requires learners to write a meaningful sentence that includes a target word. This process requires a learner to engage with the meaning of a word, thereby activating related meaning features and increasing semantic activation.

There are a number of theoretical explanations of the generation effect that propose a wide range of cognitive mechanisms for this effect. These include distinctiveness (Begg, Vinski, Frankovich, & Holgate, 1991; Kinoshita, 1989), mental effort (McFarland, Warren, & Crockard,

1985), selective rehearsal of generate items over read items (Slamecka & Katsaiti, 1987), and transfer-appropriate processing (Morris, Bransford, & Franks, 1977). However, none of these explanations is fully able to account for all of the available evidence (Bertsch et al., 2007). In the present study, we investigate two theories of the generation effect: *the two-factor theory* (Hirshman & Bjork, 1988) and the *enhanced semantic processing hypothesis* (McElroy, 1987).

One of the best-known theories of generation, the two-factor theory (Hirshman & Bjork, 1988), posits that generation both strengthens connections between stimuli and responses (i.e., L1 and L2 word forms in the context of the current study), and also enhances activation of semantic features (i.e., meaning representations). This theory assumes that any word learning activity, such as sentence reading or sentence generation, activates a certain number of lexical and semantic features associated with the key word or concept. According to this theory, generation effects arise from the combination of two observations: 1) generation activates more semantic features than sentence reading, and 2) generation strengthens the association between a stimulus and a response. This theory maps nicely onto the RHM framework, and in the context of the RHM this would mean that generation both strengthens form-form and form-meaning connections, and also strengthens meaning representations. In the context of the current study, we hypothesize that sentence generation will strengthen L1 form-L2 form connections as well as form-meaning connections during novel word learning, which will make word forms and meanings easier to retrieve during future encounters.

This theory incorporates elements of an older theory of generation effects, the *enhanced semantic processing hypothesis* (McElroy, 1987). This hypothesis proposed that the act of generation encourages meaning access (McElroy, 1987), which leads to greater semantic processing during encoding, and greater activation of associated semantic features to aid in later

retrieval. The two-factor theory incorporates this view, but extends the theory to state that generation not only enhances semantic processing, but it also enhances the connection between stimuli and responses (i.e., word forms and meanings). If we find generation effects in tasks that require strong connections between word forms, as does translation production, then we can interpret this as evidence to support the two-factor theory. If we find evidence that generation effects only emerge in tasks that require substantial meaning access, such as meaning production, this would support both the enhanced semantic processing account of generation effects and the two-factor theory. If we find that generation effects are only found in tasks that require meaning access (such as meaning production), but not in tasks that require only form representations (such as free recall) or primarily form-form connections (such as translation production), this would support the enhanced semantic processing theory but not the two-factor theory.

One important difference between the studies from which the *two-factor theory* and the *enhanced semantic processing hypothesis* were developed and the present set of experiments is the type of generation task. The *two-factor theory* and the *enhanced semantic processing hypothesis* were developed based on the results of experiments that used generation tasks that required learners to generate the second word of a word pair after being given a rule to follow. For example, Hirschman and Bjork (1998) presented learners with a word pair in which the first word was complete, but the second word was missing several letters, and learners were asked to generate the missing letters (i.e., word fragment completion). Similarly, McElroy (1987) presented learners with word pairs that were rhyming words or synonyms, in which the first word was complete and the second word was missing one or more internal vowels, and learners were asked to generate the missing letters. In contrast, the present study uses a sentence generation task in which learners are required to consider the meaning of a target word and write a semantically-appropriate sentence

that contains the target word. These types of generation are quite different, and so we briefly turn to a discussion of whether the predictions of models of generation based on word fragment completion tasks can be expected to hold for sentence generation tasks.

There is ample evidence that sentence generation engages cognitive mechanisms that are also engaged by word fragment completion generation tasks. Sentence generation tasks have been used successfully by a number of researchers who report generation effects that are in line with the type of generation effects we would expect to see based on early examinations of this effect by Hirschman and Bjork (1998) and McElroy (1987). For example, Webb (2005) demonstrated that a sentence generation task was more effective than a sentence reading task for learning L1 Japanese-L2 English word pairs, as long as the amount of time spent studying was equivalent for each method. Similarly, Tokowicz and Jarbo (2009) found that sentence generation was more effective than sentence reading for learning Dutch-English word pairs, and Eddington, Martin, and Tokowicz (2012) reported that sentence generation was more effective than sentence reading for learning German-English word pairs. Thus, we have ample reason to expect that sentence generation tasks will produce generation effects in line with those reported by studies of word fragment completion generation tasks, although this will be the first explicit attempt to extend the *two-factor theory* and the *enhanced semantic processing hypothesis* to sentence generation tasks.

1.5 Overview of experiments

The proposed experiments will investigate the use of generation during L1 and L2 vocabulary learning with ambiguous and unambiguous words. They will furthermore extend the SSD hypothesis to vocabulary learning for the first time, and test whether this extension makes

accurate predictions about how different types of ambiguous and unambiguous words are learned. The two-factor theory of generation effects proposes that the act of generation encourages meaning access, which leads to greater semantic processing during encoding, and greater activation of associated semantic features to aid in later retrieval. If this account is correct, we would expect to find differences between generated and read items such that generated items are retrieved more accurately than read items. The SSD hypothesis, reviewed in more detail above, proposes that as semantic processing increases, differential processing dynamics emerge depending on the type of ambiguity, such that polysemes benefit from cooperative activation and homonyms suffer from competitive activation. If generation increases semantic processing, and if this increased semantic processing leads to the temporal dynamics described by the SSD hypothesis for ambiguous words, we would then expect a benefit of generation for polysemes and unambiguous words, and no benefit of generation for homonyms. Finally, in all three experiments we expect generation effects and ambiguity effects to be impacted by two individual difference measures: working memory (WM) and inhibitory control, and we also expect that best performance on all measures will be observed for participants who are both high in WM and inhibitory control (Michael, Tokowicz, Degani, & Smith, 2011).

1.5.1 Description of proposed research

The proposed research aims to investigate, for the first time, the effects of encouraging semantic processing via generation for learning semantically-ambiguous and translation-ambiguous words.

In Experiment 1, we examined how the use of sentence generation vs. sentence reading affected the learning of German-English word pairs that included unambiguous, homonymous,

and polysemous pairs. We furthermore investigated how the translation-ambiguity and generation effects are impacted by individual differences in inhibitory control, as measured by the Simon task. We examined outcomes on a free recall task and an oral L1–L2 translation production task immediately after learning and after a one-week delay.

Experiment 2 examined the impact of sentence generation vs. sentence reading in L1 vocabulary learning. We taught native English speakers rare words and their definitions, using a stimulus set that comprised unambiguous words, homonyms, and polysemes. We investigated how ambiguity and generation, as well as individual difference measures, affected performance on free recall and meaning production tasks both immediately after learning and after a one-week delay.

In Experiment 3, we followed up the results of Experiment 2, and asked whether the use of definitions or context sentences during L1 rare word learning affected ambiguous and unambiguous words differently, and whether individual differences or the use of sentence generation vs. sentence reading interacted with the type of training material to which learners were exposed. We investigated performance on free recall, sentence completion, and meaning production tasks both immediately after learning and after a one-week delay.

In summary, in three experiments, this project asks the following questions:

1. Is generation more effective than reading for learning L1 and L2 ambiguous and unambiguous words, both immediately after testing and after a one-week delay?
2. Are generation effects similar for cross-language and within-language stimuli, and similar or different across different tasks?
3. Is generation more or less effective for certain types of ambiguous words (i.e., homonyms vs. polysemes)?

4. Do individual differences in WM and inhibitory control impact learning of ambiguous words or the generation effect?
5. Does the use of definitions in addition to context sentences during learning lead to better learning outcomes than the use of context sentences for semantically-ambiguous words, and is this effect modified by generation or type of ambiguity?

2.0 Experiment 1

The translation-ambiguity disadvantage describes the general finding that words that have multiple translations across languages are more difficult to learn (e.g., Degani & Tokowicz, 2010). Since the discovery of this disadvantage, a number of studies have investigated methods of mitigating or offsetting the effects of the translation-ambiguity advantage (Degani, Tseng, & Tokowicz, 2014; Ekves, 2014; Rice, Ekves, & Tokowicz, 2017). Recent work has observed that in textbooks and classrooms the multiple translations of words are often taught separately. For example, a first translation of a word might be introduced many chapters before the second translation, but this separation is actually disadvantageous for learning translation-ambiguous words (e.g., Degani, Tseng, & Tokowicz, 2014). Studies following up this line of inquiry have focused on allowing a learner to establish the correct form-meaning mappings from an initial encounter with a word. In the current study, we propose a second and distinct line of inquiry: whether using training methods that enhance the meaning representations of ambiguous words will offset the translation-ambiguity disadvantage. Additionally, this experiment also investigates the role of individual differences in L2 vocabulary learning, especially because individual differences may impact the success of training methods that aim to enhance meaning representations. We begin with a discussion of past research investigating generation effects in L2 vocabulary learning, and then turn to a discussion of past investigations of individual differences in L2 vocabulary learning.

2.1 Generation effects in L2 vocabulary learning

Although there is a vast literature investigating generation effects (e.g., Bertsch et al., 2007; Gardiner & Rowley, 1984; Lutz, Briggs, & Cain, 2003; McNamara & Healy, 2000; Pesta, Sanders, & Murphy, 1999; Slamecka & Graf, 1978), and generation effects in language learning (e.g., Basi, Thomas, & Wang, 1997; Begg et al., 1991; Johns & Swanson, 1988; McElroy & Slamecka, 1982a; Mulligan, 2002; O'Neill, Roy, & Tremblay, 1993; Payne, Neely, & Burns, 1986; Slamecka & Katsaiti, 1987), thus far there are only a small handful of studies that have investigated generation effects in adult L2 vocabulary learning (Barcroft, 2009; Coomber, Ramstad, & Sheets, 1986; Eddington, Martin, & Tokowicz, 2012; Tokowicz & Jarbo, 2009). This question is of interest because L2 vocabulary learning proceeds differently than L1 vocabulary learning (e.g., Lotto & de Groot, 1998; Tan et al., 2003), and therefore it is possible that generation effects are more or less impactful for L1 and L2 linguistic stimuli.

We first examine a study by Coomber, Ramstad, and Sheets (1986) that investigated the use of sentence generation during L2 vocabulary learning, and reported a benefit of enhancing semantic processing via generation for L2 vocabulary learning. In this experiment, the authors created an artificial vocabulary, and taught English-speaking university students novel words using one of three training methods: definition matching, example matching, and sentence generation. The definition condition required learners to match novel words with their definitions, the example matching condition required learners to select an example of how a word should be used from a list of word-example pairs, and the sentence generation condition asked participants to generate two semantically meaningful sentences that contained a target word. They reported the most accurate performance across several outcome measures for words trained in the sentence

generation condition. This experiment demonstrated that the use of sentence generation during L2 vocabulary learning can be a highly beneficial learning strategy.

An additional study in this area was conducted by Tokowicz and Jarbo (2009), who investigated whether generation improves L2 vocabulary learning in beginning adult learners. They trained native English speakers on Dutch-English word pairs, half of which were trained by reading the word pair and half of which were trained by generating a meaningful English sentence with the Dutch word inserted. Furthermore, to investigate the utility of providing definitions during training, half of the sentences in each read or generate condition were presented with a definition and half without a definition. Results showed that generation conditions led to marginally higher accuracy than read conditions on a semantic relatedness task, and that words trained without a definition but with generation led to the highest accuracy on tests of free recall and translation production. These results provide preliminary evidence that generation may helpful for learning L2 vocabulary. However, all word pairs in this study were unambiguous, and so it remains an open question whether ambiguous L2 words respond to generation in the same way as unambiguous words.

Extending this research, Eddington, Martin, and Tokowicz (2012) trained native English speakers on unambiguous German-English word pairs in four conditions: reading a definition, reading a sentence, generating a definition, and generating a sentence. They found a benefit of generation over reading for free recall a week after training. Furthermore, generating a sentence was the only condition that led to significantly-improved semantic processing of the trained words as measured by a semantic generation task. These results provide additional evidence that generation can be successful for training L2 vocabulary, and furthermore show that sentence

generation promotes greater semantic processing than definition generation, although it is still unknown if these effects would hold for ambiguous words.

However, not all studies of generation for L2 vocabulary learning have reported positive outcomes. Barcroft (2009) investigated the use of synonym generation during L2 vocabulary learning, by teaching a group of native Spanish speakers novel English words and instructing half of the learners to generate L1 synonyms for target words, whereas the other half were simply instructed to learn the target pairs. Performance on translation production tasks suggested that synonym generation actually negatively impacted learner outcomes. However, this result may have been because the training and testing tasks were not congruent in their demands (see transfer-appropriate processing model; Morris et al., 1977).

2.2 Generation effects with translation-ambiguous words

Of the handful of studies that have investigated generation effects with L2 vocabulary learning, only one study has ever investigated generation effects with translation-ambiguous words. However, even this study did not directly test whether the effectiveness of generation varies depending on the type of ambiguity (an aim of the present study). In this section we review this study, and discuss how it forms the basis of the current work.

The only study of generation effects with translation-ambiguous and unambiguous L2 words was conducted by Eddington (2015; Experiment 2). In this experiment, native English speakers were trained on German-English word pairs, half of which were translation-ambiguous, and half of which were translation-unambiguous. A sentence-generation method was employed during training, in which participants were asked to read a German-English word pair and a

definition, and write an English sentence that contained the German word. An underlying assumption of this experiment was that generation would benefit all words, regardless of ambiguity status. However, this idea had yet to be empirically tested.

After training, participants in Eddington (2015) completed a free recall task and an L2-L1 translation production task. Results for free recall showed a translation-ambiguity disadvantage, which was unexpected given past reports of a translation-ambiguity advantage in free recall (Degani et al., 2014; Ekves, 2014). Translation production results showed a translation-ambiguity disadvantage for homonyms but not polysemes in RT, and no overall ambiguity effects in accuracy analyses. Overall, the results of this study were surprising in that they only found the expected translation-ambiguity disadvantage for homonymous words, and either no translation ambiguity effect or a translation-ambiguity advantage for polysemous words. Eddington hypothesized that these results might be due to the use of the sentence generation training method, because this method of training might have emphasized the differences in meaning for homonyms, and created inhibitory activation. Therefore, a goal of the present study is to further investigate this claim, and determine if the use of sentence generation during learning may impact translation-ambiguous and translation-unambiguous words in previously-unknown ways. Whereas Eddington trained all words with the sentence generation method, the present study will test the effects of both sentence generation and the more traditional training method of reading experimenter-generated context sentences.

2.3 Individual differences in L2 vocabulary learning

Learning translation-ambiguous words poses unique challenges for learners to overcome. For instance, translation ambiguity requires learners to understand complex mappings between L1 and L2 word forms and meanings. In the case of meaning-ambiguous words, a given L1 form can translate to multiple L2 meanings that are represented by multiple L2 word forms (or vice versa from L2 to L1). To further complicate the situation, the multiple meanings vary in how related they are; whereas polyseme meanings are related and may intuitively make sense to the learner, homonym meanings are unrelated and may pose a particular challenge to leaners because it is not often obvious why two unrelated meanings correspond to the same word form. Whereas learning unambiguous words draws on general cognitive skills such as word knowledge (Elgort et al., 2015), and WM (for a review, see: Linck, Osthus, Koeth, & Bunting, 2014), learning ambiguous words may require cognitive processes in addition to WM, such as inhibitory control. Individuals vary in their cognitive abilities, and we predict that variation in WM and inhibitory control will impact the ability to learn translation ambiguous words in different ways than unambiguous words.

In this section, we review what is known about individual differences in L2 vocabulary learning. To begin, we turn to a meta-analysis of the role of WM in L2 learning conducted by Linck et al. (2014). This meta-analysis examined 79 studies that investigated different aspects of WM in L2 learning, and reported a substantial estimated population effect size of .26. In other words, across a wide range of outcome measures, WM was found to be a robust predictor of L2 outcomes. Interestingly, one key finding was that measures of WM that focused on executive processes were better predictors of L2 outcomes than WM measures that were more concerned with maintaining an active memory representation. However, this meta-analysis included studies of L2 processing and production, and was not specific to L2 vocabulary learning. We next examine

studies of L2 vocabulary learning that investigate both of these components (maintaining an active memory representation and executive processing).

Many L2 vocabulary learning studies have examined the role of WM processes concerned with building and maintaining active representations. For example, Martin and Ellis (2012) investigated the role of phonological short-term memory and WM when learners were acquiring novel vocabulary in an artificial language. They reported that, even after accounting for variance in vocabulary learning outcomes due to phonological short-term memory, WM capacity still accounted for significant variance in L2 word learning.

Tokowicz, Michael, and Kroll (2004) reported that individuals with higher WM capacity were better able to use more effective L2 communication strategies than individuals with lower WM capacity. Similarly, Michael et al. (2011) investigated WM (as measured by an operation span task) and executive control processes (as measured by the Stroop task) in native English speakers who were learning L2 Spanish, which included translation-ambiguous and translation-unambiguous words. They investigated performance on translation production tasks, and reported that individuals who were better able to inhibit task-irrelevant information were able to translate word more accurately than individuals who had lesser inhibitory control, but interestingly this finding was qualified by an interaction with WM, such that participants with a higher WM span and less Stroop interference did better than lower WM span participants and higher WM span participants with greater Stroop interference. The current experiment will investigate whether WM and inhibitory control may interact during word learning in the current experiment. Specifically, we predict that learners who are higher in WM and also better able to exert inhibitory control will display a reduced translation-ambiguity disadvantage, and this will be especially pronounced for homonyms.

2.4 Experiment overview

Experiment 1 examined how the use of generation during vocabulary training impacted the learning of translation ambiguous German-English word pairs. Participants were trained on German-English word pairs and definitions, and then instructed to practice this material by either reading and retyping an English context sentence containing a target German word *(read condition)*, or by generating their own English context sentence containing a target German word *(generate condition)*. Stimuli consisted of unambiguous words, homonyms, and polysemes. We examined how well the word pairs and their meanings were learned by testing participants on free recall immediately after training and oral L1-L2 translation production immediately after training and again after a one-week delay. Our specific research questions were: 1) is generation beneficial for learning translation-ambiguous words, 2) does the generation effect differ depending on word type (homonyms, polysemes, or unambiguous words), and 3) is generation more or less effective for learners of varied cognitive profiles.

Based on previous research (Slamecka & Graf, 1978), we expected to find a generation effect such that German-English word pairs learned using generation would be remembered more quickly and accurately on free recall and translation production tasks than German-English word pairs learned by reading and repetition. Furthermore, we expected that the strength of the generation effect would vary depending on ambiguity type. Specifically, we predicted that polysemes and unambiguous words would benefit from generation, whereas homonyms would be negatively affected by generation. We predicted this would be the case because when participants are asked to generate semantic information for the multiple meanings of ambiguous words, the related polysemous senses share semantic overlap would lead to cooperative activation and

enhances memory whereas the unrelated homonymous meanings generate competitive activation that leads to memory inhibition.

Additionally, we predicted that individual differences in inhibitory control and WM would modulate the translation-ambiguity effect, during learning or during testing, or possibly during both. Therefore, at the end of each experiment, participants completed two individual difference measures: the Waters Reading Span test (Waters & Caplan, 1996) and the Simon task (Simon & Wolf, 1963). The Waters task measures WM, which is known to correlate with L2 vocabulary ability as well as how able learners are to have multiple word meanings active simultaneously (e.g., Michael et al., 2011). The Simon task measures learners' ability to suppress task-irrelevant information, which may assist learners in selecting the relevant meaning of an ambiguous word while suppressing irrelevant meanings (e.g., Michael et al., 2011). We predicted that individuals who are higher in inhibitory control, as measured by the Simon task (Simon & Wolf, 1963), would be better able to inhibit competition from irrelevant meanings of ambiguous words and focus on the relevant meanings. We predicted this would result in a reduced homonymy disadvantage (as measured by higher accuracy or faster response times to homonyms) relative to individuals who are lower in inhibitory control. Similarly, we predicted that individuals with a higher WM span, as measured by the Waters Reading Span test (Waters & Caplan, 1996), would be better able to store and process multiple meanings for a word, and thus will show a reduced translation-ambiguity disadvantage relative to individuals with low WM span. Furthermore, based on the results of Michael et al. (2011), we predict that WM and inhibitory control may interact with word type, such that we will observe the best performance on our outcomes measures for individuals that have higher WM and also better inhibitory control.

2.5 Methods

2.5.1 Participants

Participants for this experiment were 28 native English speakers, 18 years and older, with no prior knowledge of German or Dutch. All participants were right-handed, with normal or corrected-to-normal vision, and were recruited from the Psychology Department Subject Pool and compensated with class credit and a $7.00 cash bonus if they completed both sessions. Participation time was approximately 2 hours total, divided between two sessions.

2.5.2 Design

This study used a 3 word type (unambiguous, homonym, polyseme) x 2 training condition (sentence reading, sentence generation) x 2 session (session 1, session 2) within-subjects design.

2.5.3 Stimuli

Experiment 1 used English-German word pairs: 50 English words and 70 corresponding German translations and definitions selected from a set collected by Eddington (2015). This subset contained 30 unambiguous, 10 polysemous, and 10 homonymous English words and their German translations, as well as their definitions. These different word types were matched on English word

length, $F(66) = 1.17$, $p = .33$, English frequency, $F(66) = 2.57$, $p = .06$,[1] and English concreteness, $F(66) = 1.02$, $p = .39$. See Table 1 for example stimuli, and Appendix A for stimuli characteristics.

Additionally, we generated context sentences for each word pair for use during vocabulary training. To ensure that context sentences accurately captured key meaning features of the words, we collected normative ratings for each sentence from a group of 10 native English speakers. Participants were presented with a short English sentence with one word missing, and asked to type in the first English word that came to mind to complete the sentence. The mean proportion of responses that included the intended English target word was calculated for each sentence. The proportion of sentence completions that included the target word was .65 ($SD = .31$). Appendix B describes the full norming procedures for these sentences and presents the normative ratings. Sentence ratings were matched across word types, $F(66) = 0.15$, $p = .93$.

Table 1. Example stimuli for Experiment 1

English	German	Word type	Definition
drill	Bohrer	homonym	A shaft-like object for making holes in firm materials
drill	Übung	homonym	Any strict, methodical, repetitive, or mechanical training, instruction
atmosphere	Lufthülle	polyseme	The gaseous envelope surrounding the earth
atmosphere	Stimmung	polyseme	A general pervasive feeling
recovery	Erholung	unambiguous	Restoration to a former or better condition

[1] Because this test revealed a marginally significant result (p = .06), we included English frequency in final regression models as a control variable where appropriate.

2.5.4 Procedure

The experiment consisted of two sessions of 1 hour each, spaced exactly one week apart. On Session 1, participants received training on English-German word pairs and definitions. Following training, participants completed a free recall test in which they were instructed to type any English or German words that they remembered from training. This task both encouraged participants to use retrieval to enhance memory (the "testing effect", e.g., (Pyc & Rawson, 2010; Roediger & Karpicke, 2006), and also allowed us to measure learning immediately after training. Next, participants completed a translation production test that assessed their ability to orally produce an L2 translation when presented with a trained English word.

During Session 2, which occurred after a one-week delay, participants again completed the L1-L2 translation production task. Participants also completed two individual difference measures: the Waters Reading Span test (Waters & Caplan, 1996) and the Simon task (Simon & Wolf, 1963). Finally, participants completed a Language History Questionnaire (LHQ; Tokowicz, Michael, Kroll, 2004) to collect relevant language background information (see Appendix C for LHQ data). Table 2 provides a summary of the experimental procedures and timeline.

Table 2. Experiment 1 timeline

Session	Day	Tasks
Session 1	Day 1	Training 1: Word pairs + definition (2x per word pair) Training 2: Read or generate sentence (1x per word pair) Testing 1: Free recall Testing 2: L1-L2 translation production
Session 2	Day 8	Testing 3: L1-L2 translation production Testing 4: Individual difference measures Testing 5: Language History Questionnaire (LHQ)

2.5.4.1 Vocabulary training

Vocabulary training consisted of two parts. In the first part of training, a fixation cross appeared on the screen until the participant pressed the space bar to initiate a trial, followed by a 100 ms blank screen, and finally a German-English word pair and the definition appeared for 8000 ms. The German-English word pair appeared centered in the upper third of the screen, and the definition appeared centered below the word pair in middle and lower thirds of the screen. Participants were instructed to read and attempt to memorize the word pair and definition. Each word pair appeared twice; presentation order was randomized by E-prime. Four practice trials (words not from the training set and excluded from analyses) appeared at the beginning of training to allow participants to familiarize themselves with the task. For ambiguous words, the two translations were presented on back-to-back trials (e.g., Degani, Tseng, & Tokowicz, 2014). In the second part of training, a fixation cross appeared on the screen until the participant pressed the space bar to initiate a trial, followed by a 100 ms blank screen, and finally a German-English word pair and definition appeared on the screen in one of two conditions: 1) with the German word embedded in an English context sentence and instructions to read and retype the sentence (*read condition*), or 2) without a context sentence, but with instructions for the participant to generate their own English sentence with the target German word embedded (*generate condition*).[2] Participants were told that generated sentences should capture the meaning of the word. Participants typed the sentences into the program, and their responses and total response times

[2] Although many previous generation tasks have used word fragment completion, we chose to follow Eddington et al. (2012) in using sentence generation because sentences are best suited to drawing participant's attention to the subtle differences in meaning of ambiguous words.

were recorded by E-Prime. There was no time limit to read or generate sentences. The second part of training also began with four practice trials to allow participants to familiarize themselves with the task. Each word pair appeared only once per meaning during this part of training. If a word pair was ambiguous, the multiple meanings were presented on back-to-back in the same training condition (i.e., both meanings were either trained in the generate condition or both in the read condition). There were four fixed training orders, in which words were counterbalanced across training condition (read vs. generate), and whether the dominant (i.e., more frequent) or subordinate (i.e., less frequent) translation of ambiguous words appeared first or second in training. Throughout a training session, the read and generate trials were randomly intermixed. Within each training order, the read and generate trials were randomly intermixed. Participants were randomly assigned to a training order. All vocabulary training occurred on the first day of the study, and took approximately 60 minutes to complete.

2.5.4.2 Free recall

Immediately following vocabulary training, participants completed a free recall task. Participants were given an Excel sheet, and instructed to type in all words that they remembered from training, even if they did not remember the full English-German word pair. They were told that it was acceptable to type partial words and guesses if they were not sure how a word was spelled. Participants were allowed unlimited time to complete this task.

A trained coder compared the free recall responses (English and German) that participants generated to the original stimuli. If all letters were present and in the correct order, a response was awarded one point. Accent marks and capitalization did not need to be correct in order to receive full credit. Partial credit (.5 points) was also awarded if at least one syllable was entered in the

correct position. Other responses were given zero points. A second independent coder verified the accuracy of the scoring.

2.5.4.3 Translation production

Immediately following the free recall task, participants were asked to complete an L1-L2 oral translation production task. The task began with a fixation cross in the center of the screen until participants pressed the space bar to initiate a trial. Next, E-prime randomly selected and presented one of the English words learned during training. Participants were instructed to say the German translation of the English word. The English word remained on the screen until the participant made a verbal response and triggered a voice key. After each response the word was replaced by a blank screen for 200 ms, and then a new fixation cross appeared. The task began with four practice trials to allow participants to familiarize themselves with the task. Responses were recorded on a digital recorder, and E-prime recorded the amount of time from initial presentation of the stimulus to the beginning of a verbal response. Participants were allowed unlimited time to make a response. If an English word had more than one German translation, participants were provided with two sequential prompts to translate the word, and they were informed that the order in which they provided the translations did not matter. Participants returned exactly one week later and completed this task a second time.

Two independent raters listened to and scored translation production responses. Coders were instructed to listen to and transcribe the German response exactly as it was said, and compare it to the original stimulus. If responses exactly matched the expected response the response was scored as correct and given one point. If the response had at least one syllable in the correct place partial credit (.5 points) was given. If the participant did not say anything, said something incorrect, or said something that indicated they did not know the response was given zero points. If a correct

response was preceded by something that may have triggered a voice key before the response was begun (such as a cough, pencil tapping, or saying a response so quietly that it needed to be repeated), it was coded as a voice key error. These voice key errors were treated as correct responses for accuracy, but not used in RT analyses. If an incorrect response was preceded by something that may have triggered a voice key before the response was begun, it was coded as a voice key error but was treated as incorrect for accuracy analyses and not used in RT analyses. The coders had a high level of agreement for both Day 1 (kappa = 0.84, $p < .0001$) and Day 2 (kappa = 0.80, $p < .0001$) (Gamer, Lemon, Fellows, & Singh, 2019). A third independent coder resolved disagreements.

2.5.4.4 Individual difference tasks

After completing translation production on Session 2 of the experiment, participants completed two individual difference measures: the Waters Reading Span test and the Simon task. The Waters Reading Span test measures working memory capacity via a linguistic task, and captures information about both memory span and language comprehension (Waters & Caplan, 1996). In the Waters test participants read 80 sentences, which were grouped into sets ranging from two to six sentences. There were 4 sets of each length (i.e., four sets of two sentences, four sets of three sentences, etc.). Trials began with a fixation cross displayed at the center of the screen for 1000 ms, which was replaced by a sentence. Participants were instructed to read the sentence and make a sensibility judgement by pressing "yes" with their right index finger if the sentence made sense to them, and "no" with their left index finger if the sentence did not make sense to them. Sentences remained on the screen for 5000 ms or until participants made a response. Response time and accuracy was recorded. Participants were instructed to remember the final word of each sentence, and after each set of sentences they were asked to type as many of the final words

as they could remember in the order in which they appeared, and press the escape key when they had typed in all responses that they remembered. Accuracy on the sensibility judgement task and accuracy and response time for the sentence final word recall task were recorded. Participant responses to the sensibility judgements were examined, and data from anyone who correctly made fewer than 70% of the sensibility judgements was removed from further analyses that used the Waters measure. This resulted in the removal of data from one participant who only responded correctly to sensibility judgements 65% of the time, leaving data from 27 participants remaining for analyses. For the sentence final word analyses, words were considered correct if they were spelled accurately, or if they contained minor typos that did not change the response to another orthographically correct word. For example, *intelligeince* for *intelligence* was considered correct, but alternate forms of a word were considered incorrect (e.g., incorrect tense and pluralization, such as *cat* for *cats*). Set size span was calculated as the largest set size for which a participant correctly recalled all sentence-final words for at least two of the four sets of that length (e.g., Tseng, Doppelt, & Tokowicz, 2018). Participants who did not recall all words in at least two of the four sets of size two were assigned a set size span of zero. In addition to set size span, total span (the total number of words recalled for the entire task) was also calculated. Total span was highly, but not perfectly, correlated with set size span (r = .69).

As mentioned above, the Simon task measures ability to suppress task-irrelevant information, which may assist learners in selecting the relevant meaning of an ambiguous word while suppressing irrelevant meanings (e.g., Michael et al., 2011; Yudes, Macizo, & Bajo, 2011). In this test participants saw a blank screen for 850 ms, followed by a fixation cross at the center of the screen for 350 ms, followed by a 150 ms blank screen, which was replaced by a blue or red colored square. Squares could appear on the left, right, or center of the screen. Participants were

instructed to press a response key as quickly and accurately as possible as soon as the colored square appeared. Allowable responses were either a blue key (the "a" key located on the left of the keyboard) and a red key (the "l" key located on the right of the keyboard). Thus, the position of the colored square on the screen could be congruent or incongruent with the position of the response key on the keyboard. Squares disappeared from the screen when a response was made, or after 2000 ms elapsed with no response. Additionally, there were control trials in which the colored square was presented in the center of the screen. Participants were given 24 practice trials to become familiar with the procedure and response keys, followed by a total of 126 testing trials. Testing trials were divided into 3 blocks of 42 trials each, during which each block contained 14 congruent, 14 incongruent, and 14 control trials presented in random order. E-prime was used to present stimuli and collect response times and accuracy. Responses faster than 200 ms were considered spurious and deleted (only 2 trials, or less than 1% of the data were faster than 200 ms). Additionally, responses greater than 2.5 standard deviations above or below a participant's mean were considered outliers and deleted (removing an additional 93 trials, or 2.7% of the data). Finally, for RT analyses only, all incorrect responses were removed from further analyses, because RTs for incorrect trials are not informative. This removed an additional 86 trials (2.6% of the data). After exclusions, Simon scores were calculated for each participant by subtracting the mean RT for congruent trials from the mean RT for incongruent trials. Lower Simon scores therefore represent less interference and better ability to suppress task-irrelevant information than higher Simon scores.

2.6 Results

2.6.1 Statistical approach

The main analyses described in this section are linear mixed-effects models, which allow examination of subject and item effects simultaneously (e.g., Baayen, Davidson, & Bates, 2008). The specific type of model that we used varied according to the distribution of the dependent variable. We used linear mixed-effects regression (lmer) to model RT data, for which the distribution was approximately normal. We used generalized linear mixed-effects regression (glmer) to model accuracy data for English free recall, because all responses were either 0 or 1. We used cumulative link mixed-effects models (clmm; Christensen, 2019) to model accuracy data for German free recall and meaning production, because the responses for these tasks could be 0, .5, or 1. Ninety-five-percent confidence intervals are reported for all fixed effects. We specified the maximal random effects structure for which models would converge.

We had three individual difference measures (including the two different ways of scoring the Waters task) and so we had to decide which measures to use in our final models. We examined whether these variables were correlated with each other and also with the dependent variables in each model. We selected the Waters measure that had the higher correlation with each dependent measure, and entered that variable and Simon scores into the final models as predictors. Because these variables measure similar constructs, we assessed whether entering two individual difference measures in the same model would result in potentially harmful multicollinearity. We conducted two tests of multicollinearity (condition number and VIF; see below for more details), and if these did not indicate harmful collinearity, we entered Simon score and the Waters measure with the higher correlation with the dependent into the final model as predictors.

In all analyses, the predictor variables of German length and English concreteness were mean-centered in order to eliminate nonessential multicollinearity (Frank, 2011). Study time (ms) was scaled to match the other variables by dividing by 1,000, and Simon scores and Waters total span were scaled to match other variables by dividing by 10. The baseline conditions were coded as follows: unambiguous words were the baseline to which homonyms and polysemes were compared, sentences were the baseline to which definitions were compared, and sentence reading was the baseline to which sentence generation was compared.

All analyses were conducted in R 3.6.0 using version 1.1.21 of the lme4 package (Bates, Maechler, Bolker, & Walker, 2014), version 3.1.0 of the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2017) to assess significance, and version 4.25 of the ordinal package (Christensen, 2019) to fit mixed-effects models with ordinal dependent variables.

2.6.2 Vocabulary training analyses

Before conducting analyses on any of the outcome measures, we first examined participant accuracy during training. Our goal was to ensure that participants followed training directions and completed the task as instructed. For read trials (i.e., training trials for which participants were given an English context sentence containing a German target word and instructed to read and retype it), we compared participant responses to the sentence they had been given to read and retype. Exact matches and responses containing minor typos or omissions of non-content words were awarded one point. Responses with omissions of content words or target words, or responses that did not match the given sentence were scored as zero points. Average accuracy was calculated for each participant, and responses from any participant whose accuracy was below 80% were manually inspected. For Experiment 1, two participants were removed from all further analyses

because they failed to follow training instructions (i.e., they failed to read and retype sentences in the read condition, and instead generated their own sentences for all responses), leaving data from 25 participants for analyses. Mean accuracy on read trials for the remaining participants in Experiment 1 was 92%.

For generate trials, we manually inspected a random sample of responses for each participant to ensure that they contained the target word and were a reasonable attempt to capture the meaning of the word. All participants met these criteria.

2.6.3 Free recall results

We examined free recall performance separately for English words and German words because accuracy for full word pairs was too low to permit further analyses (< 2% of English-German word pairs were correctly recalled). We first examined Pearson correlations between English and German free recall accuracy the individual differences measures (i.e., two outcome measures from the Waters task and the Simon score). Simon score significantly correlated with both English (r = -.10) and German (r = -.10) accuracy. Waters total span significantly correlated with both English (r = .06) and German (r = .06) accuracy as well. However, Waters set size span was not significantly correlated with either English or German accuracy; therefore, we used Waters total span instead of Waters set size span in all analyses. Waters total span and the Simon score were significantly correlated with each other (r = -.29). Because we had specific hypotheses about both Waters and Simon scores, we wanted to include both in the final models. To make sure we could clearly test the effects of each of these as independent predictors we assessed the risk of multicollinearity between these two predictors, and in the full model, for both the English and German free recall models.

To do this, we examined the condition number, a test of the overall amount of collinearity in a model (Baayen, 2008). For the English free recall model, the condition number indicated that there was standard collinearity ($\kappa = 20.29$), which falls below the threshold for potentially harmful collinearity ($\kappa = 30$; Baayen, 2008). We then examined the variance inflation factor (VIF), a measure of how much the individual variables in a model are affected by collinearity (Frank, 2011). We observed VIF values of 2.48 Simon scores and 2.32 for Waters total span, which were both below the point at which VIF becomes problematic (VIF = 5.00). Therefore, we proceeded to build a final English free recall model that included both Waters and Simon.

We followed the same procedures to test for multicollinearity for the German free recall data, and found similar results.[3] There was standard collinearity in the model ($\kappa = 20.29$), and VIF values for Simon (2.36) and Waters (2.34) were both below the point at which VIF becomes problematic.

We constructed two linear mixed-effects models to examine accuracy: Model 1) a glmer model to examine accuracy on English words because acceptable values for the dependent variable were either 0 or 1, and Model 2) a clmm model to examine accuracy on German words because partial credit was awarded and acceptable values for the dependent variable were 0, .5, or 1. For both models, the fixed effects of theoretical interest were training condition (read or generate), word type (unambiguous, homonym, polyseme), Simon score, Waters total span, and the four-way

[3] Note that clmm models do not support the standard methods of testing for testing collinearity (condition number and VIF). In Experiments 1-3, to test for multicollinearity in clmm models, we constructed glmer models that were otherwise identical to the clmm models. We then tested for multicollinearity with the glmer models using the procedures described above. Because clmm and glmer models yield roughly equivalent results the estimates of multicollinearity from a glmer model is informative for a clmm model.

interaction of word type, training condition, Waters total span, and Simon score. Fixed effects were also entered for the following control variables: study time during training, German word length, and English concreteness.

For both models, we specified the maximal random effects structure for which our models would converge (e.g., Barr, Levy, Scheepers, & Tily, 2013). At minimum, we included random intercepts for both subjects and items, and when model convergence allowed, we included by-subject random slopes for the interaction term that that was most central to our hypotheses: word type and training condition. We considered the other variables in the model to be control variables, and we did not specify random effects for these (see Barr et al., 2013).

2.6.3.1 English free recall accuracy

We began by examining descriptive statistics for English free recall accuracy. Mean proportion correct for all English words was .28 (*SD* = .45). The mean accuracy for words trained with sentence reading was .24 *(SD* = .43), and the mean accuracy for words trained with sentence generation was .32 (*SD* = .47). Mean accuracy was .24 (*SD* = .43) for unambiguous words, .33 for homonyms (*SD* = .47), and .35 for polysemes (*SD* = .48).

The model equation and results for English free recall accuracy (Model 1) are presented in Table 3, and the estimates of the random effects for this model are presented in Table 4. Of the fixed effects of theoretical interest, there was a significant effect of word type, such that both homonyms, b = -1.76, SE = 0.38, z = -4.59, $p < .001$, and polysemes, b = -1.47, SE = 0.37, z = -3.95, $p < .001$, were recalled more accurately than unambiguous words. There were no significant main effects of generation, Simon score, or Waters total span, but these were all involved in significant higher-order interactions. First, there was a two-way interaction of word type and Waters total span, b = 0.43, SE = 0.22, z = 1.99, $p < .05$. We probed this interaction using the

effects package in R (Fox & Weisberg, 2019), and graphed the estimated probability of correctly recalling an English word across word types at one standard deviation above and below the mean for Waters total span (Figure 4). This revealed that homonyms were recalled significantly more accurately than unambiguous words, but only for participants with a higher WM span.

There was also a significant three-way interaction of between Simon score, word type, and training condition, $b = 0.35$, $SE = 0.16$, $z = 2.14$, $p = .03$. We probed this interaction using the *effects* package in R, and graphed the estimated probability of correctly recalling an English word for read vs. generate conditions across word types, at one standard deviation above and below the mean Simon score (Figure 5). This revealed that generation was more beneficial for ambiguous words than unambiguous words, and furthermore that this benefit was significant for participants who were better able to suppress task-irrelevant information (i.e., had lower Simon scores), and not significant or marginal for participants who were less able to suppress task-irrelevant information (i.e., had lower Simon scores). The hypothesized four-way interaction of word type, training condition, Simon score, and Waters total span did not reach significance.

Table 3. Fixed effects estimates for Model 1, English free recall

			95% CI				
	Estimate	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Intercept	-1.48	0.28	-2.03	-0.94	-5.34	<.001	***
Homonyms	-1.76	0.38	-2.51	-1.01	-4.59	<.001	***
Polysemes	-1.47	0.37	-2.21	-0.74	-3.95	<.001	***
Generate condition	-0.02	0.18	-0.38	0.35	-0.08	.93	
Simon score	-0.17	0.09	-0.35	0.02	-1.76	.08	†
Waters total span	-0.05	0.16	-0.37	0.27	-0.31	.76	
Study time	0.10	0.05	0.01	0.20	2.10	.04	*
English concreteness	0.51	0.18	0.15	0.87	2.77	.01	**
Word length German	0.06	0.05	-0.04	0.16	1.15	.25	
Translation number	0.14	0.16	-0.18	0.46	0.86	.39	
Homonyms*Generation	1.22	0.29	0.65	1.79	4.21	<.001	***
Polysemes*Generation	0.86	0.28	0.31	1.41	3.07	<.001	**
Homonyms*Simon	0.32	0.12	0.09	0.56	2.66	.01	**
Polysemes*Simon	0.16	0.12	-0.07	0.38	1.35	.18	
Generate*Simon	0.05	0.10	-0.15	0.26	0.52	.60	
Homonyms*Waters	0.43	0.22	0.01	0.86	1.99	<.05	*
Polysemes*Waters	0.04	0.20	-0.35	0.44	0.22	.83	
Generate*Waters	0.22	0.19	-0.15	0.60	1.19	.24	
Simon*Waters	0.03	0.07	-0.11	0.16	0.38	.71	
Homonyms*Generation*Simon	-0.35	0.16	-0.68	-0.03	-2.14	.03	*
Polysemes*Generation*Simon	-0.31	0.16	-0.63	0.00	-1.95	.05	†
Homonyms*Generation*Waters	-0.34	0.30	-0.93	0.25	-1.13	.26	
Polysemes*Generation*Waters	-0.10	0.29	-0.67	0.47	-0.35	.73	
Homonyms*Simon*Waters	0.01	0.09	-0.16	0.19	0.13	.89	
Polysemes*Simon*Waters	0.01	0.08	-0.15	0.17	0.09	.93	
Generate*Simon*Waters	-0.08	0.08	-0.23	0.07	-1.00	.32	
Homonyms*Generation*Simon*Waters	0.08	0.12	-0.16	0.32	0.63	.53	
Polysemes*Generation*Simon*Waters	-0.01	0.12	-0.24	0.22	-0.11	.92	

Model equation. Model 1 <- glmer(EnglishACC~ 1 + WordType*TrainingCondition*Simonscore*waterstotcorr + studytime + Englishconcreteness + Germanconcreteness + transnum + (1|Subject)+(1|EnglishWord), data=d1fr, family = binomial, glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 20000)))

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 4. Random effects estimates for Model 1, English free recall accuracy

	Variance	*SD*
Item intercept	0.43	0.65
Subject intercept	0.26	0.51

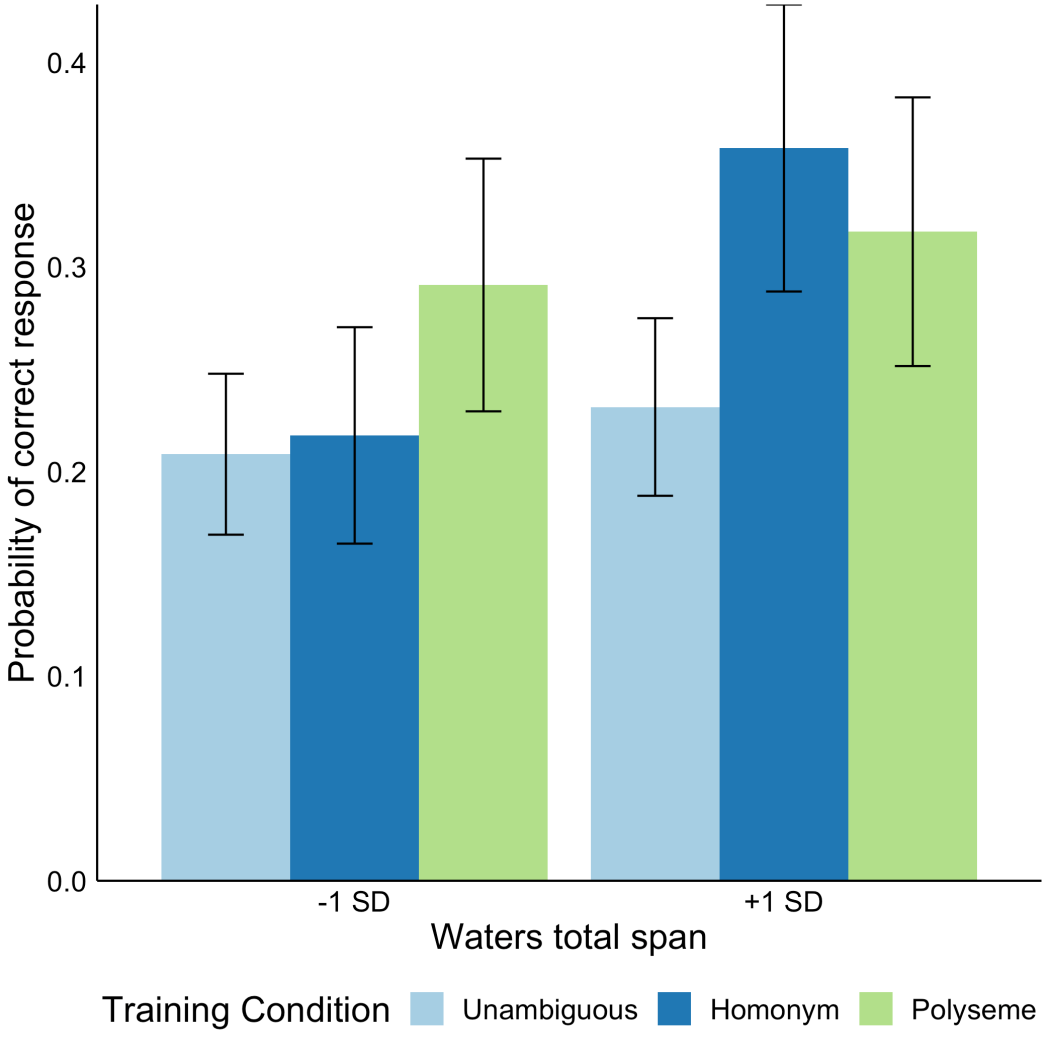

Figure 4. Estimated probability of English free recall by word type and Waters total span. Error bars represent standard error of the mean.

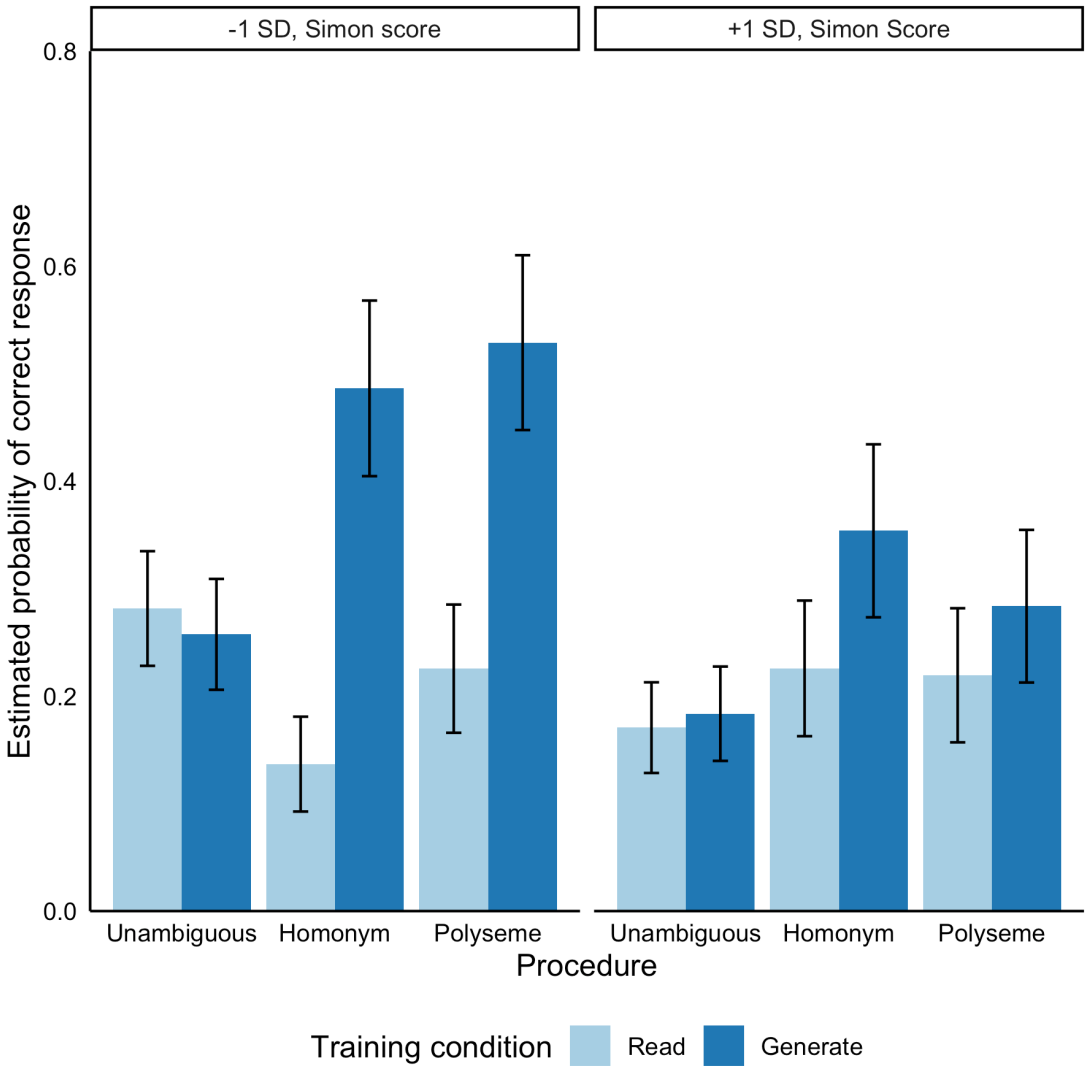

Figure 5. Estimated proportion correct for English free recall by training condition, word type, and Simon score

2.6.3.2 German free recall accuracy

We began by examining descriptive statistics for German free recall accuracy. Mean proportion correct for all German words was .08 (*SD* = .25). The mean accuracy for words trained with sentence reading was .08. *(SD* = .25), and the mean accuracy for words trained with sentence

generation was .09. (*SD* = .27). Mean accuracy was .07 (*SD* = .23) for unambiguous words, .10 for homonyms (*SD* = .27), and .12 for polysemes (*SD* = .30).

We constructed a clmm model to investigate performance on free recall of German words. The model equation and results for this model (Model 2) are presented in Table 5, and the estimates of the random effects for this model are presented in Table 6. Of the fixed effects of theoretical interest, there was a significant interaction of word type and training condition, b = 0.97, SE = 0.42, z =2.31, p = .02. We probed the interaction using the *effects* package, and graphed the estimated probability of correctly recalling a German word for read vs. generate conditions across the three different word types (Figure 6). This revealed that generation was more effective for ambiguous words than unambiguous words, but this difference was only significant for polysemes.

There was also a significant interaction of word type and Simon score, b = -0.40, SE = 0.18, z = -2.25, p = .02. Probing this interaction with the *effects* package revealed that this interaction was driven by differences in polyseme recall for learners with lower vs. higher Simon scores (Figure 7). Specifically, learners with greater ability to suppress task-irrelevant information (i.e., low Simon scores) recalled polysemes significantly more accurately than learners with weaker ability to suppress task-irrelevant information (i.e., higher Simon scores). Unambiguous words and homonyms were not significantly impacted by differences in Simon scores.

Table 5. Fixed effects estimates for Model 2, German free recall

			95% CI				
	Estimate	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	-0.28	0.46	-1.19	0.62	-0.61	0.54	
Polysemes	-0.67	0.42	-1.50	0.16	-1.58	0.11	
Generate condition	-0.17	0.24	-0.64	0.30	-0.71	0.48	
Simon score	-0.02	0.13	-0.27	0.22	-0.19	0.85	
Waters total span	0.16	0.21	-0.25	0.58	0.77	0.44	
Study time	0.06	0.06	-0.06	0.18	0.98	0.33	
English concreteness	0.54	0.21	0.14	0.95	2.62	0.01	**
Word length German	-0.08	0.06	-0.20	0.05	-1.23	0.22	
Translation number	-0.39	0.21	-0.80	0.02	-1.85	0.06	†
Homonyms*Generation	0.62	0.45	-0.26	1.49	1.39	0.17	
Polysemes*Generation	0.97	0.42	0.15	1.80	2.31	0.02	*
Homonyms*Simon	-0.03	0.20	-0.42	0.36	-0.17	0.86	
Polysemes*Simon	-0.40	0.18	-0.74	-0.05	-2.25	0.02	*
Generate*Simon	-0.11	0.13	-0.36	0.14	-0.89	0.37	
Homonyms*Waters	-0.18	0.34	-0.84	0.48	-0.54	0.59	
Polysemes*Waters	-0.19	0.32	-0.82	0.45	-0.58	0.57	
Generate*Waters	-0.11	0.23	-0.56	0.33	-0.50	0.62	
Simon*Waters	0.01	0.09	-0.16	0.18	0.15	0.88	
Homonyms*Generation*Simon	0.04	0.24	-0.43	0.51	0.17	0.86	
Polysemes*Generation*Simon	0.26	0.22	-0.18	0.70	1.14	0.25	
Homonyms*Generation*Waters	0.19	0.42	-0.63	1.00	0.45	0.66	
Polysemes*Generation*Waters	0.36	0.41	-0.45	1.17	0.87	0.39	
Homonyms*Simon*Waters	0.02	0.14	-0.25	0.29	0.13	0.89	
Polysemes*Simon*Waters	-0.07	0.13	-0.31	0.18	-0.54	0.59	
Generate*Simon*Waters	-0.09	0.09	-0.27	0.09	-0.99	0.32	
Homonyms*Generation*Simon*Waters	0.04	0.17	-0.29	0.37	0.25	0.80	
Polysemes*Generation*Simon*Waters	-0.01	0.16	-0.33	0.30	-0.08	0.94	

Model equation. Model 2 <- clmm(GerACC~ 1 + WordType*TrainingCondition*Simonscore*waterstotcorr + studytime + Englishconcreteness + Germanconcreteness + transnum + (1|Subject)+(1|EnglishWord) + (0 + WordType*TrainingCondition|Subject), data=d1fr, link = 'logit', threshold = 'flexible')
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 6. Random effects estimates for Model 2, German free recall

	Variance	*SD*
Item intercept	0.41	0.64
Unambiguous|Subject	0.60	0.78
Polysemes|Subject	0.65	0.81
Homonyms|Subject	0.23	0.48
Generate|Subject	0.05	0.23
Polysemes*Generate|Subject	1.06	1.03
Homonyms*Generate|Subject	0.19	0.44
Subject intercept	<0.001	<0.001

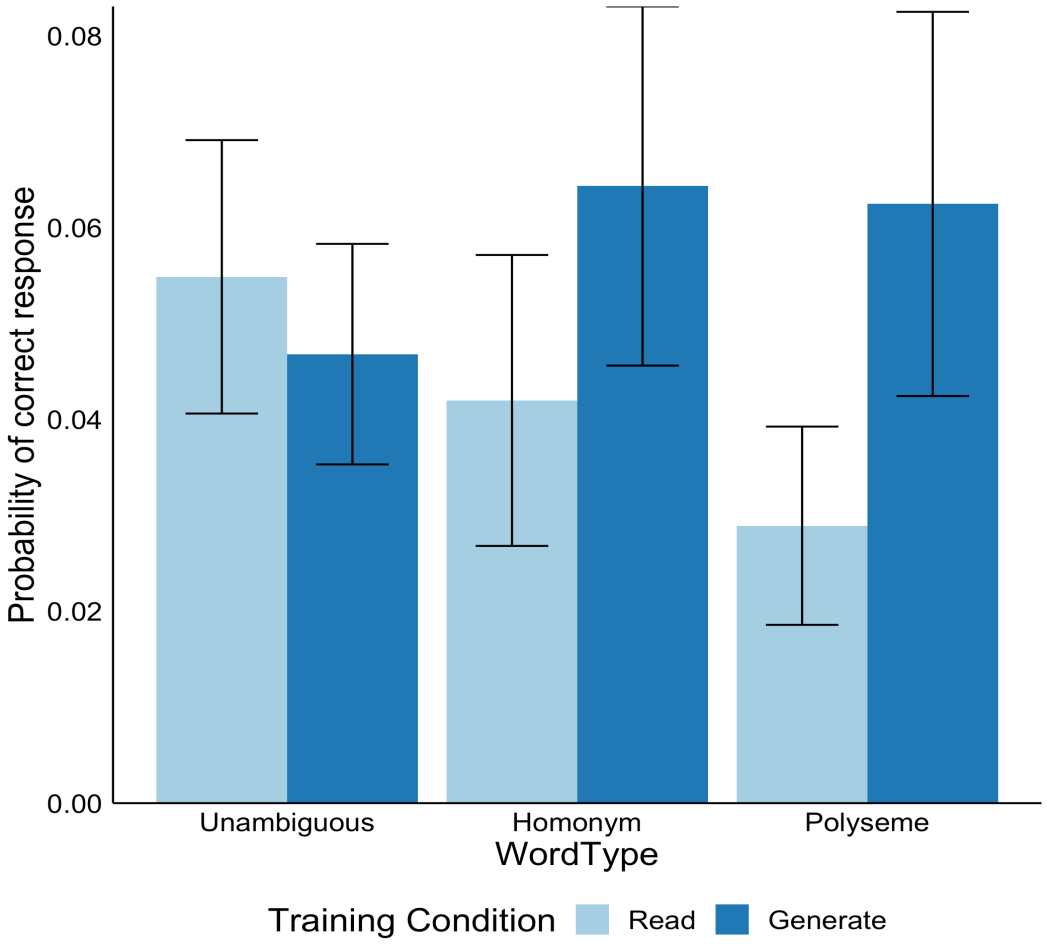

Figure 6. Predicted probability of correct response for German free recall accuracy by word type and condition. Error bars depict standard error of the mean.

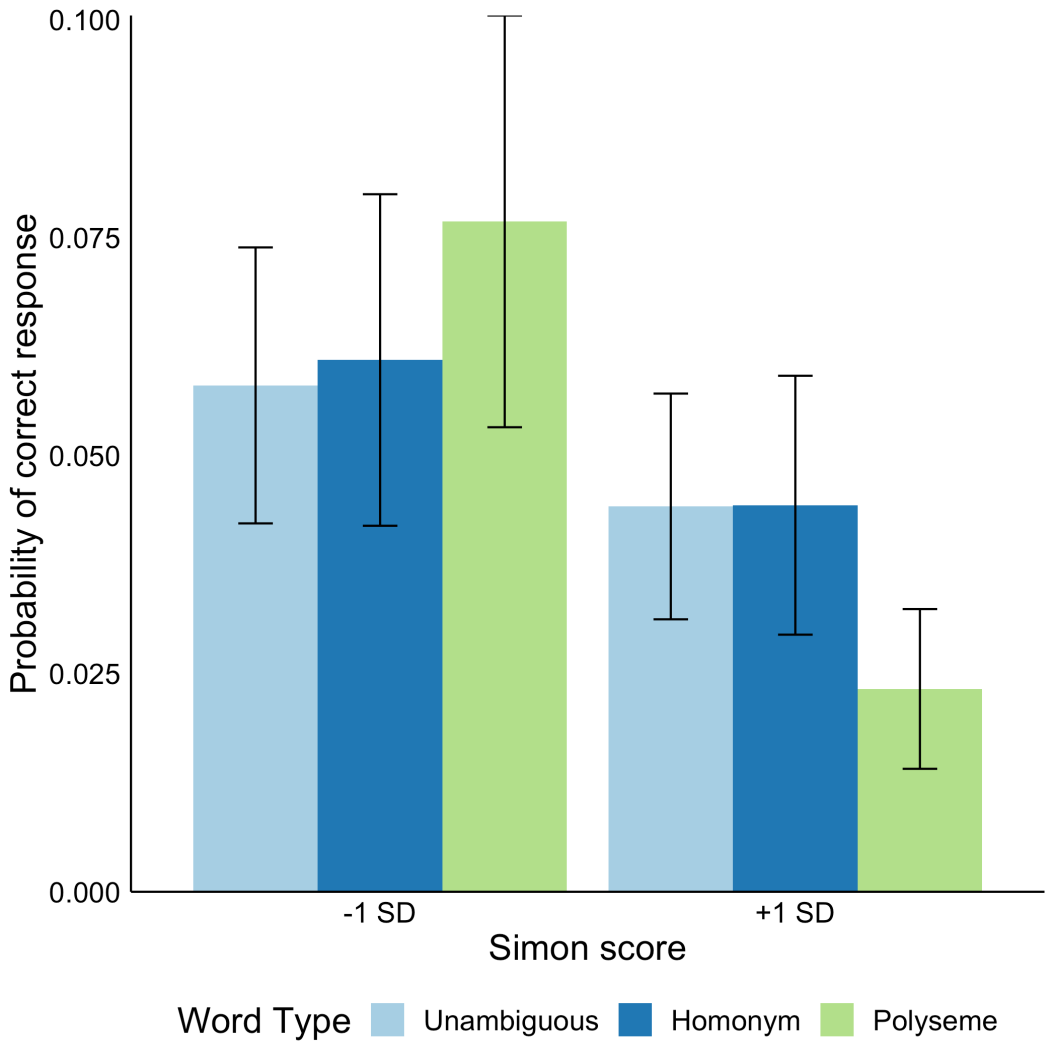

Figure 7. Estimated probability of German free recall by word type and Simon score. Error bars represent standard error of the mean.

2.6.4 Translation production results

This section describes the results of analyses of the translation production response data. In these analyses, we examined accuracy and response time data from the translation production task for both Session 1 and Session 2 using linear mixed-effects models. For both models, the fixed effects of theoretical interest were training condition (read or generate), word type (unambiguous, homonym, polyseme), Simon score, Session (1 or 2) and the interaction of word type, training condition, Simon score, and Waters total span. Fixed effects were also

entered for the following control variables: study time during training, German word length, and English concreteness.

For both models, we specified the maximal random effects structure justified by our experimental design. We included random intercepts for both subjects and items, and included by-subject random slopes for the interaction term and session. We considered the other variables in the model to be control variables, and we did not specify random effects for these (see Barr et al., 2013).

2.6.4.1 Translation production accuracy analyses

We removed data for one participant for both sessions because Session 1 accuracy was 0%, indicating the participant was not paying attention or was not following the directions. In addition to this exclusion, there were some missing data. Due to an audio recorder malfunction, Session 2 translation production data for one participant were not recorded. We removed all data from this participant, leaving data from 24 participants for analyses.

Next, we examined descriptive statistics for the translation production analyses. Mean proportion correct in Session 1 was .25, (SD = 0.16), and mean proportion correct in Session 2 was .12*,* (*SD* = 0.12). Mean proportion correct differed across session, word type, and training condition (see Figure 10). The overall proportion correct was higher for Session 1 (*M* = .25*, SD* = 0.16) than Session 2 (*M* = 0.12*, SD* = 0.12), for unambiguous words (*M* = .23, *SD* = 0.39) than polysemes (*M* = .13, *SD* = 0.31) or homonyms (*M* = .17, *SD* = 0.35), and for generate (*M* = .19, *SD* = 0.36) than read (*M* = .18, *SD* =0.36) training.

We next examined Pearson correlations between translation production accuracy and the individual difference measures (i.e., two outcome measures from the Waters task and the Simon score). Simon score significantly correlated with translation production accuracy (*r* = -.08), and so

did Waters total span (r = .07). However, Waters set span size was not significantly correlated with translation production accuracy, and so we decided to use Waters total span instead of Waters set size span in all analyses. Following procedures described above, we tested for potentially problematic multicollinearity between Waters total span and Simon scores, with tests for condition number and VIF. The results of these tests indicated that there was standard collinearity (κ = 21.06), which falls below the threshold for potentially harmful collinearity, and the VIF values for Waters total span (1.31) and Simon (1.54) were below the point at which VIF becomes problematic. Therefore, we proceeded to build a final translation production accuracy model that included both Waters and Simon.

The model equation and results for Model 3 are presented in Table 7, and the random effects are presented in Table 8. Of the fixed effects of theoretical interest there was a significant effect of Session, b = -1.20, SE = 0.10, z = -12.18, $p < .001$, such that performance on Session 1 was significantly better than performance on Session 2. There was a main effect of word type such that correct translations of polysemes were produced significantly less often than unambiguous words, b = -0.96, SE = 0.32, z = -2.97, $p < .001$.

Additionally, there was also a significant interaction of word type and training condition, b = 0.73, SE = 0.32, z = 2.30, p = .02. We probed this interaction using the *effects* package in R, and graphed the estimated probability of correctly producing a translation for read vs. generate conditions across the three different word types (see Figure 8). This revealed that there were no significant differences in the estimated probability of producing a correct response for either unambiguous words or homonyms in either the read or the generate condition. Instead, the word type by training condition interaction was driven by significant differences in the estimated proportion correct for polysemes that were trained in the read vs. generate conditions. Specifically,

there was no training effect for unambiguous words or homonyms, although unambiguous words were translated most accurately overall. However, there was a training effect for polysemes, such that polysemes trained with generation were translated more accurately than polysemes trained without generation, and more accurately than homonyms trained with either sentence generation or sentence reading.

There was also a significant interaction of word type and Simon score, b = -0.29, SE = 0.13, z = -2.32, p = .02. We probed the interaction using the *effects* package, and graphed the estimated probability of correctly producing a translation for read vs. generate conditions at one standard deviation below and above the mean Simon score (Figure 9). This revealed that polysemes were recognized significantly more often when participants were better able to suppress task-irrelevant information (i.e., had lower Simon scores). In other words, the translation-ambiguity disadvantage for polysemes was offset by greater inhibitory control.

The effects of Waters total span did not reach significance, and the hypothesized four-way interaction of word type, training condition, Simon, and Waters was not significant.

Table 7. Fixed effects estimates for Model 3, translation production accuracy

			95% CI				
	Estimate	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	-0.46	0.32	-1.09	0.16	-1.45	.15	
Polysemes	-0.96	0.32	-1.59	-0.32	-2.97	<.001	**
Generate condition	-0.09	0.17	-0.42	0.24	-0.52	.60	
Simon score	-0.12	0.13	-0.38	0.13	-0.94	.35	
Waters total span	0.08	0.22	-0.35	0.51	0.35	.72	
English frequency	0.48	0.20	0.08	0.88	2.33	.02	*
Session	-1.20	0.10	-1.40	-1.01	-12.18	<.001	***
Study time	0.13	0.04	0.04	0.22	2.89	<.001	**
Translation number	0.23	0.14	-0.05	0.51	1.61	.11	
English concreteness	0.27	0.15	-0.02	0.56	1.82	.07	†
German word length	-0.16	0.04	-0.24	-0.07	-3.46	<.001	***
Definition length	-0.04	0.02	-0.08	0.00	-1.77	.08	†
Homonyms*Generation	-0.12	0.31	-0.72	0.48	-0.38	.70	
Polysemes*Generation	0.73	0.32	0.11	1.34	2.30	.02	*
Homonyms*Simon	0.04	0.12	-0.18	0.27	0.38	.70	
Polysemes*Simon	-0.29	0.13	-0.54	-0.05	-2.32	.02	*
Generate*Simon	-0.02	0.09	-0.20	0.15	-0.27	.79	
Homonyms*Waters	0.10	0.20	-0.30	0.50	0.49	.62	
Polysemes*Waters	-0.28	0.20	-0.68	0.12	-1.39	.16	
Generate*Waters	0.03	0.16	-0.28	0.33	0.17	.87	
Simon*Waters	-0.12	0.09	-0.29	0.05	-1.37	.17	
Homonyms*Generation*Simon	-0.01	0.15	-0.31	0.29	-0.05	.96	
Polysemes*Generation*Simon	0.27	0.17	-0.06	0.59	1.58	.11	
Homonyms*Generation*Waters	-0.17	0.27	-0.70	0.36	-0.62	.53	
Polysemes*Generation*Waters	0.35	0.28	-0.20	0.90	1.24	.22	
Homonyms*Simon*Waters	0.05	0.08	-0.10	0.19	0.61	.54	
Polysemes*Simon*Waters	-0.09	0.08	-0.24	0.06	-1.17	.24	
Generate*Simon*Waters	-0.03	0.06	-0.15	0.09	-0.52	.60	
Homonyms*Generation*Simon*Waters	0.04	0.10	-0.16	0.25	0.39	.70	
Polysemes*Generation*Simon*Waters	0.11	0.11	-0.10	0.33	1.05	.30	

Model equation. Model 3 <- clmm(GerACC~ 1 + WordType*TrainingCondition*Simon*Waters + EngFreq + studytime + transnum + Session + Englishconcreteness + wordlengthgerman + transnum + deflen+ (1|Subject)+(1|EnglishWord) + (0 + WordType*ReadorGen|Subject), data=transprod, link = 'logit', threshold = 'flexible')

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 8. Random effects estimates for Model 3, translation production accuracy

	Variance	*SD*
Item intercept	0.28	0.53
Unambiguous|Subject	1.01	1.00
Polysemes|Subject	0.65	0.81
Homonyms|Subject	1.23	1.11
Generate|Subject	0.11	0.34
Polysemes*Generate|Subject	0.24	0.49
Homonyms*Generate|Subject	0.48	0.70
Subject intercept	<0.001	<0.001

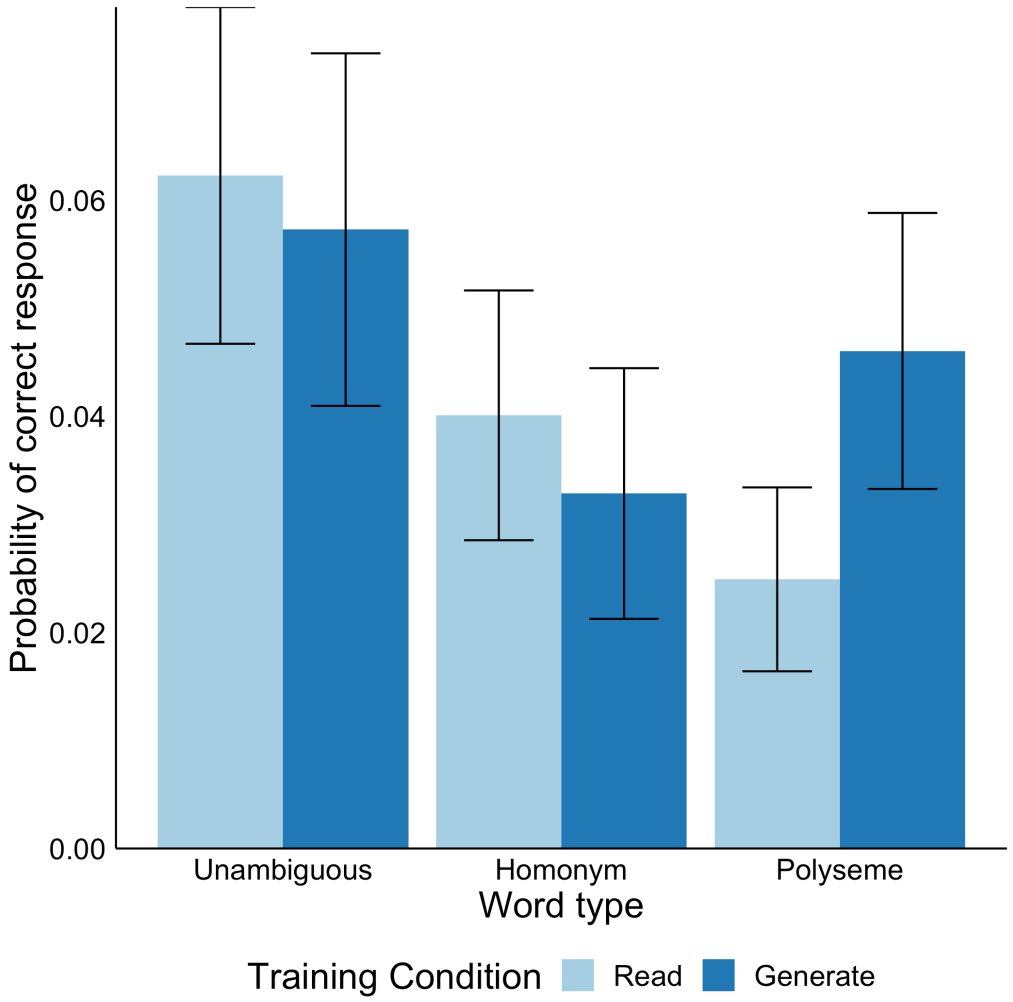

Figure 8. Estimated probability of correct translation production by word type and training condition. Error bars represent standard error of the mean.

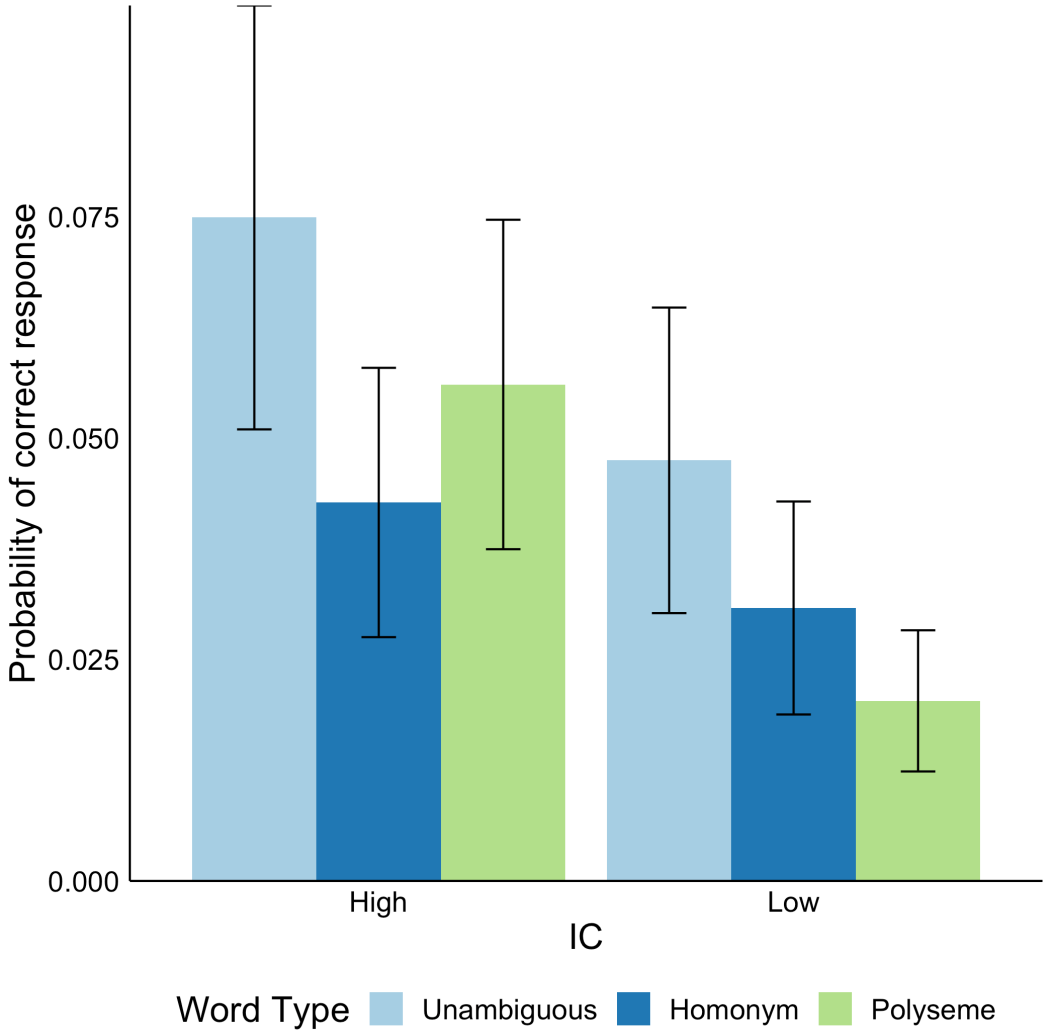

Figure 9. Estimated probability of correct translation by word type and Simon score. Error bars represent standard error of the mean.

2.6.4.2 Translation production RT results

For RT analyses only, we excluded incorrect responses (77.3% of the total trials). An additional 43 responses (5.6% of correct trials) were excluded because of voice key errors (i.e., the participant made an extraneous noise, such as "umm", before producing their response, which triggered the voice key before they actually responded. Responses faster than 200 ms were considered spurious and therefore excluded, which removed an additional three trials (< 1% of

correct trials). Trials with response times 2.5 standard deviations or greater above or below a participant's mean response time or a word's mean response time were excluded as outliers and treated as missing values, which resulted in the removal of 37 trials (4.8% of the correct word trials).

During testing, participants were asked to provide both translations of ambiguous words. However, only RTs for the first translations were included in this set of analyses. Because there were very few trials (< 2% of all trials) for which participants successfully produced both translations of a word, it was not possible to conduct reliable analyses to test for differences between first and second translations. Therefore, we excluded all trials for translations produced second from all RT analyses, which removed 7.9% of all correct word trials.

We examined descriptive statistics for RTs across Session, word type, and training condition. Response times were longer for Session 2 (M = 2929 ms, SD = 1887) than Session 1 (M = 3001 ms, SD = 1896), for unambiguous words (M = 3003 ms, SD = 2006) than homonyms (M = 2952 ms, SD =1729) or polysemes (M = 2847 ms, SD = 1773), and for read trials (M = 3001 ms, SD = 1998) than generate trials (M = 2907 ms, SD = 1787).

We next examined Pearson correlations between translation production RTs and the three individual difference measures. In contrast to the accuracy analyses, none of the individual difference variables were significantly correlated with translation production RTs, although as before Waters total span was more strongly correlated with RT (r = .05) than Waters set size span (r = .01). As before, we decided to use Waters total span instead of Waters set size span in all analyses. Following procedures described above, we tested for potentially problematic multicollinearity between Waters total span and Simon scores, and found only standard collinearity (κ = 22.98) and the VIF values (Simon = 1.71, Waters = 1.63) were below the point at which VIF

becomes problematic. Therefore, we proceeded to build a final translation production accuracy model that included both Waters and Simon.

We constructed a lmer model to examine translation production response time data. The fixed effects of theoretical interest were training condition (read or generate), word type (unambiguous, homonym, polyseme), session, Simon score, and the interaction of word type, training condition, Simon score, and Waters total span. The amount of time spent studying a word in training, German word length, English concreteness, and translation number (translation viewed first or second in training) were included as control variables.

The model equation and fixed effects estimates for Model 4 are presented in Table 9, and the random effects estimates are presented in Table 10. None of the fixed effects of theoretical interest were significant predictors of response time in this model. The main effects of word type and training condition were not significant, and there was no significant interaction between these variables. The four-way interaction of word type, training condition, Waters, and Simon was not significant.

Table 9. Fixed effects estimates for Model 4, translation production RT

			95% CI				
	Estimate	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Intercept	4710.50	367.39	3990.41	5430.58	12.82	0.00	***
Homonyms	122.82	391.23	-644.00	889.63	0.31	0.75	
Polysemes	667.84	392.36	-101.18	1436.86	1.70	0.09	†
Generate condition	-86.43	203.80	-485.87	313.01	-0.42	0.67	
Simon score	-161.36	134.44	-424.87	102.15	-1.20	0.24	
Waters total span	-143.90	223.93	-582.80	294.99	-0.64	0.52	
English frequency	-396.58	249.64	-885.86	92.71	-1.59	0.12	
Study time	-15.83	67.22	-147.58	115.93	-0.24	0.81	
Translation number	-1521.05	241.43	-1994.25	-1047.85	-6.30	0.00	***
Session	185.78	150.85	-109.89	481.44	1.23	0.22	
English concreteness	-510.17	176.33	-855.77	-164.56	-2.89	0.01	**
German word length	3.69	64.84	-123.39	130.78	0.06	0.95	
Definition length	4.38	25.76	-46.11	54.86	0.17	0.87	
Homonyms*Generation	146.24	397.68	-633.21	925.69	0.37	0.71	
Polysemes*Generation	-95.95	402.32	-884.49	692.60	-0.24	0.81	
Homonyms*Simon	-9.17	150.81	-304.76	286.41	-0.06	0.95	
Polysemes*Simon	197.37	170.00	-135.82	530.56	1.16	0.25	
Generate*Simon	158.64	111.72	-60.33	377.61	1.42	0.16	
Homonyms*Waters	182.84	280.23	-366.40	732.09	0.65	0.51	
Polysemes*Waters	304.15	284.23	-252.95	861.25	1.07	0.29	
Generate*Waters	306.05	200.95	-87.81	699.90	1.52	0.13	
Simon*Waters	-40.61	85.64	-208.46	127.25	-0.47	0.64	
Homonyms*Generation*Simon	-217.05	216.83	-642.03	207.94	-1.00	0.32	
Polysemes*Generation*Simon	-129.53	222.98	-566.58	307.52	-0.58	0.56	
Homonyms*Generation*Waters	-608.60	396.65	-1386.03	168.83	-1.53	0.13	
Polysemes*Generation*Waters	-597.65	382.61	-1347.56	152.26	-1.56	0.12	
Homonyms*Simon*Waters	-75.09	96.79	-264.80	114.63	-0.78	0.44	
Polysemes*Simon*Waters	59.94	99.06	-134.22	254.11	0.61	0.55	
Generate*Simon*Waters	18.42	73.13	-124.93	161.76	0.25	0.80	
Homonyms*Generation*Simon*Waters	23.27	143.37	-257.74	304.27	0.16	0.87	
Polysemes*Generation*Simon*Waters	-96.05	140.20	-370.83	178.73	-0.69	0.49	

Model equation. Model 4 <- lmer(GerACC~ 1 + WordType*TrainingCondition*Simonscore*waterstotcorr + EngFreq+ studytime + transnum + Session + Englishconcreteness + wordlengthgerman + transnum + DefLen+ (1|Subject)+(1|EnglishWord), data=transprodRT, link = 'logit', threshold = 'flexible')

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 10. Random effects estimates from Model 4, translation production RT

	Variance	*SD*
Item intercept	279383	529
Subject intercept	711695	844
Residual	2703844	1644

2.7 Discussion

This experiment examined how the use of generation during vocabulary L2 training impacted the learning of translation-ambiguous German-English word pairs, as measured by performance on free recall and translation production tasks both immediately after learning and after a one-week delay. We hypothesized that we would find an overall benefit of generation for both free recall and translation production tasks, and an overall translation-ambiguity advantage in free recall but an overall translation-ambiguity disadvantage in translation production. Critically, we expected that the generation and the translation ambiguity effects would interact in novel ways. As outlined in more detail in the Introduction, we generated a specific set of predictions based on a combination of two previously unrelated hypotheses: the *semantic settling dynamics* hypothesis (Armstrong & Plaut, 2016) and the *two-factor theory of generation* (Hirshman & Bjork, 1988). Specifically, we expected that polysemes and unambiguous words would benefit from generation, whereas homonyms would be negatively affected by generation. Additionally, we expected that performance on the immediate testing (Session 1) would be better than performance on the delayed testing (Session 2). Finally, we hypothesized that individual differences in WM (as measured by the Waters Reading Span task) and in the ability to suppress task-irrelevant information (as measured by the Simon task) would interact with each other, as

well as with word type and training condition. Specifically, we predicted that learners who were better able to exert inhibitory control (i.e., had lower Simon scores) would display a reduced translation-ambiguity disadvantage, whereas learners who were less able to exert inhibitory control might display a reduced generation effect. And finally, we predicted an interaction of individual differences, word type, and generation, such that learners who are higher in WM and also better able to exert inhibitory control will display a reduced translation-ambiguity disadvantage when words are trained with generation, and this will be especially pronounced for homonyms.

Results from the free recall task partially support these predictions, although it is important to remember that extremely few full German-English word pairs were recalled and so we examined free recall for English and German words separately. The results in the following paragraphs should be interpreted as tentative given the atypical nature of the free recall analyses, and the reader should keep in mind that although we conducted analyses on English words and German words separately, these came from the same, not separate, tasks. With that said, free recall for English words showed a significant interaction of word type and training condition. Probing this interaction revealed that training condition did not impact free recall accuracy for unambiguous words, but generation significantly benefitted both types of ambiguous words. However, this interaction was qualified by a three-way interaction of word type, training condition, and Simon score. Probing this interaction revealed that learners who could exert greater inhibitory control benefitted more from generation for both types of ambiguous words, whereas the benefit of generation was only apparent for homonyms for learners who had weaker inhibitory control. Unambiguous words were not significantly affected by training condition, but learners with weaker inhibitory control performed more poorly on unambiguous words than learners with stronger inhibitory control.

The results from the English free recall task partially support our hypotheses, while also diverging in a few key ways. First, consistent with past reports, we observed a general benefit for translation-ambiguous words, rather than a translation-ambiguity disadvantage. Degani et al. (2014) reported a translation-ambiguity advantage in free recall, and hypothesized that this might be due either to the distinctiveness of words with multiple meanings (DeLosh & McDaniel, 1996), or because the open-ended format and unlimited time allowed for a free recall task might not elicit the same processing dynamics as a more time-constrained task like translation production. This translation-ambiguity advantage in free recall was replicated by Ekves (2014). However, we show for the first time that this translation-ambiguity advantage in free recall is heightened by the generation effect. For free recall of English words, there were no differences between ambiguous and unambiguous words when they were trained without generation, but when they were trained with generation an ambiguity advantage emerged. Additionally, we demonstrate for the first time that there is an interaction of word type and training condition in free recall, and furthermore that this is sensitive to individual differences in inhibitory control: learners with greater ability to suppress task-irrelevant information perform significantly better for homonyms and polysemes when trained with generation, whereas for participants with lower ability to suppress task-irrelevant information this advantage of generation is constrained only to homonyms, and not polysemes.

Free recall for German words again showed a significant interaction of word type and training condition. Probing this interaction revealed that there was no generation effect for unambiguous words or homonyms, but there was a significant benefit of generation for polysemes. This partially supports our hypotheses – we predicted that generation would benefit polysemes in particular. However, we did not find evidence in this task that generation harms recall of

homonyms. This discrepancy might be due to task demands, because a free recall task does not require a participant to access the meanings of words. We expected to see a disadvantage for homonyms trained with generation because generation enhances semantic processing, and therefore may draw attention to the multiple competing meaning of homonyms and induce difficulty settling on a meaning to produce. However, in free recall these dynamics may not have come into play because learners were required only to retrieve the form of the words and not the meaning. An alternate possibility is that because there was no time limit for free recall, learners may simply have had sufficient time to overcome any competitive dynamics that emerged during homonym retrieval, and were able to produce a correct response.

Finally, although Simon scores did emerge as a significant predictor of overall German free recall accuracy, there was no three-way interaction of word type, training condition, and Simon score like there was for English free recall. There was however, an interaction between Simon score and word type, such that individual differences in inhibitory control had little impact effect on unambiguous words or homonyms, but there was a significant difference in how learners with better and worse inhibitory control performed for polysemes. Specifically, learners with better inhibitory control performed significantly better on polysemes than did learners with worse inhibitory control. This agrees with the English free recall results for polysemes, but whereas homonyms were affected by Simon score in English free recall they were not in German free recall. This difference might be explained by the Revised Hierarchical Model. This model assumes that L1 forms are strongly connected to meanings, whereas L2 forms are only weakly connected to meanings, especially at lower levels of proficiency. Therefore, free recall of German words may be a worse measure of meaning access than free recall of English words. Although in general free recall does not require meaning access, English word forms are strongly connected to meaning

representations for native English speakers, and so some amount of meaning access likely occurred when recalling the English words. In contrast, especially given the low levels of proficiency of learners in this study, it is unlikely that meaning access occurred when recalling the German words. In English free recall, learners with better ability to inhibit task-irrelevant information (such as competing meanings) performed better on homonyms than did learners with less inhibitory control, but in German free recall we did not find such an effect. This may be because in German free recall there was a lesser need to inhibit any task-irrelevant information due to lesser amounts of potentially competitive meaning activation.

Results from translation production accuracy analyses showed a significant main effect of word type, but this was qualified by a significant interaction of word type and training condition. Probing this interaction revealed that there was no generation effect for unambiguous words or homonyms, but there was a significant benefit of generation for polysemes. As predicted, there was a significant effect of Session, such that accuracy was higher for Session 1 than Session 2.

We primarily expected to find results in the accuracy data, and not the RT data, based on past reports (i.e., Degani et al., 2014) and the low number of trials available for RT analyses. Indeed, we found little of interest to report in these analyses other than the expected effect of Session, such that RTs were slower in Session 2 than Session 1.

2.7.1 Future directions

There are a number of interesting future directions suggested by the results of this study, which can be grouped into three main categories: addressing study limitations, future directions for laboratory studies, and implications for classroom studies and applications.

First, there are a number of limitations of the current study that future research should take into account. One such limitation is that accuracy for both the free recall and translation production tasks was quite low, which may have impacted our ability to detect the effects of interest. Future studies should consider ways to increase learning and retention. There are a variety of methods that might be used to accomplish that, including: increasing the number of encounters with a word, increasing the number of training sessions, including multimodal components of training (i.e., audio recordings, speech production tasks), or using different testing tasks. Due to time limitations, we were only able to test participants on free recall and translation production tasks, but other tasks that test receptive, rather than productive, word knowledge might show stronger or different effects of the training manipulations. For example, translation recognition tasks might measure aspects of word learning that the present study did not, such as cases when even though a participant does not remember a word and its translation well enough to produce the word, they might be able to recognize if a translation pair is correct or incorrect.

Additionally, it is interesting that although we did find the hypothesized benefit of generation for polysemous words, we did not find evidence for the reverse – that generation negatively impacts homonyms. One potential reason we did not observe this effect might be due to the tasks we selected: free recall and translation production. Free recall tasks do not require meaning access, and so it is possible that we did not observe the predicted negative effects of generation in the free recall data because our testing task did not require learners to engage in processes that would cause the predicted dynamics to emerge. However, if this were the case we would still expect to see a negative effect of generation for homonyms during L1-L2 translation production, because this task is known to require meaning access (Kroll & Stewart, 1994).

Although we see a trend in that direction, it did not reach significance. Therefore, the amount of meaning access required by the testing task is not a sufficient explanation for these results.

One additional possibility concerns the amount of time learners had to produce a word in the free recall and translation production tasks. According to the SSD hypothesis, different processing dynamics are evoked over time, as a function of word type. For example, semantic activation increases quickly for polysemes due to cooperative activation from multiple related meanings, but semantic activation increases more slowly for homonyms due to competition from the multiple unrelated meanings. However, given enough time, the amount of semantic activation reaches an equivalent point for unambiguous words, homonyms, and polysemes, and the competitive and cooperative dynamics at play in the early stages of word processing are resolved because the learner has selected a word (in the case of unambiguous words) or a meaning (in the case of ambiguous words). Both the free recall and translation production tasks allowed learners an unlimited amount of time to produce words, and so the short-term competitive and cooperative dynamics may have resolved and no longer been relevant to task success by the time a learner settled on a word to produce. Future studies should consider including tasks that require greater meaning access and require faster responses.

One important application of the present study is how translation-ambiguous words are taught in second language classrooms. The present results provide some insight into how the use of generation may be best be introduced into L2 classrooms. The current study provides evidence that generation positively impacts learning for polysemous words, and does not hinder unambiguous words or homonyms. Therefore, generation can be widely used for training L2 vocabulary, because it will help a subset of words and it is at least as effective as traditional repetition approaches for other words.

2.7.2 Conclusions

Overall, the current study is the first to demonstrate that generation during word learning may affect different types of ambiguous words in different ways. We examined this question both immediately after learning and after a one-week delay, as well as with a task that does not require meaning access (free recall) and one that requires greater meaning access (L1-L2 translation production). Generally speaking, we found evidence that generation is particularly effective for polysemous words, and neither helps nor hurts unambiguous words or homonyms. Furthermore, we report multiple instances in which individual differences in inhibitory control modulate these effects. In general, learners who have better inhibitory control fare better on novel word learning, but this effect seems to be heightened for polysemous words.

3.0 Experiment 2

Whereas Experiment 1 examined how learners acquire translation ambiguous and unambiguous L2 vocabulary words, the present experiment examines how learners acquire semantically-ambiguous and unambiguous L1 vocabulary words. Semantic ambiguity occurs when words have multiple meanings or senses, depending on the degree of meaning relatedness. For example, homonyms are words with multiple related meanings (e.g., BANK can mean either a river bank or a financial institution) whereas polysemes are words with multiple related senses (e.g., FOOT can mean either a body part or a unit of measure). This is a current topic of interest in the literature, because a large number of words across many languages are ambiguous (e.g., Klein & Murphy, 2001). However, most research in this area has investigated how semantically-ambiguous words are processed and produced, and relatively little work has investigated how semantically-ambiguous words are *learned*.[4] In the current study, we aim to address gaps in our knowledge by investigating how semantic ambiguity affects novel L1 word learning in adults. Based on predictions generated from models of semantic ambiguity resolution during the

[4] Although a substantial body of literature investigates semantically-ambiguous word learning in children (e.g., Doherty, 2004; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Markman & Wachtel, 1988; Mazzocco, 1997), we elect to not review these studies because word learning in adults involves different cognitive processes than in children (Gillette, Gleitman, Gleitman, & Lederer, 1999; Takashima, Bakker-Marshall, van Hell, McQueen, & Janzen, 2019). However, these studies do provide evidence that learning to map multiple meanings to a single word form presents a challenge for successful vocabulary learning.

processing of ambiguous words (e.g., the SSD hypothesis; Armstrong & Plaut, 2016), we predict that competitive and cooperative dynamics will emerge during ambiguous word learning that will result in an advantage for polysemes and a disadvantage for homonyms relative to ambiguous words. Furthermore, because these dynamics are evoked as a result of semantic processing, we predict that the use of a vocabulary training method known to enhance semantic processing, the generation effect, will cause these effects to emerge (in other words, we expect to see these effects when words are trained with generation, but not necessarily when words are not trained with generation). Additionally, this experiment also investigates the role of individual differences in novel L1 vocabulary learning, and asks whether learner characteristics impact the success of the generation effect when applied to semantically-ambiguous word learning. We begin with a discussion of past findings regarding semantic ambiguity, and then turn to a discussion of generation effects with semantically-ambiguous words, and close with an overview of the current experiment.

3.1 Semantic ambiguity effects

A large number of studies have observed differences in how unambiguous words and ambiguous words are processed, and more specifically differences in processing between homonyms, polysemes and unambiguous words (e.g., Azuma & van Orden, 1997; Borowsky & Masson, 1996; Eddington & Tokowicz, 2015; Hino, Lupker, & Pexman, 2002; Hino, Lupker, Sears, & Ogawa, 1998; Hino, Pexman, & Lupker, 2006; Jager & Cleland, 2016; Jaztrzembski, 1981; Kellas, Ferraro, & Simpson, 1988; Klepousniotou, 2002; Rice et al., 2019; Rodd et al., 2002; Tokowicz & Kroll, 2007). In this section we briefly review the key findings from these studies.

Overall, words with words with fewer meanings and words with unrelated meanings were recognized more slowly during lexical decision than words with a greater number of meanings or meanings that were related (Azuma & van Orden, 1997). Polysemes specifically are recognized more quickly than homonyms or unambiguous words (Beretta, Fiorentino, & Poeppel, 2005; Rodd et al., 2002), although this advantage appears only for words low in concreteness (Jager & Cleland, 2014) and is further affected by the complex interactions of number of meanings/senses and a large number of psycholinguistic variables, including: context availability, word frequency, and orthographic neighborhood features (Rice et al., 2019).

A variety of accounts of semantic ambiguity have been proposed for these findings, some of which focus on lexical factors such as orthography and phonology, and some of which focus on semantic factors, such as semantic activation and semantic feedback. In the present experiment, we focus primarily on exploring the semantic factors that contribute to semantic ambiguity effects during novel word learning. As described in more depth in the General Introduction, the SSD hypothesis (e.g., Armstrong, 2012; Armstrong & Plaut, 2016) predicts a polyseme advantage and a homonym disadvantage during processing due to temporal settling dynamics that interact with the semantic characteristics of ambiguous words. Specifically, this hypothesis proposes that the main determinant of ambiguity effects is how much semantic processing has occurred. Semantic processing begins when a word form is encountered, and excitatory activation increases gradually, as potential meanings of the word are activated. Number of meanings/senses and meaning relatedness both impact how quickly a word will be recognized. Most importantly for the present experiment, this account predicts that polysemes will benefit from cooperative activation whereas homonyms suffer from competition due to having multiple unrelated meaning features, and so tend to be recognized more slowly than polysemes and unambiguous words. The SSD was developed

to explain L1 semantic ambiguity resolution but these dynamics may underlie the learning of semantically-ambiguous words as well. Therefore, we will support the extension of the SSD hypothesis to L1 vocabulary learning if we find that during novel vocabulary learning generation benefits polysemes and unambiguous words but harms homonyms.

A central feature of the SSD hypothesis is that increasing semantic activation allows ambiguity effects to emerge. Therefore, the present experiment uses the generation effect to increase semantic activation during novel vocabulary learning. In this next section we describe some previous research that investigates the effects of using generation to enhance L1 vocabulary learning.

3.2 Generation effects and word learning

Although many studies of the generation effect have used lexical items as stimuli, most of these studies investigate the memorial effects of generation for words that are already known, and very few studies have investigated generation effects for learning novel L1 words. However, there are a handful of studies that have investigated the use of generation for learning nonword stimuli, which, although different than learning novel L1 words due to differences in orthographic and phonological processing of words and nonwords (Coltheart & Ulicheva, 2018; Plaut, McClelland, Seidenberg, & Patterson, 1996), may still be informative for the purposes of this study. This section reviews the findings of these studies.

Overall, findings from studies that use generation with nonwords are mixed. Whereas some studies report a generation effect with nonwords (Nairne & Widner, Jr., 1987), the majority of studies fail to find a generation effect with nonwords (e.g., McElroy & Slamecka, 1982; Nairne,

Pusen, & Widner, 1985; Payne et al., 1986). We first review the available evidence, and then discuss the theoretical questions that arise from the findings of these studies.

McElroy and Slamecka (1982) presented word pairs and nonword pairs to undergraduate participants at an English-speaking university, in either a read condition or a generate condition. In the read condition, the full word pair was visible on the screen, and in the generate condition the first word of the pair was visible, and the participants were required to follow a simple rule (such as generating an antonym or transposing letters) to generate the second word of the pair. The experiment also manipulated whether the trials were timed or self-paced. Participants were tested on word pair recognition (Experiment 1), and free recall (Experiment 2) and results showed a generation effect for the word pairs, but no generation effect for the nonword pairs. The authors interpreted their results as providing evidence that the semantic processing was necessary for the generation effect to occur. These findings were later replicated by Nairne et al. (1985).

Additionally, Payne et al. (1986) investigated whether these failures to find a generation effect with nonwords might be due to experimental design factors rather than to specific properties of nonwords. They presented word-word, nonword-word, word-nonword, and nonword-nonword pairs to English speaking participants who were instructed to either use a read task or a generate task to remember the word pairs. Over three experiments, they reliably observed generation effects when the response item was a word (i.e., for word-word and nonword-word pairs) but not when it was a nonword (i.e., for nonword-nonword and word-nonword pairs). Their findings supported the conclusion that generation effects are only obtainable when items to be recalled have a semantic representation.

The findings of these studies were of concern for the present study, because to teach novel L1 vocabulary words to adult learners, we had to select lexical items that were extremely low-

frequency and for which participants likely did not have pre-existing lexical or semantic representations.[5] If McElroy and Slamecka (1982) were correct that pre-existing semantic representations are necessary for generation effects to emerge, then we would not expect generation to be a beneficial training method for rare L1 vocabulary words.

However, in contrast to the studies above, Nairne and Widner, Jr. (1987) investigated whether failures to find generation effects with nonwords might be due to a lack of congruency between training and testing tasks. In the tasks described above, the training task required learners to manipulate surface features of the second word/nonword in the word pair, but the testing tasks (free recall and cued recall) asked learners to retrieve both the first and second word/nonwords from memory. Nairne and Widner proposed making training and testing tasks congruent would result in finding a generation effect for nonwords. They conducted two experiments to test this hypothesis, and in both studies, they reported robust generation effects for nonwords as well as the words. This study provided evidence that generation can be an effective technique for learning novel word forms that are not already in a participant's lexicon. Therefore, the present study ensures that the training and testing conditions are congruent, and provides an additional test of whether generation is effective for novel L1 word learning.

Although we do not have direct examples of the use of generation effects in training novel L1 vocabulary words, the studies reviewed in this section provide an interesting basis from which to begin to investigate this question. If we ultimately fail to find generation effects with this

[5] This was not of concern for Experiment 1, both because the words in that study were more common than the current stimulus set, and also because although the L2 words did not have pre-existing lexical representations, they did have pre-existing semantic representations (i.e., they were known concepts), and furthermore mapped to existing L1 lexical representations.

particular set of rare words, this would constitute further evidence that generation is not a suitable training technique for words that do not have pre-existing semantic representations. However, if generation effects do emerge for this particular set of words then this will constitute the first available evidence that generation is a suitable technique for training novel L1 vocabulary words. Futhermore, as in Experiment 1, it is an open question whether generation effects are modulated by semantic ambiguity. The next section discusses research related to this question.

3.3 Generation effects with semantically-ambiguous words

As mentioned above, it is an open question whether generation effects are modulated by semantic ambiguity. Other than evidence from the L2 vocabulary learning studies reviewed in Experiment 1, there is a paucity of research addressing this question. To the best of our knowledge, only one previous study by McElroy (1987), investigated generation effects with semantically-ambiguous words.

McElroy (1987) tested the effects of generation on memory for homographs paired with either rhyming cues (Exp 1; e.g., the homograph *duck* could be paired with *luck*) or semantically-related cues (Exp 2; e.g., the homograph *coal* could be paired with *mine*). Participants studied the words in one of two conditions: In the *read* condition, participants read lists of word pairs and in the *generate* condition participants saw the word pairs with a vowel removed from the target word (i.e., luck - d-ck), and were asked to complete the word by writing it in a booklet. Memory for targets was tested using a cued recall task in which half of the cues biased the homograph meaning that was trained, and half biased the untrained meaning. Results showed a benefit for the generation

condition in both experiments, but this effect was limited to cases in which the retrieval cues biased the meaning of the homograph encountered during training.

These results are of interest because they demonstrate that generation of one specific meaning of a semantically-ambiguous word can enhance recall for that specific meaning. However, the word fragment completion task used by McElroy (1987) does not require a learner to engage meaning representations, and so might not be as effective for encouraging semantic processing as the sentence generation tasks such as those used by Tokowicz and Jarbo (2009) and Eddington et al. (2012). Further, McElroy investigated recall for only one meaning of an already-known word. The present study will extend this research by investigating how learners perform when trained on multiple meanings of novel semantically-ambiguous words, and when using a sentence completion generation task to encourage semantic activation during learning.

3.4 Individual differences in ambiguous word learning

Successful language learning is not predicted by word characteristics alone. In fact, a large amount of variance in language learning outcomes is a product of word characteristics and learner characteristics (see Daneman & Merikle, 1996). This section addresses the later part of the equation: how individual differences affect vocabulary learning in general, and the generation effect and ambiguous word learning in particular.

Learners vary greatly in their cognitive abilities, including differences in executive functioning, inhibitory control, WM, and attention. Language learning, comprehension, and production place demands on these abilities, and these demands are heightened by the challenges presented by semantic ambiguity. To the extent that learners vary in cognitive abilities such as

WM and executive control, they may be more or less successful in learning and processing ambiguous words. A number of studies have explored this area, and we review several in the following section.

Individual differences in WM capacity are important for semantic ambiguity processing and resolution (e.g., Bornkessel, Fiebach, & Friederici, 2004; Gunter, Wagner, & Friederici, 2003; Miyake, Just, & Carpenter, 1994). When readers encounter a word with multiple meanings, those with greater WM abilities may be better able to hold the multiple meanings in memory. Miyake, Just, and Carpenter (1994) demonstrated that when readers encounter an ambiguous word in a sentence that is not disambiguated until several words later, those with higher WM span are better able to maintain multiple interpretations of the ambiguous word until they reached the disambiguating context than readers with lower WM span.

Additionally, individual differences in executive control may affect how learners process the multiple meanings of ambiguous words. For example, in Experiments 2 and 3 of their study, Lev-Ari and Keysar (2014) tested whether individual differences in bilinguals' executive control predicted how they perceived the different meanings of homonyms and polysemes. They reported that individuals with poorer inhibitory control were less able than individuals with better executive control to perceive differences in meaning between homonymous word meanings, and additionally were less able to perceive similarities in the related meanings of polysemes. In the context of the current study, this suggests that participants with weaker ability to suppress task-irrelevant information (i.e., higher Simon scores), may not display as strong a disadvantage for homonyms because they are not as aware of competing meanings as participants with stronger ability to suppress task-irrelevant information (i.e., lower Simon scores).

3.5 Experiment overview

Experiment 2 examined how the use of generation during vocabulary training impacted the learning of rare L1 vocabulary words that were unambiguous, or had ambiguous homonym or polyseme meanings/senses. Participants were trained on novel L1 words and their definitions, and then asked to practice this material by either reading and retyping an experimenter-generated sentence that contained the target word or generating their own semantically-appropriate sentence containing the target word. We examined how well participants were able to recall the trained words on tests of free recall and meaning production immediately after training and after a one-week delay. Our specific research questions were: 1) is generation beneficial for learning semantically-ambiguous words, 2) is generation more beneficial for some word types than others? and 3) is generation is more or less effective for learners depending on WM and inhibitory control?

Based on previous research, we expected to find a generation effect such that words trained with sentence generation were remembered better than words trained with sentence reading, both immediately after testing and after a one-week delay. Furthermore, we expected that generation may be more helpful for some words than others. Based on the predictions of the SSD, we predicted that polysemes and unambiguous words would benefit from generation, but homonyms would be negatively affected by generation. Additionally, we predicted that individual differences in WM and inhibitory control would interact with both semantic-ambiguity and generation. As in Experiment 1, we predicted that individuals who are higher in inhibitory control would be better able to inhibit competition from irrelevant meanings of ambiguous words and focus on the relevant meanings, resulting in a reduced homonymy disadvantage relative to individuals who are lower in inhibitory control. Similarly, we predicted that individuals with a higher WM span would be better

able to store and process multiple meanings for a word, and thus would show a reduced ambiguity disadvantage relative to individuals with lower WM span.

3.6 Methods

3.6.1 Participants

A total of 30 native English speakers were tested for this experiment. Data from three participants were excluded from final analyses due to equipment malfunctions during training and testing, and data from one additional participant were excluded for low accuracy on the Waters Reading Span test, leaving data from 26 participants for analyses. All participants were 18 years and older, right-handed, with normal or corrected-to-normal vision, and were recruited from the Psychology Department Subject Pool and compensated with class credit and a $7.00 cash bonus if they completed both sessions. Participation time was approximately 2 hours total, divided equally between two sessions.

3.6.2 Design

This study used a 3 word type (unambiguous, homonym, polyseme) x 2 training condition (sentence reading, sentence generation) x 2 session (session 1, session 2) within-subjects design.

3.6.3 Stimuli

Experiment 2 used 40 very low frequency English words. These words consisted of 20 nouns selected from stimuli created by Frishkoff, Collins-Thompson, Perfetti, and Callan (2008), and 20 adjectives selected from stimuli created by Balass (2011). None of the words in this set had more than one meaning (this is likely because rare words generally have few meanings, e.g., Balass, 2011), although approximately half were polysemes and had two or three related senses according to WordNet (Parks, Ray, & Bland, 1998). To examine how adult learners acquire ambiguous words in their first language, we needed words that were likely to be unknown to learners but also included unambiguous words, polysemes, and homonyms. Therefore, we selected a set of 20 unambiguous words, and 10 words that were naturally polysemous, but to obtain very rare homonyms we had to create novel meanings for 10 additional unambiguous words. To accomplish this, we created novel unrelated meanings for these words by selecting meanings from the unused words in the original stimuli sets. This ensured that meanings are actual word meanings belonging to words of similarly low frequency. To ensure that new meanings were not related to the existing meanings, three native English speakers independently reviewed the original and novel meanings and indicated whether word meanings seemed to be related. All raters independently agreed that none of the novel word meanings were related to the original meanings.

We extracted definitions for the rare words from online dictionaries (e.g., Merriam Webster online dictionary, dictionary.com, WordNet), and edited to be concise. Words were balanced so that the number of words per definition was matched across word type for definition 1: $F(37) =$

3.04, p = .06,[6] and definition 2: $F(18) = 1.77$, p = .20, and were furthermore balanced for part of speech across word type. Table 21 contains example words and definitions for Experiment 2.

Table 11. Example stimuli for Experiment 2

Word	Word type	Definition 1	Definition 2
irenic	unambiguous	tending to promote peace	-
sapid	polyseme	agreeable taste or flavor	agreeable to the mind
reboant	homonym	marked by reverberation	highly absorbent

Two context sentences were created for each unambiguous word, and four context sentences were created for each ambiguous word (i.e., two sentences per word meaning). To ensure that context sentences accurately captured key meaning features of the words, we collected normative ratings for each sentence. Appendix D describes the norming procedures for these sentences and presents the sentences and their ratings. Sentence ratings were matched across word types for both sentence 1: $F(37) = 0.44$, p = .65, and sentence 2: $F(17) = 1.84$, p = .19.

3.6.4 Procedure

The experiment consisted of two sessions of 1 hour each, spaced one week apart. In Session 1, participants received training on rare English words and their definitions. Following training, participants completed a free recall test in which they were instructed to type any words that they remembered from training. This task encouraged participants to use retrieval to enhance memory

[6] Because this test revealed a marginally significant result (p = .06), we included definition length in final regression models as a control variable where appropriate.

(the "testing effect", e.g., (Pyc & Rawson, 2010; Roediger & Karpicke, 2006), and provided a measure of learning immediately after training. Next, participants completed a meaning production test which assessed their ability to produce a word meaning when presented with the word as a prompt.

During Session 2, which occurred after a one-week delay, participants again completed the meaning production task. Participants also completed two individual difference measures: the Waters Reading Span test (Waters & Caplan, 1996) and the Simon task (Simon & Wolf, 1963). Finally, participants completed a Language History Questionnaire to collect relevant language background information. Table 12 provides a summary of the experiment procedures and timeline.

Table 12. Experiment 2 timeline

Session	Day	Tasks
Session 1	Day 1	Training 1: Familiarity check Word + definition (2x per word) Training 2: Read or generate sentence (1x per word) Testing 1: Free recall Testing 2: Meaning production
Session 2	Day 8	Testing 3: Meaning production Testing 4: Individual difference measures Testing 5: Language History Questionnaire (LHQ)

3.6.4.1 Familiarity check

To ensure that the novel words were not already known to participants, the initial training session began with a familiarity check (e.g., Balass, 2011). Trials began with a 500ms blank screen, following which words appeared in random order in the center of the screen for 1000 ms. After that time, the word disappeared from the screen and was replaced with a question mark. When the question mark appeared, participants were instructed to press a button with their left index finger

if they were *not* familiar with the word or a different button with their right index finger if they were familiar with the word. After participants entered a response the program advanced to the next trial. E-Prime was used to present stimuli and collect responses.

3.6.4.2 Vocabulary training

After completing the familiarity check, participants moved on to training, which followed the same procedures as Experiment 1. Two training orders were used to counterbalance whether words appeared in the *read* or *generate* training condition. Participants were randomly assigned to one of these two orders. Training took approximately 50 minutes to complete.

3.6.4.3 Free recall

After completing training, learners were asked to complete a free recall task in which they were instructed to type every target word they could remember, in any order. They were instructed to type a word even if they only remembered a part of the word and even if they were not confident about how to spell the word.

A trained coder compared the free recall responses that participants generated to the original stimuli. If all letters were present and in the correct order, a word was awarded one point. Partial credit (.5 points) was also awarded if at least one syllable was entered in the correct position. Other responses were given zero points. A second independent coder verified the accuracy of the scoring. There were no missing data for this task.

3.6.4.4 Meaning production

After completing the free recall task, learners were asked to complete a meaning production task in which they were presented with a target word they had encountered during training and

were asked to type in the definition. Words were presented by e-Prime in random order and remained on the screen until participants made a response. e-Prime collected responses and response times.

One trained coder scored the response data by comparing responses to the original training definitions as well as first-order synonyms extracted from the online Merriam-Webster Thesaurus for the content words from the original definitions. The coding protocol was adapted from Balass (2011). Responses were scored as correct and awarded one point if they accurately captured the meaning of the word and included at least one content word from the original definition. Responses were scored as partially correct and a half point was awarded if the response captured some, but not all, of the meaning of the word, and included synonyms of the content words from the original definition. Responses were scored as incorrect if the response did not capture the meaning of the word and did not include any content words or synonyms of content words from the original definition. A second coder examined a randomly-selected subset of 20% of all trials to verify coding accuracy. There were no missing data for this task.

3.6.4.5 Individual difference tasks

After completing the meaning production task, participants moved on to complete the individual difference measures, which followed the same procedures as Experiment 1. We examined accuracy on the Waters sensibility judgements to ensure participants were paying attention to the task. We screened participant responses to ensure that all participants correctly responded to at least 70% of all sensibility judgements. We found that one participant only correctly responded to only 54% of sensibility judgements, and so we removed data from this participant from all analyses, leaving data from 26 participants for analyses.

3.7 Results

3.7.1 Statistical approach

As in Experiment 1, the analyses described in this section used linear mixed-effects models, which allowed us examine subject and item effects simultaneously (e.g., Baayen et al., 2008). The specific type of model that we used varied according to the distribution of the dependent variable. We used glmer to model accuracy data when the responses followed a binomial distribution (either 0 or 1). We used cumulative link mixed models to model accuracy data when the responses were ordinal (either 0, .5, or 1) (Christensen, 2019).

Ninety-five-percent confidence intervals are reported for all fixed effects. We entered random effects for subjects and items, and specified the maximal random effect structure for which models would still converge.

Because we had specific hypotheses about the roles of both WM and inhibitory control, but these variables are known to measure similar concepts, tests of multicollinearity (condition number and VIF) were conducted to determine whether multiple individual difference measures could be entered in to the same model. If we determined this was acceptable, we selected the Waters measure (total span or set size span) with the higher correlation with the dependent measure and entered that measure and Simon score into the final model as predictors.

In all analyses, word length was mean-centered to eliminate nonessential multicollinearity (Frank, 2011). Simon scores and Waters total span were scaled to match other variables by dividing by 10. Because the read and generate portion of training was self-paced, we collected response times to be used as a fixed effect to control for variation introduced due to participants studying a word for varying amounts of time. This variable, study time during training (ms), was scaled to

match the other variables by dividing by 1,000. The baseline conditions were coded as follows: unambiguous words were the baseline to which homonyms and polysemes were compared, and sentence reading was the baseline to which sentence generation was compared.

All analyses were conducted in R 3.6.0 using version 1.1.21 of the lme4 package (Bates et al., 2014), version 3.1.0 of the lmerTest package (Kuznetsova et al., 2017) to assess significance, and version 4.25 of the ordinal package (Christensen, 2019) to fit mixed-effects models with ordinal dependent variables.

3.7.2 Familiarity check

To establish whether the stimuli we selected were sufficiently rare as to be unfamiliar to participants, the initial training session included a familiarity check. Following procedures from Balass (2011), words that were consistently rated as familiar by participants were removed from further analyses. We calculated the percent of participants that indicated familiarity with each word, and examined the distribution of this variable. Three words emerged as clearly more familiar than the rest of the set, with more than 40% of participants indicating familiarity: evanescence (42% familiar), levity (50% familiar), and discinct (62% familiar). Evanescence and levity were removed from all further analyses. It seems unlikely that more than 60% of participants were actually familiar with the extremely rare word *discinct*, and more likely that given the rapidity of the task (i.e., words were only viewed for 1000 ms), participants instead believed they had viewed the common and orthographically-similar word *distinct.* Because the goal of this experiment is to examine novel word learning, we elected to also remove *discinct* from analyses because we were unsure if participants would proceed as if they were learning a new word or a new meaning to a known word (i.e., distinct).

3.7.3 Vocabulary training analyses

Before conducting analyses on any of the outcome measures, we first examined participant accuracy during training following the same procedures outlined above in Experiment 1. For Experiment 2, one participant was removed from all further analyses because of errors on read trials, leaving data from 26 participants remaining for analyses. Mean accuracy on read trials for the remaining participants in Experiment 2 was 98%. All participants met criteria for generate trials.

3.7.4 Free recall results

This section describes the results of analyses of the free recall response data. We began by examining descriptive statistics for free recall accuracy. The mean proportion correct for all words was .16 (*SD* = 0.09). The mean accuracy for words trained with sentence reading was .16 *(SD* = .34), and the mean accuracy for words trained with sentence generation was .17 (*SD* = .35). Mean accuracy was .13 (*SD* = .31) for unambiguous words, .21 for homonyms (*SD* = .40), and .20 for polysemes (*SD* = .38).

We examined Pearson correlations between free recall accuracy and psycholinguistic variables and individual difference measures, and found that all three individual difference measures were significantly correlated with free recall accuracy (Waters set size span: r = .09, Waters total span: r = .10, and Simon score: r = .11). Because the two Waters variables are strongly correlated with one another (r = .77), we decided to use the Waters measure with the higher correlation with free recall accuracy: Waters total span. Because we had specific hypotheses about both Waters and Simon scores, we wanted to include both in the final models. To make sure we

could clearly test the effects of each of these as independent predictors we assessed the risk of multicollinearity between these two predictors. To do this, we examined the condition number. The condition number that indicated that there was standard collinearity ($\kappa = 7.60$), which falls below the threshold for potentially harmful collinearity ($\kappa = 30$; Baayen, 2008). We then examined VIF, and discovered values of 1.08 for both Waters and Simon, which is below the point at which VIF becomes problematic (VIF = 5.00). Therefore, we proceeded to build a final model that included both Waters and Simon.

We constructed a cumulative link mixed model (clmm) to examine free recall accuracy. The fixed effects of theoretical interest were training condition (read or generate), word type (unambiguous, homonym, polyseme), Simon score, Waters total span, and the interaction of word type by training condition by Simon score by Waters total span. Word length in letters was included as a control variable. We specified the maximal random effects structure for this model, including random intercepts for both subjects and items, and included by-subject random slopes for the interaction term, for the individual difference measure (Simon score), and word length. There were no missing data.

The model equation and fixed effects estimates for Model 5 are presented in Table 13, and the random effects estimates are presented in Table 14. Of the fixed effects of theoretical interest, there was a significant main effect of Simon score, such that participants with higher inhibitory control recalled words more accurately. Furthermore, there was an interaction of Simon score and generation, $b = 0.40$, $SE = 0.18$, $z = 2.27$, $p = .02$, such that participants with higher inhibitory control actually recalled more words trained in the read condition than in the generate condition, and conversely, participants with lower inhibitory control recalled more words trained in the generate condition than in the read condition. However, this two-way interaction was qualified by

a significant three-way interaction of homonyms, generation, and Simon score, b = -0.61, SE = 0.28, z = -2.16, p = .03. We probed this interaction using the *effects* package in R, which uses the regression equation to generate estimated probabilities of producing a correct response at designated Simon score values for both read and generate conditions. In this case we chose to probe the interaction at 1 SD above and below the mean Simon score (see Figure 10). Low Simon scores indicate better suppression of task-irrelevant information, and so Figure 10 shows that participants who were better able to suppress task-irrelevant information were more accurate for unambiguous words when trained with sentence reading, whereas participants who were less able to suppress task-irrelevant information were most accurate on unambiguous words when trained in the sentence generation condition. There were no significant effects of Waters total span, and the four-way interaction of training condition, word type, Simon, and Waters was not significant.

Table 13. Fixed effects estimates for Model 5, free recall

			95% CI				
	Estimate	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	0.57	0.42	-0.25	1.40	1.37	0.17	
Polysemes	0.57	0.38	-0.18	1.32	1.50	0.13	
Generate condition	0.14	0.28	-0.41	0.68	0.49	0.63	
Simon score	-0.39	0.15	-0.68	-0.10	-2.65	0.01	**
Waters total span	0.23	0.21	-0.20	0.65	1.05	0.29	
Word length	-0.02	0.08	-0.18	0.14	-0.27	0.79	
Homonyms*Generation	-0.04	0.47	-0.96	0.88	-0.09	0.93	
Polysemes*Generation	-0.15	0.43	-1.00	0.70	-0.34	0.74	
Homonyms*Simon	0.34	0.21	-0.07	0.75	1.64	0.10	
Polysemes*Simon	0.16	0.19	-0.21	0.54	0.86	0.39	
Generate*Simon	0.40	0.18	0.05	0.75	2.27	0.02	*
Homonyms*Waters	0.25	0.31	-0.35	0.86	0.82	0.41	
Polysemes*Waters	-0.06	0.29	-0.63	0.50	-0.22	0.82	
Generate*Waters	-0.28	0.27	-0.80	0.25	-1.04	0.30	
Simon*Waters	0.21	0.21	-0.20	0.62	1.02	0.31	
Homonyms*Generation*Simon	-0.61	0.28	-1.17	-0.06	-2.16	0.03	*
Polysemes*Generation*Simon	-0.35	0.26	-0.87	0.16	-1.34	0.18	
Homonyms*Generation*Waters	-0.11	0.42	-0.94	0.72	-0.26	0.80	
Polysemes*Generation*Waters	0.34	0.40	-0.43	1.12	0.87	0.39	
Homonyms*Simon*Waters	-0.15	0.29	-0.72	0.42	-0.51	0.61	
Polysemes*Simon*Waters	-0.30	0.25	-0.80	0.19	-1.19	0.23	
Generate*Simon*Waters	-0.13	0.24	-0.61	0.34	-0.56	0.58	
Homonyms*Generation*Simon*Waters	-0.06	0.36	-0.76	0.64	-0.17	0.87	
Polysemes*Generation*Simon*Waters	0.17	0.33	-0.48	0.81	0.51	0.61	

Model equation. Model 5 <- clmm(ACC~ 1 + WordType*TrainingCondition*Simonscore*waterstotcorr + wordlength + (1|Subject)+(1|EnglishWord), data=freerecall2, link = 'logit', threshold = 'flexible')

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 14. Random effects estimates for Model 5, free recall

	Variance	*SD*
Item intercept	0.31	0.55
Subject intercept	0.19	0.43

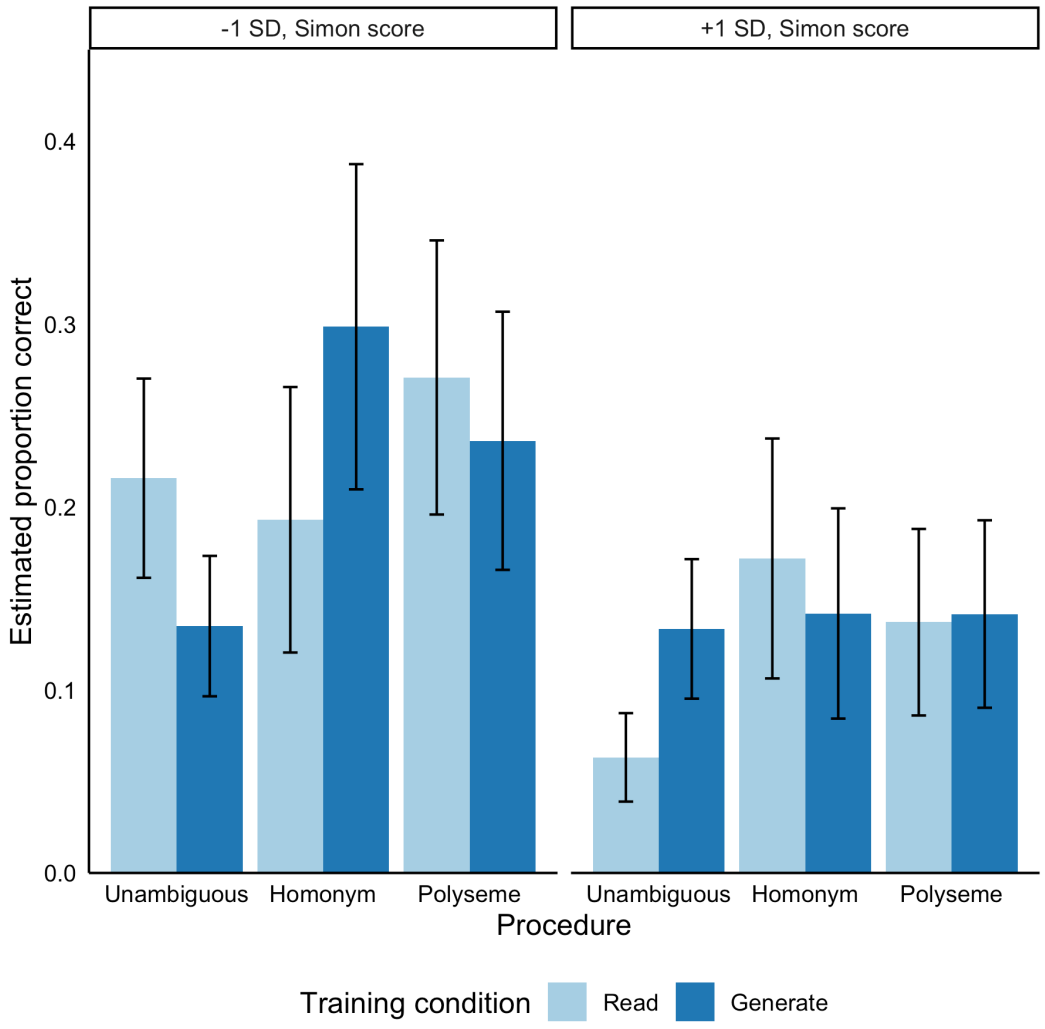

Figure 10. Estimated proportion correct for free recall by word type, training condition, and Simon score. Error bars represent standard error of the mean.

3.7.5 Meaning production results

We examined accuracy data from the meaning production task for both Session 1 and Session 2, using linear mixed-effects models. The fixed effects of theoretical interest were training condition (read or generate), word type (homonym, polyseme, unambiguous), Simon score, Session (1 or 2) and the interaction of word type by training condition by WM and inhibitory control. Fixed effects were also entered for the following control variables: study time during training, training order (trained first vs. second), and training sentence length. We specified the maximal random effects structure justified by our model for which the model would converge. There were no missing data for these analyses.

We began by examining the proportion of correct responses for Session 1 and Session 2. Overall, the mean proportion correct in Session 1 was .41 (SD = 0.16), and the mean proportion correct in Session 2 was .27 (SD = 0.12). The mean accuracy for words trained with sentence reading was .30 (*SD* = .43), and the mean accuracy for words trained with sentence generation was .38 (*SD* = .46). Mean accuracy was .37 (*SD* = .45) for unambiguous words, .31for homonyms (*SD* = .44), and .33 for polysemes (*SD* = .44).

We examined Pearson correlations between proportion correct on the meaning production task, psycholinguistic variables, and the individual difference measures. As before, all three individual difference measures were correlated with meaning production accuracy (Waters set size span: r = .12, Waters total span: r = .18, and Simon score: r = .19). Because the two Waters variables are strongly correlated with one another (r = .77), we decided to use the Waters measure with the higher correlation with free recall accuracy: Waters total span. Because we had specific hypotheses about both Waters and Simon scores, we wanted to include both in the final models.

To make sure we could clearly test the effects of each of these as independent predictors we assessed the risk of multicollinearity between these two predictors, using the procedures outlinedin Experiment 1. The condition number indicated that there was standard collinearity ($\kappa = 7.60$), and we observed VIF values of 1.08 for both Waters and Simon, which is well below the point at which VIF becomes problematic (VIF = 5.00). Therefore, we proceeded to build a final model that included both Waters and Simon.

We constructed a clmm to examine meaning production accuracy. The fixed effects of theoretical interest were word type, training condition, Session, Waters total span, and the interaction of word type by training condition by Simon score by Waters total span. Word length, definition length, meaning number (first or second trained), and study time during training were included as control variables. We specified the maximal random effects structure for which this model would converge, which meant including random intercepts for both subjects and items as well as subject random slopes for the interaction of word type and training condition.

The model equation and fixed effects estimates for Model 6 are presented in Table 15, and the random effects estimates are presented in Table 16. Of the fixed effects of theoretical interest, there was a significant effect of Simon score, such that lower Simon scores were associated with significantly higher meaning production accuracy, $b = -0.33$, $SE = 0.12$, $z = -2.78$, $p = .01$. There was a significant main effect of Waters total span, $b = 0.40$, $SE = 0.20$, $z = 2.04$, $p = .04$, such that higher WM scores were associated with significantly higher meaning production accuracy. There was a significant effect of Session, such that performance was significantly better in Session 1 than Session 2, $b = -0.94$, $SE = 0.09$, $z = -10.23$, $p < .001$. None of the interactions, including the four-way interaction of word type, training condition, Simon, and Waters, were significant.

Table 15. Fixed effects estimates for Model 6, meaning production

			95% CI				
	Estimate	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	-0.68	0.57	-1.79	0.44	-1.19	.23	
Polysemes	-0.29	0.50	-1.26	0.69	-0.58	.56	
Generate condition	0.35	0.20	-0.03	0.73	1.79	.07	†
Simon score	-0.33	0.12	-0.56	-0.10	-2.78	.01	**
Waters total span	0.40	0.20	0.02	0.79	2.04	.04	*
Session	-0.94	0.09	-1.12	-0.76	-10.23	<.001	***
Word length (letters)	-0.26	0.13	-0.52	0.00	-1.94	.05	†
Definition length (words)	0.08	0.03	0.03	0.13	3.06	<.001	**
Meaning	-0.16	0.11	-0.38	0.06	-1.41	.16	
Study time	0.02	0.02	-0.02	0.07	1.04	.30	
Homonyms*Generation	0.56	0.34	-0.11	1.23	1.63	.10	
Polysemes*Generation	-0.13	0.28	-0.67	0.41	-0.48	.63	
Homonyms*Simon	-0.12	0.13	-0.38	0.14	-0.89	.38	
Polysemes*Simon	0.01	0.10	-0.19	0.21	0.06	.95	
Generate*Simon	0.05	0.11	-0.17	0.27	0.47	.64	
Homonyms*Waters	0.01	0.20	-0.38	0.40	0.05	.96	
Polysemes*Waters	-0.01	0.16	-0.33	0.32	-0.03	.97	
Generate*Waters	0.12	0.19	-0.25	0.49	0.62	.53	
Simon*Waters	0.14	0.15	-0.16	0.45	0.94	.35	
Homonyms*Generation*Simon	0.15	0.20	-0.23	0.53	0.76	.44	
Polysemes*Generation*Simon	0.18	0.15	-0.12	0.48	1.16	.25	
Homonyms*Generation*Waters	-0.23	0.32	-0.86	0.40	-0.72	.47	
Polysemes*Generation*Waters	0.10	0.27	-0.42	0.63	0.38	.71	
Homonyms*Simon*Waters	-0.26	0.19	-0.62	0.11	-1.37	.17	
Polysemes*Simon*Waters	-0.06	0.15	-0.34	0.23	-0.41	.68	
Generate*Simon*Waters	-0.14	0.16	-0.45	0.17	-0.91	.37	
Homonyms*Generation*Simon*Waters	0.13	0.26	-0.39	0.64	0.48	.63	
Polysemes*Generation*Simon*Waters	-0.09	0.21	-0.50	0.33	-0.41	.68	

Model equation. Model 6 <- clmm(ACC ~ 1 + WordType*ReadorGen*SimonScore*waterstotcorr + Session + LenEng+ DefLen+ meaning + study time + (1|Subject)+(1|EnglishWord) + (0 + WordType*ReadorGen|Subject),data=meanprod.both, link = 'logit', threshold = 'flexible')
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 16. Random effects estimates for Model 6, meaning production

	Variance	*SD*
Item intercept	1.38	1.17
Unambiguous|Subject	0.71	0.84
Polysemes|Subject	1.23	1.11
Homonyms|Subject	0.52	0.72
Generate|Subject	0.27	0.52
Polysemes*Generate|Subject	1.07	1.03
Homonyms*Generate|Subject	0.54	0.73
Subject intercept	<0.001	<0.001

3.8 Discussion

This experiment examined how the use of generation affected novel English word learning for rare English words. We hypothesized that we would find an overall benefit of generation for both free recall and meaning production tasks, but that this benefit would be influenced by word type. Specifically, based on predictions derived from the SSD hypothesis (Armstrong & Plaut, 2016), we predicted that polysemes and unambiguous words would benefit from generation, but homonyms would be negatively affected by generation. Furthermore, we expected that performance would be better for immediate (Session 1) than delayed (Session 2) testing. Finally, we predicted that individual differences would modulate the translation-ambiguity effect, and specifically that individuals with higher inhibitory control (as measured by lower scores on the Simon task) would exhibit a reduced homonymy disadvantage, and individuals with higher WM would perform better than individuals with lower WM on ambiguous words specifically. Furthermore, we predicted that Waters and Simon would interact with each other, such that best performance would be observed for individuals with high WM and better ability to suppress task-irrelevant information.

Results from the free recall task partially supported these predictions. In these data we observed a significant main effect of Simon score, such that participants higher inhibitory control recalled words more accurately. Furthermore, there was an interaction of Simon score and generation, such that participants with higher inhibitory control recalled more words trained in the read condition than in the generate condition, and conversely, participants with lower inhibitory control recalled more words trained in the generate condition than in the read condition. However, this two-way interaction was qualified by a three-way interaction of homonyms, generation, and Simon score, which revealed that the two-way interaction reported above was primarily driven by differences in unambiguous words. Specifically, participants with higher inhibitory control were more accurate for unambiguous words when trained with sentence reading, whereas participants with lower inhibitory control were most accurate on unambiguous words when trained in the sentence generation condition.

The finding of an interaction between word type, Simon score, and the generation effect is novel. We first discuss some past research that has investigated interactions between individual difference and the generation effect, and then turn to a discussion of how the interaction between word type and generation conforms or does not conform to our hypotheses.

Relatively few studies have investigated individual differences in the generation effect, but the available literature suggests that generation is not equally effective, nor perhaps even suitable, for all learners or all word types. For example, Schindler, Schindler, and Reinhard (2019) investigated whether the success of the generation effect depended on an individual's *need for cognition* – the desire to engage in effortful learning activities and to enjoy being challenged during learning. They used a word generation task in which participants were given a context word and a semantically-related target word. Half of the target words were complete and participants were

instructed to study them. The other half of the words were incomplete, and participants were instructed to generate the missing letters. When tested on cued recall of the target words, participants with a higher need for cognition showed a greater generation effect than participants with a lower need for cognition.

The results of Schindler et al. (2019) are generally in agreement with our finding of a reduced generation effect for participants higher in inhibitory control. Although inhibitory control and need for cognition are distinct concepts, they share some overlap in that they both reflect aspects of general cognitive abilities. Therefore, it is possible that generation is best suited for participants with weaker cognitive abilities, and that participants with stronger cognitive abilities do not benefit enough to make the extra time and effort it takes to generate learning material worthwhile. Alternatively, it may be that participants with greater cognitive abilities are simply already using advanced learning strategies like generation, and so for these participants the difference between the sentence reading and sentence generation tasks might not have been as pronounced as for participants with weaker cognitive abilities.

The three-way interaction of word type, generation, and Simon score is especially interesting given the absences of a two-way interaction of word type and generation, which was predicted by our hypotheses. This suggests that the predicted differences in the efficacy of generation may only appear for participants who are higher in inhibitory control, making individual differences necessary for this interaction to emerge. However, the way in which the interaction of word type and generation emerged for participants higher in inhibitory control did not fully follow our hypotheses. We predicted that unambiguous words and polysemes would benefit from generation, whereas homonyms would be harmed by generation. This turned out not to be the case: for participants lower in inhibitory control only, unambiguous words benefitted

from generation, but generation was detrimental for unambiguous words for participants with higher inhibitory control. There were no differences in the effectiveness of generation for either group of participants for polysemes. And contrary to predictions, generation actually benefitted homonyms, although only for participants with higher inhibitory control. One potential explanation for these unexpected results is the nature of the task. Results from free recall are known to show a reversal of the effects of translation ambiguity: whereas there is a translation-ambiguity disadvantage in production tasks, there is a translation-ambiguity advantage in free recall (Degani, Tseng, & Tokowicz, 2014; Ekves, 2014). This is thought to be because ambiguous words typically show a disadvantage in processing (i.e., they are harder to learn), and this difficulty makes them distinctive to learners (Loess & Duncan, 1952). It is possible that a similar effect occurs with semantically-ambiguous words, but only appears for participants higher in inhibitory control.

Results from the meaning production task partially supported our predictions. Consistent with past research (e.g., Bertsch et al., 2007), there was an overall generation effect that was robust across all word types. We observed the expected effect of Session, such that performance was better on the immediate testing than the delayed testing. Consistent with findings from free recall, there was an effect of Simon score, such that participants with higher inhibitory control performed better overall. There was also a marginally significant (p = .09) interaction of word type and training condition, such that homonyms benefited more from generation than did polysemes or unambiguous words. This interaction did not reach significance and should not be interpreted as strong evidence for our hypotheses, but nevertheless may be suggestive of the fact that some word types may respond to generation in different ways. This is also notable because the same finding emerged in the free recall data.

Overall, the results of Experiment 2 provide limited support for the hypothesis that generation effects vary by word type, although as in Experiment 1, the effects did not manifest in the predicted directions. There are a number of reasons why this may have been the case. One important factor to keep in mind is that accuracy in this experiment was low for both free recall and meaning production. Learners may have struggled to learn the rare words trained in this experiment because they were simply too difficult, and that may have resulted in such low accuracy that it was difficult to find differences between conditions unless individual differences were considered. By considering individual differences we were in some ways creating subsets of participants who did well on the test (had higher accuracy) and participants who did less well (had lower accuracy). When we did this, we were able to see some of the effects we predicted, and these were more pronounced for participants in the higher accuracy groups. Future research should investigate methods of increasing overall learning and accuracy, which we will discuss further in the next section.

Finally, there is one additional possibility for why we did not observe the exact pattern of interactions of generation and word type that we predicted. One of the ramifications of teaching L1 speakers rare L1 words that they have not previously encountered, is that this task requires participants to establish a new meaning representation, as well as learn a new word form. This is in contrast to Experiment 1, in which participants were learning new L2 word forms that mapped to existing meaning representations and L1 word forms. There are a number of reports in the literature of failures to find a generation effect when target words do not correspond to a known semantic representation (e.g., McElroy & Slamecka, 1982a) or when the words used are low-frequency (e.g., Nairne et al., 1985). However, both Nairne and Widner, Jr. (1987) and Johns and Swanson (1988) later suggested that these findings may have been confounded due to a lack of

congruency between the testing task and the training task. And furthermore, although some of the words in the stimulus set do not correspond to existing semantic representations (for example, it is difficult to imagine a synonym for *dumose*: *filled with bushes*), other words have common synonyms with which learners may have been familiar, and thus the novel words could be mapped to these existing similar meaning representations (for example, *evanescence* shares meaning features with the common word *disappearance*).

In contrast to the predictions of McElroy and Slamecka (1982) and Nairne et al. (1985), we did find strong evidence of a generation effect in both free recall and meaning production. This is of interest, because it adds to the small amount of literature that has investigated whether generation effects can occur in the absence of pre-existing semantic representations, and demonstrates that pre-existing semantic representations are not necessary for generation effects to emerge. However, there remains the possibility that generation effects manifest differently for low-frequency words, and this might be one reason why we failed to find the hypothesized patterns of interaction of word type and training condition. Future research should investigate this possibility.

3.8.1 Future directions

This is the first study to examine whether generation effects manifest differently across different types of semantically-ambiguous and unambiguous words. As such, there are a number of interesting future directions suggested by this study. In the following section, we discuss three main future directions: limitations of the current study, potential moderating variables for future exploration, and applications for laboratory and classroom studies of L1 vocabulary learning.

First, it is important to acknowledge that there are a number of psycholinguistic variables that were not investigated in this study that are known to affect the learning and processing of

semantically-ambiguous words. These include meaning dominance, number of meanings, and more fine-grained information about meaning relatedness (i.e., the semantic similarity of the multiple meanings of ambiguous words). Because this was the first exploration of whether generation effects are modulated by ambiguity, we were unable to include all variables of interest because our models contained a large number of variables of potential theoretical importance, and larger datasets would be advisable to test the effects of additional predictor variables. We hope that this study will provide a starting point for research in this area, and allow future research to develop hypotheses that may include testing the effects of some of these potentially important variables. For example, there was an interesting trend in the meaning production data that suggested that homonyms may benefit from generation more than other types of ambiguous and unambiguous words. This was contrary to hypotheses, and future research might investigate whether this result might be an artifact of some uncontrolled for source of variation.

Second, as discussed above, there is the possibility that generation effects manifest differently for low-frequency words and words without pre-existing semantic representations. A particularly interesting future direction would be to directly test both of these possibilities. This is complicated by the fact that it is not easy (or perhaps even possible) to truly teach common L1 words to adult learners, because neurotypical adult learners will have already acquired these words by the time they adulthood. Studies in this area may have to be conducted with younger learner populations who have not yet acquired a rich L1 vocabulary.

Finally, there are some interesting results in this study that demonstrate that generation effects are impacted by individual differences in inhibitory control. This suggests that future research should take individual differences into account when using generation-based tasks. Furthermore, the finding that participants with higher inhibitory control did not benefit from

generation, whereas participants with lower inhibitory control did benefit from generation bears further investigation. First, it would be important to replicate this finding in other samples and with other tasks. If this finding does replicate (and corroborating evidence from Schindler et al. (2019) suggests that it will), then future laboratory and classroom approaches should consider using different training mechanisms for learners of varied cognitive abilities.

3.8.2 Conclusions

Overall, Experiment 2 contributed to our understanding of generation effects with semantically-ambiguous words. We found evidence in both free recall and meaning production tasks that generation effects do occur with semantically-ambiguous words, and mixed evidence that the strength of generation effects varies across unambiguous words, homonyms, and polysemes. Our results also provided the first evidence that individual differences impact generation effects, and we suggest a number of future directions based on this interesting result.

4.0 Experiment 3

This experiment builds on Experiment 2, and investigates generation effects with semantically-ambiguous L1 words that are either trained with definitions and context sentences (as in Experiment 2) or only context sentences. This question is of interest because previous research has demonstrated that training novel words with definitions and context sentences was more effective than with context sentences only (Bolger et al., 2008). However, it is unknown whether these findings hold for different types of ambiguous and unambiguous words, although there is some reason to expect that different types of ambiguous and unambiguous words may benefit from different combinations of definitions and context sentences during training. For example, a definition may be more helpful for learning polysemes than homonyms because polyseme senses are related, and precise definitions might illuminate fine differences in meaning better than context sentences. In contrast, homonyms might respond to training manipulations in the same ways as unambiguous words, and benefit the most from the combined use of definitions and context sentences. Furthermore, past research suggests that, at least for L2 words, the success of generation during learning may be impacted by whether definitions are used to train novel L2 words (e.g., Tokowicz & Jarbo, 2009). This study will be the first to extend this question to L1 vocabulary word learning, and ask if generation effects manifest differently when words are trained with definitions or context sentences. In the following section we first review the literature that informs these questions, and then turn to a discussion of the aims and predictions of this experiment.

4.1 Word learning from context

Research on word learning describes a process in which word forms and meanings are learned gradually and incrementally (e.g., Bolger et al., 2008; Frishkoff, Collins-Thompson, Perfetti, & Callan, 2008; Reichle & Perfetti, 2003). When learners first encounter a word, they establish a contextualized and incomplete semantic representation (e.g., Durso & Shore, 1991; Shore & Durso, 1990), which becomes decontextualized and complete over time, as the word is encountered multiple times in multiple contexts (Bolger et al., 2008; Nagy, Anderson, & Herman, 1987). The present work is guided by this *instance-based framework of word learning* (Bolger et al., 2008).

The majority of word learning is thought to be incidental; that is, word learning primarily occurs when readers encounter a word in a context sentence during reading, and without explicit instruction, they infer at least part of the meaning (Jenkins, Stein, & Wysocki, 1984; Nagy, 1995; Nagy et al., 1987; Nagy, Herman, & Anderson, 1985). Although implementing truly naturalistic reading conditions in the laboratory is not possible, we know that readers are predisposed to easily learn words from context (Swanborn & de Glopper, 1999), and so we have ample reason to expect that vocabulary training methods that make use of context sentences may be successful.

A number of research studies have supported context-based approaches to vocabulary learning. For example, Gipe (1980) compared four methods of teaching vocabulary to elementary school children: 1) an association method in which a novel word was paired with a known synonym (graphite – pencil-lead), 2) a category method that asked learners to sort known and unknown words into categories, 3) a context method in which learners read context sentences containing unknown target words, and 4) a dictionary method, in which participants looked up unknown words in a dictionary, and then practiced writing definitions and context sentences for the word.

Participants learned words using one of the four methods one time per word, with the words spread out over the course of eight weeks. At the end of eight weeks, performance on a fill-in-the-blank sentence task showed that learning was best for words in the context condition and worst for words in the definition condition.

However, not all contexts are reliable ways of learning word meanings. Beck, McKeown, and McCaslin (1983) demonstrated that although word learning can be accomplished through teaching words embedded in context sentences, not all contexts are equally beneficial for learners, and in fact some contexts can be misleading, unhelpful, or simply incorrect. The authors developed an evidence-based classification of four types of contexts: misdirective (contexts that direct readers to an incorrect assumption about word meaning), nondirective (contexts that are of no particular use to readers in learning the meanings of a word), general (contexts that provide information that allows readers to understand the general category in which a word belongs), and directive (contexts that lead the reader to the correct meaning of a word). They described a view of word learning in which the best contexts are directive, and furthermore point out that learners should encounter words repeatedly and in a variety of contexts.

Although traditionally vocabulary learning was thought to best be accomplished through the use of context sentences (Beck et al., 1983), incidental learning is not the only way to acquire novel vocabulary. One alternative method is the use of an explicit, definition-based approach. Whereas context-based approaches provide highly contextualized and specific knowledge about how a word is used, dictionary definitions provide learners with decontextualized and specific information about the core meaning of a word (Drum & Konopak, 1987)(Drum & Konopak, 1987). Research findings are mixed concerning the relative effectiveness of context sentences vs. definition use for vocabulary instruction. Although some studies have found definitions to be

unhelpful for word learning (e.g., Gipe, 1980; Nash & Snowling, 2006), other studies have found that definitions can be helpful (Fischer, 1994; Jenkins, Matlock, & Slocum, 1989), especially if they are paired with context sentences (e.g., Badgett, 2003; Bolger et al., 2008), or if the words being learned are previously completely unknown to the learner (e.g, Shore & Durso, 1990). Studies with children have demonstrated that providing definitions to elucidate the meanings of words encountered in context can facilitate word learning (Elley, 1989; Wilkinson & Houston-Price, 2013).

There is some evidence that these findings hold true for adult learners as well. For example, Bolger et al. (2008) conducted a word learning study that investigated learning words from definitions and context sentences. This study trained native English speakers on rare English words and manipulated whether they were presented with context sentences without a definition, or with context sentences and a definition. Each word was viewed four times, and words were presented either in the *varied* condition, in which each word had four different context sentences, or in the *repeated* condition, in which each word had one context sentence that was repeated four times. Whether they had a varied or repeated context sentence, the sentences were always randomly intermixed throughout training. Participants completed a meaning generation task in which they were shown a word on the screen, and asked to provide a one- or two-word definition. They found differences in overall accuracy depending on whether sentences were trained with a definition or not. If words were trained with a definition, there was no benefit of varied over repeated presentation, but if words were trained without a definition the varied condition yielded higher accuracy than the repeated condition. This study demonstrated that when definitions are present to provide explicit knowledge about decontextualized word meanings then it does not help to have

varied contexts. Conversely, when learners have not received an explicit definition it does help to have multiple instances of context to draw from.

There are a number of additional factors that may impact the effectiveness of context and definition training methods. These include the use of particular training strategies such as generation as well as interactions between learner characteristics and instructional methods. We turn now to a discussion of these factors.

Only one previous study has examined whether the use of context sentences and definitions during learning impacts generation effects during vocabulary learning. Badgett (2003) examined novel vocabulary learning using the generation effect, and trained undergraduates at an English-speaking university on unfamiliar English vocabulary words taken from materials created by Dempster (1987). Participants were assigned to one of three training conditions: 1) reading and rewriting definitions multiple times, 2) reading and copying a definition one time, and then generating a meaningful context sentence, and 3) reading context sentences and then generating a definition. Participants were tested on a cued-recall task immediately after learning, after a 48-hour delay, and after a 21-day delay. Badgett reported that the sentence generation group remembered the definitions for significantly more words than did the other two groups on all post-tests. Additionally, across all three posttests the read and copy group remembered the fewest definitions.

Badgett (2003) also assessed individual differences via a verbal intelligence test which required participants to select a synonym for a given vocabulary word (using a separate list than the stimulus set for the main study). The author reported strong correlations (*r* ranging from .46 to .56) between verbal intelligence and performance on all three assessments, which demonstrated the large impact that individual differences can have on vocabulary learning.

Pertinent to the aims of this study, Nash and Snowling (2006) examined vocabulary learning in children with poor vocabulary knowledge. They taught novel vocabulary words with either a definition-based approach or a context-sentence approach. They reported that definitions were significantly less helpful than context sentences for low vocabulary learners. The study did not include a high vocabulary comparison group, so it is unknown if context sentence are equally helpful for all learners. However, Jenkins et al. (1984) examined vocabulary acquisition in fifth grade children using context passages presented a varying number of times. They reported that better readers benefitted more from context than did poorer readers. Although both of these studies examined novel word learning with children, not adults, and furthermore examined different individual difference measures than the present study, they do provide some evidence that the effectiveness of training novel vocabulary words with context sentences vs. definitions may be affected by individual differences.

4.2 Experiment overview

This experiment builds on Experiment 2, and examines the effect of using generation with either definitions or context sentences to learn semantically-ambiguous words, which has not yet been investigated. A few previous studies, reviewed above, have investigated the relative benefits of training novel unambiguous vocabulary with definitions or context sentences, but this is the first study to extend this research to specifically examine ambiguous words, and the first to also examine the impact of sentence generation as well. The stimuli and experimental design for this study were identical to those of Experiment 2, other than the addition of a new training

manipulation (definition vs. context sentence training), and the addition of a new outcome measure (a forced-choice sentence completion task).

Our research questions for this experiment were: 1) Does training novel L1 vocabulary with definitions lead to better learning of semantically-ambiguous words than training L1 vocabulary with context sentences? 2) Does the effectiveness of training vocabulary with definitions and context sentences dependent on either semantic ambiguity or the use of generation during training? and 3) Are the training methods in questions 1 and 2 more or less effective for individuals who are higher vs. lower in WM and inhibitory control?

Our predictions for this experiment included the predictions described in Experiment 2, as well as several additional hypotheses about the impact of the definition vs. sentences condition. Based on previous research, we expected to find that definitions are more helpful than context sentences for learning the meanings of polysemes than homonyms or unambiguous words. Conversely, we expect to find that context sentences are more useful than definitions for learning the meaning of homonyms and unambiguous words. Given the absence of any past literature regarding the use of generation with definitions vs. context sentences, we do not make any specific predictions about whether these variables may interact. Rather, we view this portion of the experiment as exploratory. Similarly, we make no specific predictions about whether the effects of definitions and context sentences are the same for individuals with higher and lower WM and inhibitory control, but will simply explore these questions.

4.3 Methods

4.3.1 Participants

Participants for this experiment were 31 native English speakers, 18 years and older, with no prior knowledge of German or Dutch. One participant was removed from all analyses due to failure to complete all of the tasks, and one other participant was removed due to failure to follow the directions during training, leaving data from 29 participants for analyses. All participants were right-handed, with normal or corrected-to-normal vision, and were recruited either from the Psychology Department Subject Pool and compensated with class credit and a $7.00 cash bonus if they completed both sessions. Participation time was approximately 2 hours total, divided equally between two sessions.

4.3.2 Design

This study used a 3 word type (unambiguous, homonym, polyseme) x 2 training condition (sentence reading, sentence generation) x 2 training materials (context sentences, definitions), 2 session (session 1, session 2) within-subjects design.

4.3.3 Stimuli

The stimuli used in Experiment 3 were identical to those used in Experiment 2 (see Appendix D).

4.3.4 Procedure

Experiment 3 consisted of two sessions of 1 hour each, spaced one week apart. In Session 1, participants completed a familiarity check to ensure that the novel words were not already known to participants. Next, participants received training on rare English words that were paired with their definitions or with context sentences. Words were randomly assigned to be trained with either a context sentence or a definition, and this condition was counterbalanced across training orders (described below). Following training, participants completed a free recall test in which they were instructed to type any words that they remembered from training. This task encouraged participants to use retrieval to enhance memory (the "testing effect", e.g., Pyc & Rawson, 2010; Roediger & Karpicke, 2006), and provided a measure of learning immediately after training. Next, participants completed a meaning production test which assessed their ability to produce a word meaning when presented with the word as a prompt.

During Session 2, which occurred after a one-week delay, participants again completed the meaning production task. Next, participants completed a forced choice sentence completion task, which assessed their knowledge of context-sensitive word knowledge. Participants also completed two individual difference measures: the Waters Reading Span test (Waters & Caplan, 1996) and the Simon task (Simon & Wolf, 1963). Finally, participants completed a Language History Questionnaire to collect relevant language background information. Table 31 provides a summary of the experiment procedures and timeline.

Table 17. Experiment 3 timeline

Session	Day	Tasks
Session 1	Day 1	Training 1: Familiarity check Word + definition or context sentence (2x per word) Training 2: Read or generate sentence (1x per word) Testing 1: Free recall Testing 2: Meaning production
Session 2	Day 8	Testing 3: Meaning production Testing 4: Forced choice sentence completion Testing 5: Individual difference measures Testing 6: Language History Questionnaire (LHQ)

4.3.4.1 Familiarity check

To ensure that novel words were not already known to participants, the initial training session began with a familiarity check (e.g., Balass, 2011). The procedure was identical to Experiment 2.

4.3.4.2 Vocabulary training

After completing the familiarity check, participants moved on to training, which followed largely the same procedures as Experiment 2. However, this experiment introduced a new training manipulation: in addition to random assignment to the sentence reading or sentence generation condition, words were randomly assigned to be trained with definitions or with context sentences. This meant that there were four training orders to counterbalance these conditions, and participants were randomly assigned to one of these four orders. Training took approximately 50 minutes to complete.

4.3.5 Free recall

After completing training, learners were asked to complete a free recall task in which they were instructed to type every target word they could remember, in any order. The procedures and scoring for this task were identical to those used in Experiment 2.

4.3.6 Meaning production

After completing the free recall task, learners were asked to complete a meaning production task in which they were presented with a target word they had encountered during training and were asked to type in the definition. The procedures for this task were identical to those used in Experiment 2. The scoring procedures were identical to Experiment 2, except that in Experiment 2 answers were considered correct if they accurately captured the meaning of the word and included at least one content word from the training definition, but in Experiment 3 they were considered correct if they accurately captured the meaning of the word and included at least one content word from the training definitions or sentences.

4.3.7 Forced choice sentence completion

After completing the meaning production task in Session 2, participants were asked to complete a forced choice sentence completion task. In this task participants viewed sentences with one word missing, and were instructed to read the sentence and pick the word that fits best in the sentence from a list of five options. Participants were told that they would only have a short amount of time to read the sentence and respond before the program moved to the next trial.

Trials began with a fixation cross at the center of the screen for 1000 ms, followed by a blank screen for 100 ms, and then the sentence and word choices appeared on the screen for 18000 ms or until a response was made. Participants were instructed to press one of five buttons on a response box to make a response. Sentences appeared in a random order, and one sentence per target word was presented.

The sentences in this task had not previously been viewed by participants, but were written and normed following the same procedures as the sentences used for training. Foils consisted of other words that were trained in the experiment, selected randomly from the entire list. The position of the correct word within the list of five options was randomized. E-Prime was used to present the trials and collect information about reaction time and accuracy.

4.3.8 Individual difference tasks

After completing the sentence completion task, participants completed the individual difference measures, which followed the same procedures as in Experiments 1 and 2. We examined accuracy on the Waters sensibility judgements to ensure participants were paying attention to the task. All participants in this experiment correctly responded to at least 70% of all sensibility judgements, and so all data were used.

4.4 Results

4.4.1 Statistical approach

As in Experiments 2 and 3, the main analyses described in this section are linear mixed-effects models, which allowed us examine subject and item effects simultaneously (e.g., Baayen et al., 2008). The statistical approach is identical to procedures described above in Experiment 2.

4.4.2 Familiarity check

To establish whether the stimuli we selected were sufficiently rare as to be unfamiliar to participants, the initial training session included a familiarity check. Following procedures from Balass (2011), words that were consistently rated as familiar by participants were removed from further analyses. We calculated the percent of participants that indicated familiarity with each word, and examined the distribution of this variable. The same three words that were removed from Experiment 2 based on the familiarity check results also emerged as clearly more familiar than the rest of the set in Experiment 3, with roughly the same percentage of participants indicating familiarity: evanescence (61% familiar), levity (47% familiar), and discinct (46% familiar). As before, these three words were removed from further analyses.

4.4.3 Vocabulary training analyses

Before conducting analyses on any of the outcome measures, we first examined participant accuracy during training following the same procedures outlined above in Experiment 1. For

Experiment 3, one participant was removed from all further analyses because of errors on read trials (sentence generation was used instead of sentence reading for approximately 25% of trials), leaving data from 29 participants remaining for analyses. Mean accuracy on read trials for the remaining participants in Experiment 3 was 99%. All participants met criteria for generate trials.

4.4.4 Free recall results

This section describes the results of analyses of the free recall response data. Mean proportion correct for all words was .20 (*SD* = 0.12). The mean accuracy for words trained with sentence reading was .15 *(SD* = .33), and the mean accuracy for words trained with sentence generation was .25 (*SD* = .40). Mean accuracy was .15 (*SD* = .33) for unambiguous words, .30 for homonyms (*SD* = .44), and .23 for polysemes (*SD* = .39).

We examined Pearson correlations between free recall accuracy and individual difference measures. We found that of the three individual differences measures, only Waters set size span was significantly correlated with free recall accuracy (*r* = -.11). Therefore, we decided to use Waters set size span instead of Waters total span in the final models. Because we had specific hypotheses about both Waters and Simon scores, we wanted to include both in the final models. Following procedures above, we assessed the risk of multicollinearity between these two predictors. The condition number indicated that there was standard collinearity ($\kappa = 8.79$), which falls well below the threshold for potentially harmful collinearity ($\kappa = 30$; Baayen, 2008). We observed VIF values of 1.01 for both Waters and Simon, which is well below the point at which VIF becomes problematic (VIF = 5.00). Therefore, we proceeded to build a final model that included both Waters and Simon.

We constructed a cumulative link mixed model (clmm) to examine free recall accuracy. The fixed effects of theoretical interest were training condition (read or generate), training material (sentence or definition), word type (unambiguous, homonym, polyseme), Simon score, Waters set size span, and the five-way interaction of word type, training condition, training material, Waters set size span, and Simon score. Word length was included as a control variable. The baseline conditions were coded as follows: unambiguous words were the baseline to which homonyms and polysemes were compared, sentences were the baseline to which definitions were compared, and sentence reading was the baseline to which sentence generation was compared. We specified the maximal random effects structure for which this model would converge, which meant including random intercepts for both subjects and items, and random by-subject slopes for the interaction of word type and training condition.

The model equation and fixed effects estimates for Model 7 are presented in Table 18 and the random effects in Table 19. Of the fixed effects of theoretical interest, there were significant main effects of homonyms, $b = 1.20$, $SE = 0.61$, $z = 1.98$, $p < .05$, and polysemes, $b = 1.36$, $SE = 0.53$, $z = 2.58$, $p = .01$, but these were qualified by higher-order interactions. There was also a significant two-way interaction of word type and Waters set size span, $b = -0.97$, $SE = 0.39$, $z = -2.50$, $p = .01$, but this was qualified by a higher order three-way interaction of word type, Waters set size span, and training condition, $b = 1.44$, $SE = 0.64$, $z = 2.26$, $p = .02$. We probed this interaction by using the regression equation from Model 7 to generate estimated proportion correct at all levels of word type, training condition, and at one standard deviation above and below the mean Waters set size (see Figure 11). This revealed differences in generation effects across word types and across participants with lower vs. higher WM spans. Specifically, for lower WM span

participants there was a generation effect for unambiguous words and homonyms, but for higher WM span participants there was a generation effect for homonyms only.

Table 18. Fixed effects estimates for Model 7, free recall

			95% CI				
	Est	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	1.20	0.61	0.01	2.38	1.98	.05	*
Polysemes	1.36	0.53	0.33	2.40	2.58	.01	**
Definition	-0.12	0.62	-1.34	1.11	-0.19	.85	
Waters size	0.31	0.35	-0.38	1.01	0.89	.38	
Generate	0.44	0.55	-0.65	1.52	0.79	.43	
Simon score	-0.08	0.17	-0.41	0.25	-0.48	.63	
Word length	-0.10	0.09	-0.27	0.07	-1.15	.25	
Homonyms*Definitions	-0.35	0.82	-1.95	1.26	-0.42	.67	
Polysemes*Definitions	-0.87	0.73	-2.31	0.57	-1.18	.24	
Homonyms*Waters	-0.51	0.46	-1.42	0.39	-1.12	.26	
Polysemes*Waters	-0.97	0.39	-1.74	-0.21	-2.50	.01	*
Definition*Waters	-0.41	0.54	-1.46	0.64	-0.76	.45	
Homonyms*Generate	0.74	0.72	-0.68	2.16	1.02	.31	
Polysemes*Generate	-0.25	0.72	-1.66	1.17	-0.34	.73	
Definition*Generate	0.95	0.81	-0.64	2.54	1.17	.24	
Waters*Generate	-0.63	0.50	-1.60	0.35	-1.27	.21	
Homonyms*Simon	-0.28	0.23	-0.73	0.18	-1.19	.23	
Polysemes*Simon	0.13	0.19	-0.24	0.49	0.68	.49	
Definition*Simon	-0.26	0.34	-0.93	0.41	-0.75	.45	
Waters*Simon	-0.02	0.12	-0.26	0.21	-0.18	.85	
Generate*Simon	0.09	0.25	-0.39	0.58	0.37	.71	
Homonyms*Definition*Waters	0.25	0.71	-1.14	1.65	0.36	.72	
Polysemes*Definition*Waters	0.69	0.67	-0.63	2.02	1.03	.30	
Homonyms*Definition*Generate	-0.41	0.99	-2.35	1.53	-0.42	.68	
Polysemes*Definition*Generate	-0.17	1.00	-2.13	1.80	-0.17	.87	
Homonyms*Waters*Generate	0.99	0.65	-0.28	2.26	1.53	.13	
Polysemes*Waters*Generate	1.44	0.64	0.19	2.68	2.26	.02	*
Definition*Waters*Generate	0.31	0.71	-1.08	1.70	0.44	.66	
Homonyms*Definition*Simon	0.19	0.47	-0.74	1.12	0.41	.68	
Polysemes*Definition*Simon	0.02	0.40	-0.76	0.80	0.05	.96	
Homonyms*Waters*Simon	0.06	0.17	-0.28	0.40	0.36	.72	
Polysemes*Waters*Simon	-0.03	0.13	-0.30	0.23	-0.25	.80	
Definition*Waters*Simon	0.02	0.22	-0.41	0.45	0.10	.92	
Homonyms*Generate*Simon	0.23	0.35	-0.46	0.92	0.66	.51	
Polysemes*Generate*Simon	-0.17	0.34	-0.85	0.50	-0.50	.62	
Definition*Generate*Simon	0.17	0.42	-0.66	1.00	0.40	.69	
Waters*Generate*Simon	-0.11	0.20	-0.50	0.29	-0.53	.60	
Homonyms*Definition*Waters*Generate	-0.58	0.86	-2.28	1.11	-0.68	.50	
Polysemes*Definition*Waters*Generate	-1.47	0.91	-3.25	0.31	-1.62	.11	

Homonyms*Definition*Waters*Simon	0.18	0.29	-0.39	0.76	0.62	.53
Polysemes*Definition*Waters*Simon	-0.04	0.31	-0.65	0.57	-0.14	.89
Homonym*Definition*Generate*Simon	-0.06	0.54	-1.13	1.01	-0.11	.91
Polyseme*Definition*Generate*Simon	0.06	0.52	-0.96	1.08	0.12	.90
Homonym*Waters*Generate*Simon	-0.02	0.26	-0.53	0.48	-0.09	.93
Polyseme*Waters*Generate*Simon	-0.01	0.25	-0.50	0.47	-0.04	.97
Definition*Waters*Generate*Simon	0.10	0.29	-0.48	0.67	0.33	.74
Homonym*Definition*Waters*Generate*Simon	-0.18	0.35	-0.87	0.51	-0.51	.61
Polyseme*Definition*Waters*Generate*Simon	0.08	0.39	-0.69	0.85	0.21	.84

Model equation. Model 7 <- clmm(ACC ~ 1 + WordType*SentorDef *ReadorGen*waterssetsize*SimonScore + EnglishLen + SimonScore + (1|Subject) + (1|Item) + (0+WordType*ReadorGen|Subject), data=freerecall, link = 'logit', threshold = 'flexible')
†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 19. Random effects for Model 7, free recall

	Variance	*SD*
Item intercept	0.43	0.65
Unambiguous|Subject	1.16	1.08
Polysemes|Subject	0.23	0.48
Homonyms|Subject	0.28	0.53
Generate|Subject	0.67	0.82
Polysemes*Generate|Subject	1.38	1.17
Homonyms*Generate|Subject	1.29	1.13
Subject intercept	<0.001	<0.001

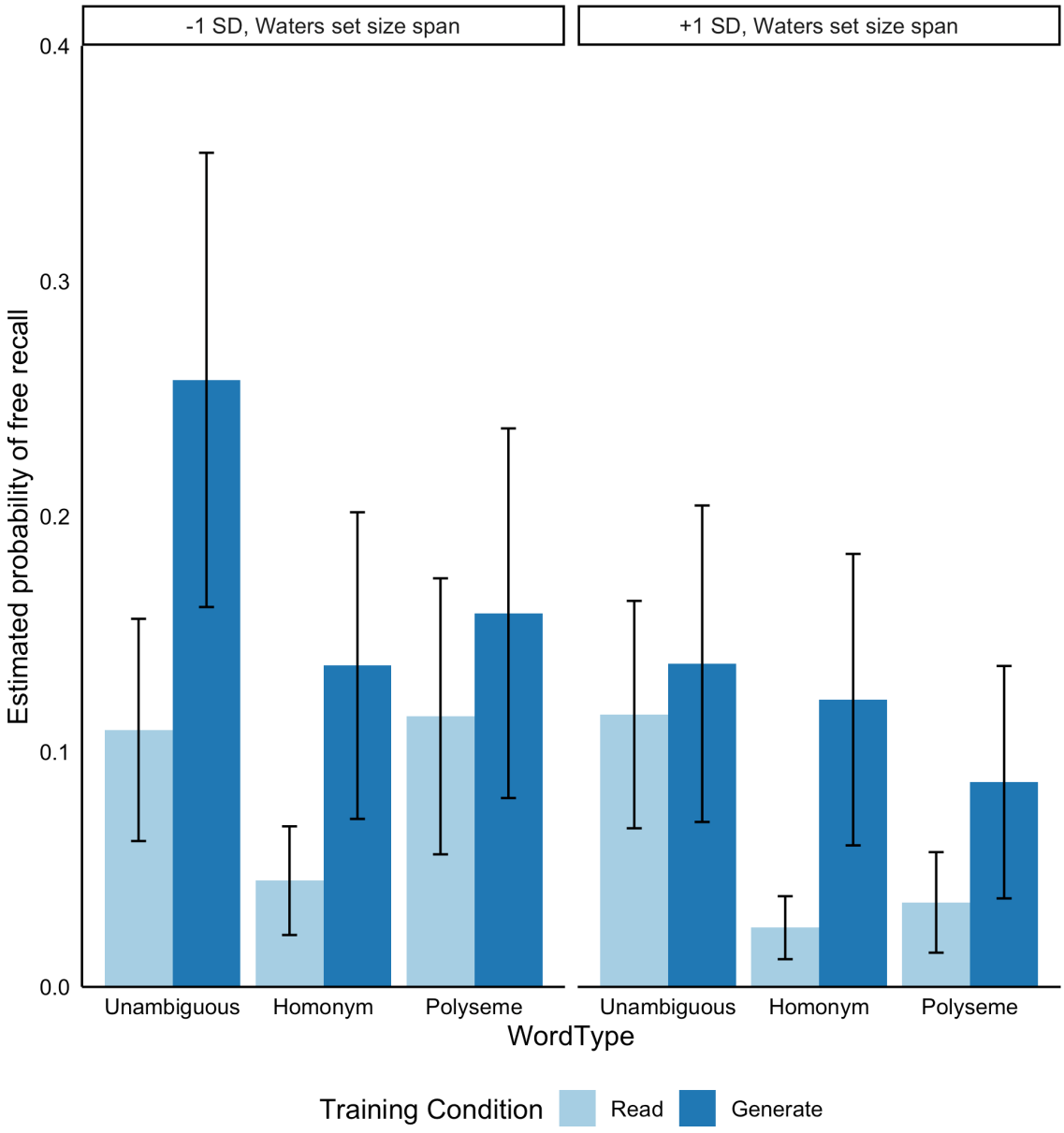

Figure 11. Estimated probability of free recall by word type, training condition, and Waters set size span. Error bars represent standard error of the mean.

4.4.5 Meaning production results

This section examined accuracy data from the meaning production. We began by examining descriptive statistics. Overall, the mean proportion correct in Session 1 was .43 (SD = 0.16), and the mean proportion correct was .27 (SD = 0.12) in Session 2. The mean accuracy for words trained with sentence reading was .31 (*SD* = .41), and the mean accuracy for words trained with sentence generation was .39 (*SD* = .46). Mean accuracy was .35 (*SD* = .44) for unambiguous words, .35 for homonyms (*SD* = .45), and .35 for polysemes (*SD* = .43).

Next, we examined Pearson correlations between proportion correct on the meaning production task and the individual difference measures. We found that both the Waters total span and Simon scores were significantly correlated with meaning production accuracy (Waters: r = .08, Simon: r = -.07), whereas Waters set size span was not (r = .02). On this basis we decided to use Waters total span instead of Waters set size span in the final models. As before, we examined the condition number and VIF of these two individual difference measures to assess the risk of multicollinearity. The condition number indicated that there was no more than standard collinearity (κ = 9.77), and the VIF levels for both variables were 1.11, well below the point at which VIF becomes problematic. Therefore, we built a final model that included both Waters and Simon.

We constructed a clmm model to examine meaning production accuracy. The fixed effects of theoretical interest were training condition (read or generate), training material (sentence or definition), word type (homonym, polyseme, unambiguous), session (day 1 or day 2) and the interaction of word type, training condition, training material, Simon score, and Waters total span. Fixed effects were also entered for the following control variables: training order (trained first vs. second), training sentence rating, training definition length, and word length in letters.

We specified the maximal random effects structure for which the model would converge. This consisted of subject and item random intercepts, as well as random slopes for the interaction of word type and training condition. There were no missing data for these analyses.

The model equation and fixed effects estimates for Model 8 are presented in Table 20, and the random-effects estimates are presented in Table 21. Of the fixed effects of theoretical interest there was a significant effect of Session, such that accuracy was higher on Session 1 than Session 2, $b = -1.05$, $SE = 0.09$, $z = -12.09$, $p < .001$. There was a significant interaction of training condition and training material, $b = 1.62$, $SE = 0.73$, $z = 2.21$, $p = .03$. We probed this interaction using the *effects* package (Figure 12), which revealed that words trained with generation and definitions performed significantly better than any other training procedures. There was also a significant interaction of Waters and Simon, $b = -0.22$, $SE = 0.08$, $z = -2.91$, $p < .001$. We probed this interaction (Figure 13), which revealed that performance was highest for individuals who were higher in WM and who were also higher in inhibitory control. Finally, there was also a marginal three-way interaction of word type, Simon, and Waters total span, $b = 0.16$, $SE = 0.10$, $z = 1.69$, $p = .09$. We probed this interaction (Figure 14), which revealed that for all three word types, performance was best when learners were both higher in WM and higher in inhibitory control. However, this interaction appeared to be driven by learners who were low in working memory but high in inhibitory control: interestingly, these learners performed unexpectedly well on homonym meaning performance. This suggests that when learners both cannot remember homonym meanings well but also have the ability to suppress the inappropriate meanings when they do remember them, they are able to do well on homonym meaning production. The four- and five-way interactions did not reach significance.

Table 20. Fixed-effects estimates for Model 8, meaning production

			95% CI				
	Est	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	0.29	0.57	-0.82	0.29	0.51	.61	
Polysemes	0.22	0.48	-0.73	1.33	0.45	.65	
Waters	0.16	0.24	-0.30	1.11	0.69	.49	
Generate	-0.24	0.44	-1.10	0.22	-0.54	.59	
Definition	0.12	0.35	-0.56	0.99	0.35	.72	
Simon	-0.10	0.09	-0.27	0.59	-1.10	.27	
Session	-1.05	0.09	-1.22	-0.87	-12.09	<.001	***
Meaning	-0.30	0.11	-0.52	-0.13	-2.63	.01	**
Word length	-0.26	0.12	-0.50	-0.04	-2.11	.04	*
Definition length	0.10	0.02	0.05	0.34	4.23	<.001	***
Study time	-0.02	0.03	-0.07	0.03	-0.61	.54	
Homonyms*Waters	-0.13	0.31	-0.74	-0.08	-0.42	.67	
Polysemes*Waters	-0.43	0.23	-0.88	0.18	-1.87	.06	†
Homonyms*Generate	0.02	0.43	-0.82	0.47	0.04	.97	
Polysemes*Generate	-0.35	0.38	-1.11	0.49	-0.92	.36	
Waters*Generate	0.47	0.45	-0.42	1.22	1.04	.30	
Homonyms*Definition	0.04	0.45	-0.83	0.93	0.10	.92	
Polysemes*Definition	-0.06	0.33	-0.70	0.82	-0.17	.86	
Waters*Definition	0.03	0.38	-0.71	0.68	0.08	.94	
Generate*Definition	1.62	0.73	0.18	2.36	2.21	.03	*
Homonyms*Simon	0.02	0.11	-0.21	1.45	0.14	.89	
Polysemes*Simon	-0.04	0.08	-0.20	0.18	-0.46	.65	
Waters*Simon	-0.22	0.08	-0.37	-0.06	-2.91	<.001	**
Generate*Simon	0.10	0.19	-0.28	0.25	0.54	.59	
Definition*Simon	0.18	0.15	-0.13	0.55	1.14	.26	
Homonyms*Waters*Generate	-0.22	0.45	-1.10	0.08	-0.49	.63	
Polysemes*Waters*Generate	0.24	0.41	-0.56	1.12	0.59	.56	
Homonyms*Waters*Definition	-0.09	0.50	-1.07	0.71	-0.19	.85	
Polysemes*Waters*Definition	0.23	0.38	-0.50	1.21	0.62	.54	
Homonyms*Generate*Definition	-0.40	0.60	-1.58	0.34	-0.66	.51	
Polysemes*Generate*Definition	-0.09	0.49	-1.04	1.09	-0.18	.86	
Waters*Generate*Definition	-0.36	0.78	-1.88	0.59	-0.46	.65	
Homonyms*Waters*Simon	0.16	0.10	-0.03	1.69	1.69	.09	†
Polysemes*Waters*Simon	0.12	0.08	-0.03	0.31	1.53	.13	
Homonyms*Generate*Simon	-0.16	0.19	-0.53	-0.01	-0.84	.40	
Polysemes*Generate*Simon	0.02	0.17	-0.31	0.39	0.12	.91	
Waters*Generate*Simon	0.14	0.16	-0.18	0.47	0.85	.39	
Homonyms*Definition*Simon	-0.22	0.20	-0.61	0.10	-1.12	.26	

Polysemes*Definition*Simon	-0.05	0.15	-0.34	0.33	-0.35	.72
Waters*Definition*Simon	0.21	0.14	-0.06	0.50	1.50	.13
Generate*Definition*Simon	-0.27	0.32	-0.90	0.00	-0.86	.39
Homonyms*Waters*Generate*Definition	0.40	0.64	-0.86	1.03	0.63	.53
Polysemes*Waters*Generate*Definition	-0.10	0.52	-1.12	1.16	-0.19	.85
Homonyms*Waters*Generate*Simon	-0.16	0.15	-0.46	0.86	-1.04	.30
Polysemes*Waters*Generate*Simon	-0.04	0.14	-0.31	0.26	-0.26	.79
Homonyms*Waters*Definition*Simon	-0.03	0.18	-0.38	0.24	-0.15	.88
Polysemes*Waters*Definition*Simon	-0.15	0.14	-0.42	0.20	-1.09	.27
Homonyms*Generate*Definition*Simon	0.37	0.26	-0.13	0.64	1.44	.15
Polysemes*Generate*Definition*Simon	0.06	0.21	-0.34	0.57	0.29	.77
Waters*Generate*Definition*Simon	-0.23	0.27	-0.76	0.18	-0.82	.41
Homonyms*Waters*Generate*Definition*Simon	0.07	0.22	-0.37	0.61	0.31	.76
Polysemes*Waters*Generate*Definition*Simon	0.06	0.18	-0.28	0.50	0.36	.72

Model equation. Model 8 <- clmm(ACC ~ 1 + WordType*SentorDef *ReadorGen*waterssetsize*SimonScore + Session + transnum + Word length + DefLen + studytime + (1|Subject)+(1|EnglishWord) + (0+WordType*ReadorGen|Subject), data=freerecall, link = 'logit', threshold = 'flexible')

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 21. Random effects estimates for Model 8, meaning production

	Variance	*SD*
Item intercept	1.19	1.09
Unambiguous|Subject	0.49	0.70
Polysemes|Subject	0.83	0.91
Homonyms|Subject	0.74	0.86
Generate|Subject	0.40	0.63
Polysemes*Generate|Subject	0.73	0.86
Homonyms*Generate|Subject	0.72	0.85
Subject intercept	<0.001	<0.001

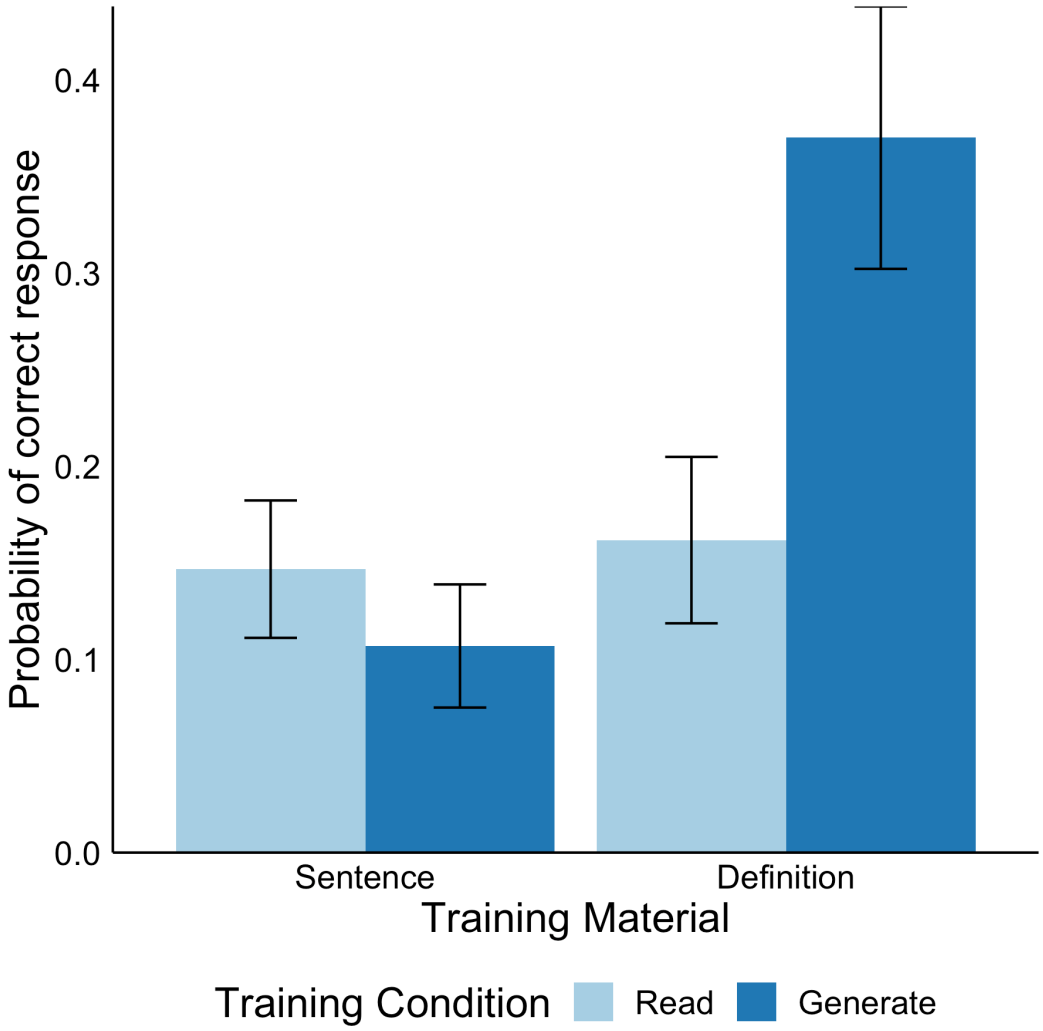

Figure 12. Estimated probability of a correct response by training condition and training material. Error bars represent standard error of the mean.

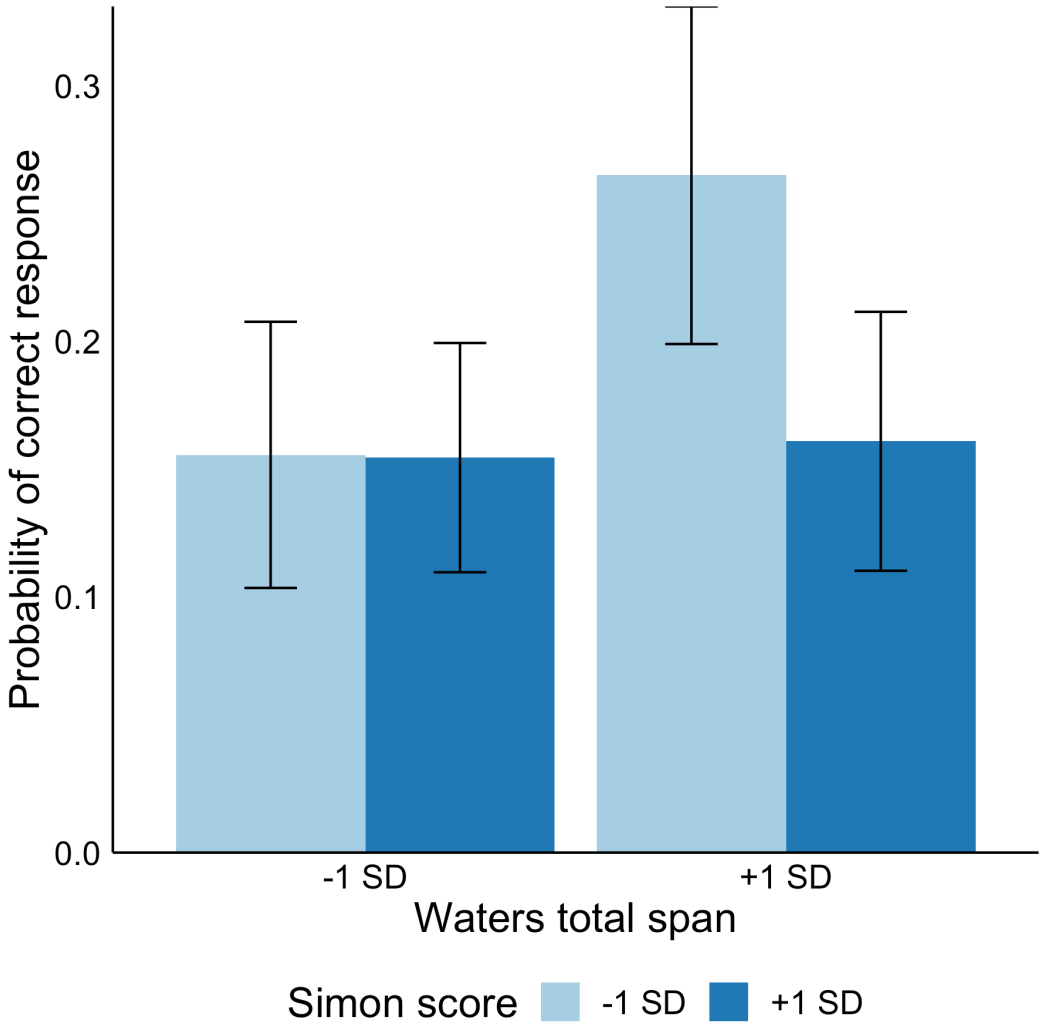

Figure 13. Estimated probability of correct response by Waters and Simon scores. Error bars represent standard error of the mean.

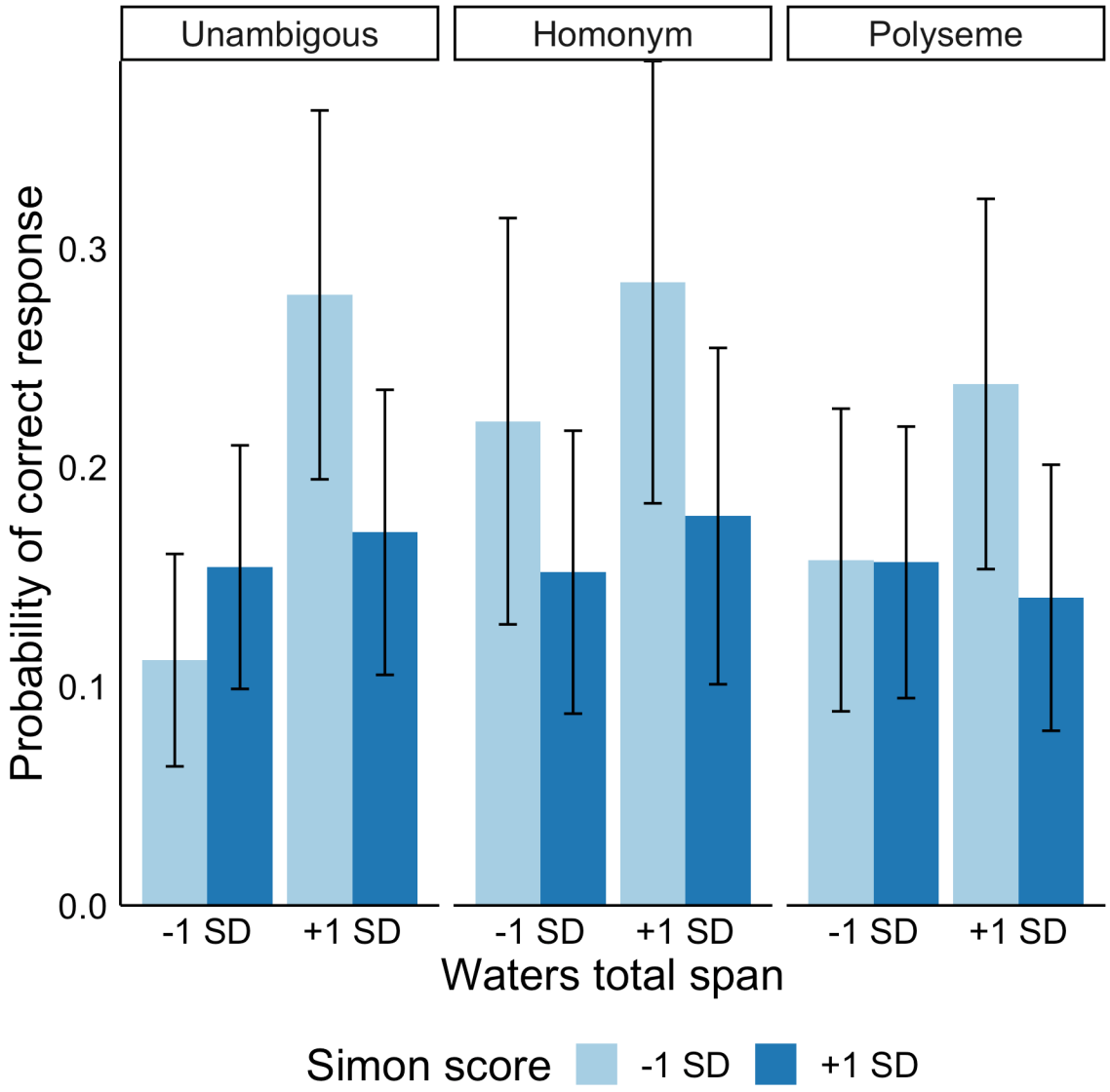

Figure 14. Probability of correct response by word type, Waters total span, and Simon score

4.4.6 Forced choice sentence completion results

We examined accuracy data from the forced choice sentence completion task using a clmm model. We began by examining descriptive statistics. The mean accuracy for words trained with sentence reading was .64 (*SD* = .48), and the mean accuracy for words trained with sentence generation was .66 (*SD* = .48). The mean accuracy for words trained with context sentences was

.65 *(SD* = .48), and the mean accuracy for words trained with definitions was .65 (*SD* = .48). Mean accuracy was .65 (*SD* = .48) for unambiguous words, .63 for homonyms (*SD* = .48), and .65 for polysemes (*SD* = .48).

We examined Pearson correlations between proportion correct on the sentence completion task and the individual difference measures. We found that Waters total span was significantly correlated with sentence completion accuracy (*r* = .08), whereas Waters set size span (*r* = .05) and Simon score were not (*r* = -.05). The condition number for the final model that included both individual difference measures indicated a moderate, but not problematic amount of collinearity in the model ($\kappa = 30.09$), and the VIF scores for both Waters and Simon were 1.05, which is under the 5.0 threshold for problematic collinearity. Therefore, we included both Waters total span and Simon scores in our final model.

We built a glmer model to examine accuracy on the sentence completion task. The fixed effects of theoretical interest were training condition (read or generate), training material (sentence or definition), word type (homonym, polyseme, unambiguous), Session (1 or 2) and the interaction of word type by training condition by waters total span, as well as the interaction of training condition, training material, Simon, and Waters. Fixed effects were also entered for the following control variables: training order (trained first vs. second), training sentence rating, training definition length, and word length in letters.

We specified the maximal random effects structure justified by our model for which the model would converge. This consisted of subject and item random intercepts, but no random slopes because these caused convergence problems for the model. There were no missing data for these analyses.

The model equation and fixed effects estimates for Model 9 are presented in Table 22, and the random-effects estimates are presented in Table 23. Of the fixed effects of theoretical interest there was a significant main effect of training material, but this was qualified by a higher-order interaction. Specifically, there was a significant interaction of training condition and training material, $b = 1.31$, $SE = 0.60$, $z = 2.18$, $p = .03$. We probed this interaction with the *effects* package, which revealed that words trained with definitions and generation were responded to correctly more often than any other type (Figure 15). Finally, there were three marginally significant interactions: a two-way interaction of word type and training condition, $b = 0.86$, $SE = 0.44$, $z = 1.95$, $p = .05$, a three-way interaction of word type, training condition, and training material, $b = -1.11$, $SE = 0.62$, $z = -1.79$, $p = .07$, and a three-way interaction of word type, training material, and Waters total span, $b = -0.81$, $SE = 0.47$, $z = -1.73$, $p = .08$. Although these interactions did not reach significance, we nevertheless highlight them because they are consistent with results found in the free recall and meaning production data, and so may be indicative of trends that bear further examination in future studies. We probed the marginal interaction of word type, training condition, and training material (Figure 16), which revealed that the interaction was driven by differences in generation effects for sentence vs. definition training for unambiguous words and polysemes. Specifically, there was a generation effect for unambiguous words and polysemes, but only when these words were trained with definitions. When these words were trained with sentences, there was a generation disadvantage.

We also probed the marginal interaction of word type, training material, and Waters total span (Figure 17), which revealed that this interaction was driven by differences in how learners with lower vs. higher WM learn polysemes. Specifically, when polysemes were trained with definitions, learners with higher WM performed better than learners with lower WM.

Table 22. Fixed effects estimates for Model 9, sentence completion

			95% CI				
	Est	SE	Lower bound	Upper bound	*z*	*p*	Sig?
Homonyms	-0.01	0.53	-1.05	1.04	-0.01	.99	
Polysemes	-0.19	0.50	-1.18	0.80	-0.38	.71	
Generate	-0.51	0.38	-1.26	0.23	-1.36	.17	
Definition	-0.95	0.37	-1.69	-0.22	-2.56	.01	*
Waters	0.01	0.27	-0.52	0.54	0.04	.97	
Simon	-0.08	0.10	-0.28	0.11	-0.84	.40	
Study time	-0.01	0.03	-0.08	0.05	-0.39	.70	
Meaning	-0.01	0.15	-0.30	0.29	-0.05	.96	
Word length	-0.28	0.11	-0.48	-0.07	-2.60	.01	**
Definition length	0.16	0.03	0.10	0.22	4.91	.00	***
Homonyms*Generate	0.53	0.45	-0.36	1.41	1.17	.24	
Polysemes*Generate	-0.08	0.43	-0.92	0.77	-0.18	.86	
Homonyms*Definition	0.86	0.44	-0.01	1.73	1.95	.05	†
Polysemes*Definition	0.41	0.42	-0.42	1.24	0.97	.33	
Generate*Definition	1.31	0.60	0.13	2.49	2.18	.03	*
Homonyms*Waters	0.06	0.33	-0.58	0.70	0.19	.85	
Polysemes*Waters	0.32	0.31	-0.29	0.94	1.04	.30	
Generate*Waters	0.33	0.40	-0.46	1.11	0.82	.42	
Definition*Waters	0.48	0.41	-0.31	1.28	1.19	.23	
Homonyms*Simon	0.04	0.12	-0.19	0.28	0.37	.71	
Polysemes*Simon	0.14	0.12	-0.09	0.37	1.22	.22	
Generate*Simon	-0.05	0.17	-0.38	0.27	-0.32	.75	
Definition*Simon	0.10	0.16	-0.22	0.41	0.59	.56	
Waters*Simon	-0.03	0.09	-0.19	0.14	-0.30	.76	
Homonyms*Generate*Definition	-1.11	0.62	-2.32	0.11	-1.79	.07	†
Polysemes*Generate*Definition	-0.09	0.59	-1.25	1.06	-0.15	.88	
Homonyms*Generate*Waters	-0.11	0.48	-1.05	0.83	-0.23	.82	
Polysemes*Generate*Waters	-0.54	0.46	-1.43	0.36	-1.17	.24	

Homonyms*Definition*Waters	-0.45	0.49	-1.41	0.51	-0.92	.36	
Polysemes*Definition*Waters	-0.81	0.47	-1.73	0.11	-1.73	.08	†
Generate*Definition*Waters	-0.56	0.64	-1.81	0.69	-0.88	.38	
Homonyms*Generate*Simon	0.01	0.20	-0.39	0.40	0.03	.98	
Polysemes*Generate*Simon	-0.14	0.19	-0.52	0.24	-0.74	.46	
Homonyms*Definition*Simon	-0.14	0.20	-0.53	0.25	-0.70	.48	
Polysemes*Definition*Simon	-0.29	0.19	-0.66	0.08	-1.52	.13	
Generate*Definition*Simon	-0.09	0.26	-0.61	0.43	-0.34	.73	
Homonyms*Waters*Simon	-0.04	0.11	-0.25	0.17	-0.39	.70	
Polysemes*Waters*Simon	0.04	0.11	-0.17	0.25	0.36	.72	
Generate*Waters*Simon	0.09	0.14	-0.18	0.35	0.63	.53	
Definition*Waters*Simon	-0.02	0.15	-0.31	0.27	-0.15	.88	
Homonyms*Generate*Definition*Waters	0.32	0.66	-0.97	1.61	0.49	.63	
Polysemes*Generate*Definition*Waters	0.67	0.63	-0.56	1.91	1.07	.29	
Homonyms*Generate*Definition*Simon	0.24	0.27	-0.30	0.78	0.87	.38	
Polysemes*Generate*Definition*Simon	0.26	0.26	-0.25	0.77	1.01	.31	
Homonyms*Generate*Waters*Simon	-0.05	0.16	-0.37	0.26	-0.33	.74	
Polysemes*Generate*Waters*Simon	-0.07	0.16	-0.38	0.23	-0.48	.63	
Homonyms*Definition*Waters*Simon	0.28	0.19	-0.09	0.64	1.50	.13	
Polysemes*Definition*Waters*Simon	0.13	0.18	-0.23	0.49	0.71	.48	
Generate*Definition*Waters*Simon	-0.24	0.23	-0.69	0.22	-1.03	.30	
Homonyms*Generate*Definition*Waters*Simon	-0.01	0.24	-0.48	0.47	-0.02	.98	
Polysemes*Generate*Definition*Waters*Simon	-0.10	0.23	-0.55	0.35	-0.44	.66	

Model equation. Model 9 <- glmer(ACC ~ 1 + WordType*SentorDef *ReadorGen*waterssetsize*SimonScore + Session + transnum + Word length + DefLen + studytime + (1|Subject)+(1|EnglishWord) + (0+WordType*ReadorGen|Subject), data=FCR, family = binomial, glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 20000)))

†$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$

Table 23. Random-effects estimates for Model 9, sentence completion

	Variance	*SD*
Item intercept	0.80	0.89
Subject intercept	0.28	0.53

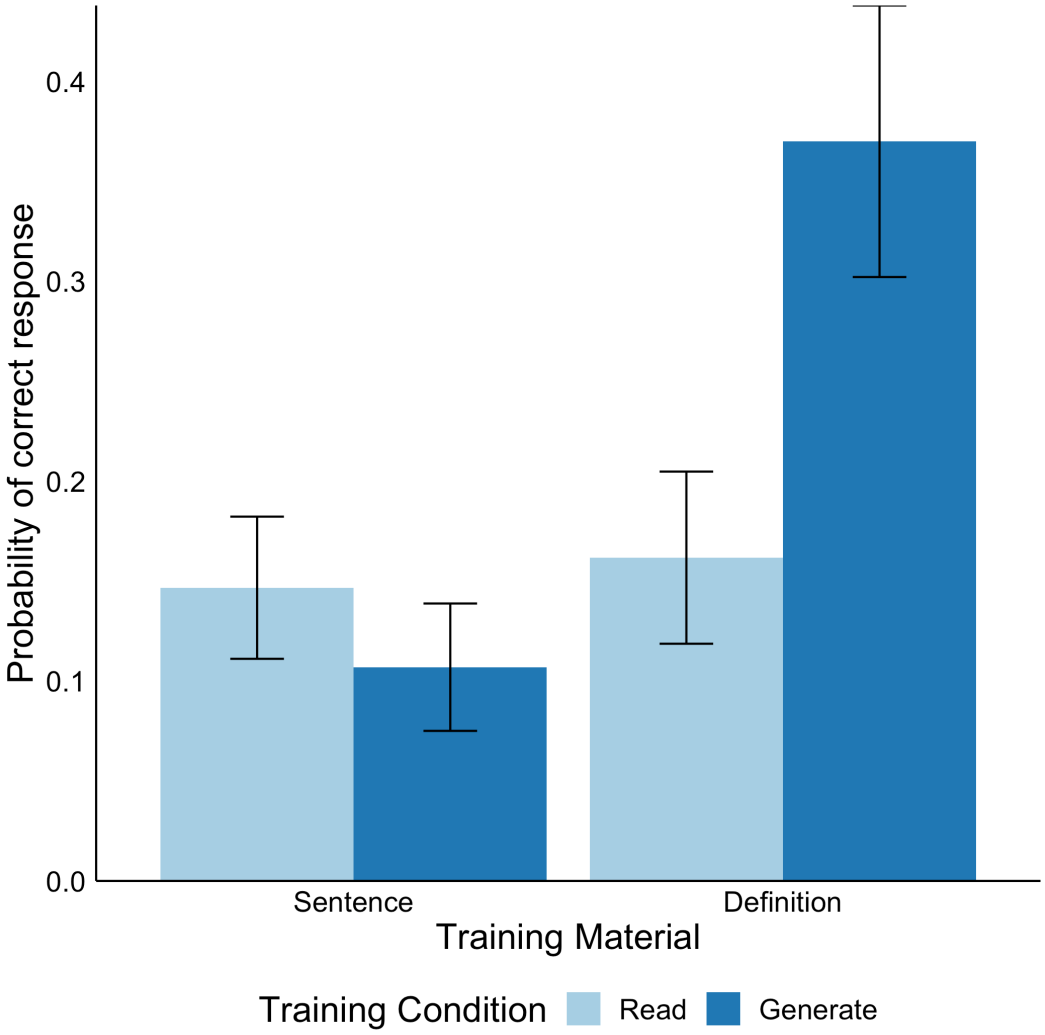

Figure 15. Estimated probability of correct sentence completion by training condition and training material. Error bars represent standard error of the mean.

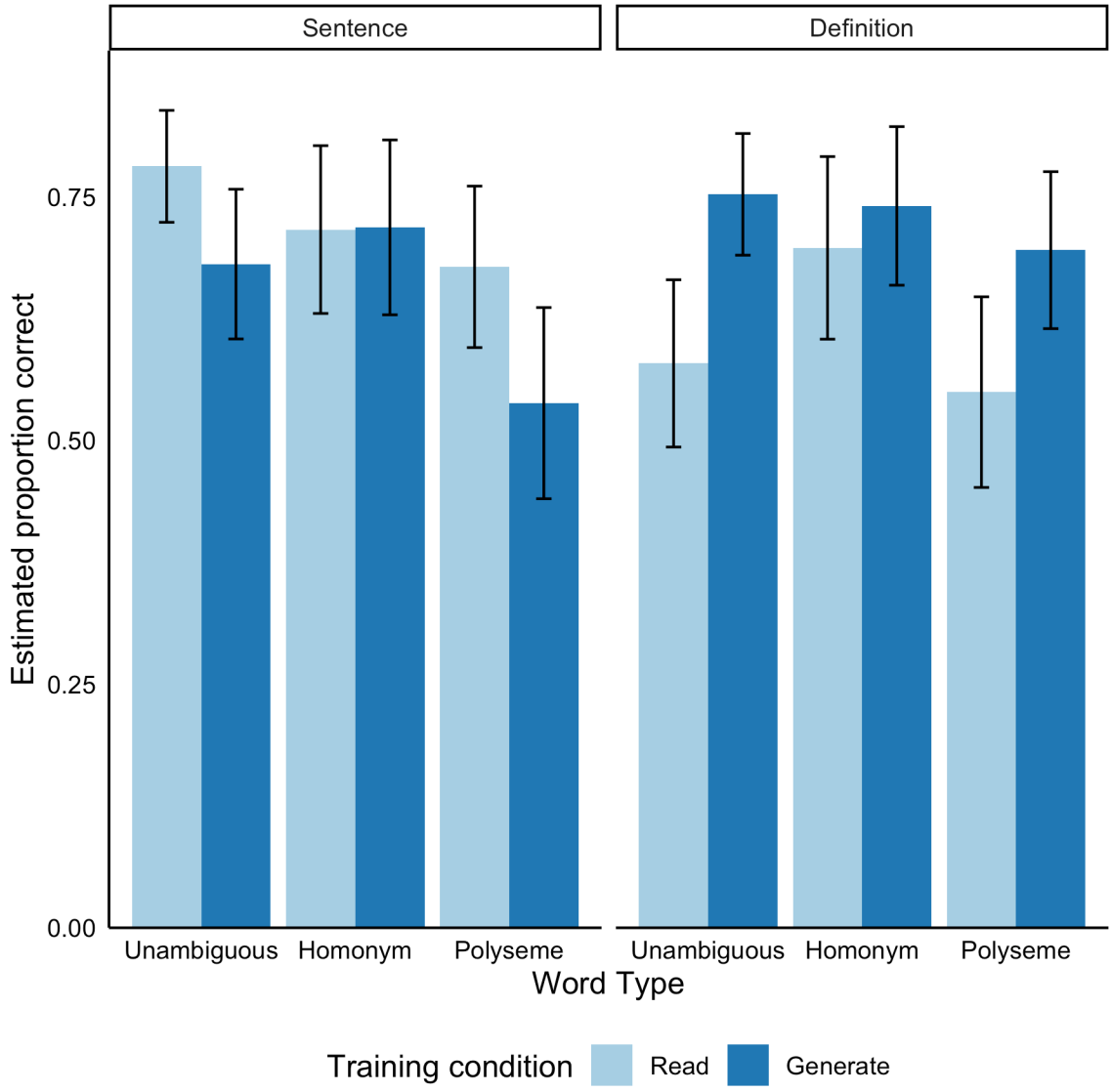

Figure 16. Estimated proportion correct by training condition, word type, and training material

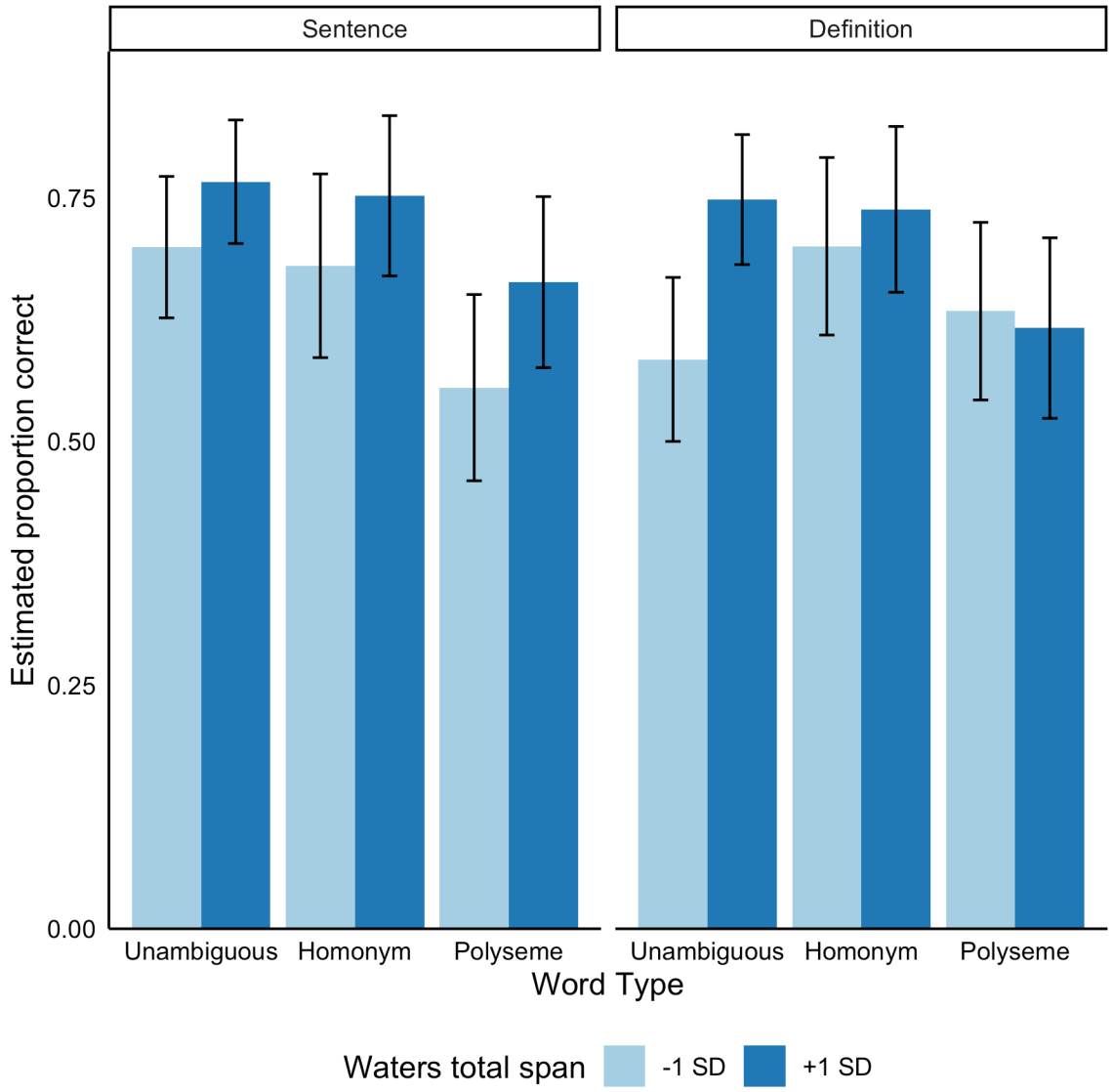

Figure 17. Estimated proportion correct by word type, Waters total span, and training material

4.5 Discussion

We now turn to a discussion of the results of Experiment 3. We begin by reviewing the hypotheses of this study, and then we turn to highlighting the significant findings and discussing how they support or fail to support our hypotheses, as well as how they agree with or diverge from

past findings. Next, we discuss possible theoretical mechanisms that may explain our findings, as well as how our findings impact existing models of generation effects and ambiguity effects. In the final section, we discuss future directions that researchers in this area may wish to explore, as well as outlining some limitations of the current approach.

Experiment 3 examined how three factors impacted novel English vocabulary learning: 1) generation, 2) context sentence vs. definition training, and 3) ambiguity. We expected to find an overall benefit of generation, but hypothesized that this would be modified by word type. Specifically, we expected generation to benefit polysemes and unambiguous words, but either not benefit or hurt homonyms. We also hypothesized that the effectiveness of training with context sentences or definitions would vary across word type, such that definitions would more effective than context sentences for polysemes than for unambiguous words or homonyms. Two exploratory aims of the study were to 1) examine whether the use of generation is more or less effective with context sentences than definitions, and 2) whether individual differences in WM and inhibitory control impact the effectiveness of context sentences vs. definitions.

Across the three outcome measures in this study, we find mixed support for our various hypotheses. We first examine support for the hypothesis that generation benefits learning. Across all three outcome measures we found evidence that generation was an effective learning strategy. In the free recall data, we found a significant benefit of training words with generation. In the meaning production data, we found a benefit of generation although this was restricted to words that were trained with definitions, and did not appear for words that were trained with context sentences. In the sentence completion results we again found a benefit of generation, but this was restricted to words trained with definitions, and furthermore only to unambiguous words and polysemes, but not homonyms. Although the interactions of generation with other variables are

novel findings, which will be discussed more in what follows, the finding of generation effects largely agrees with past literature (e.g., Slamecka & Graf, 1978). It is also interesting to note the implications of these findings for theories of the generation effect that state that generation effects occur only for words that have pre-existing semantic representations (e.g., McElroy & Slamecka, 1982b; Nairne et al., 1985; Payne et al., 1986). The current study clearly demonstrates that pre-existing semantic representations are not a necessary factor for generation effects to emerge.

Our second hypothesis was that the generation effect would be modified by word type. Evidence from all of the outcome measures examined in this experiment generally supported this hypothesis, although this effect always interacted with individual difference variables so it is not straightforward to interpret. These findings are novel and have not been previously reported. In the free recall data, we report that the generation effect did manifest differently for different word types, but only when individual differences in WM were taken into account. Specifically, for individuals with lower WM span participants there was a generation effect for unambiguous words and homonyms, but for individuals with higher WM span there was a generation effect for homonyms and polysemes. In the meaning production data, we found only marginal effects of word type and training condition, and again these were qualified by a marginal interaction with WM span. Finally, in the sentence completion data we reported a significant interaction of word type, generation, and definition training, such that generation was significantly more effective than sentence reading only for words trained with definitions, and of those, only for unambiguous words and polysemes.

Our third hypothesis was that the effectiveness of training with context sentences or definitions would vary across word type, such that definitions would be more effective than context sentences for polysemes than for unambiguous words or homonyms. Again, the results were mixed

and sometimes supported this hypothesis, although this effect was sometimes qualified by interactions with other variables. These novel results are reported here for the first time. Results from the free recall task revealed a marginal four-way interaction of training condition, training method, generation, and waters set size span, which suggests a trend in which word type may interact with definition vs context sentence training. However, because this interaction was not fully significant, we did not probe or interpret it, and further research will be needed to determine if this interaction would appear in other datasets. Results from the meaning production task revealed a significant interaction of training condition and training material, but word type did not interact with training material. Finally, results from the sentence completion task showed that definitions were more effective than context sentences for unambiguous words and polysemes, but only when words were also trained with generation. This finding partially supported our hypotheses—we expected that polysemes would benefit from definition training, but it was unexpected that unambiguous words would as well. This highlights the interesting finding that definitions are not beneficial for training homonyms. Perhaps learners are generally able to understand how a word can be used in multiple different contexts, but understanding that a word can have multiple competing meanings is more difficult. In other words, the definitions condition may have caused the predicted competition between competing homonyms meanings, whereas context sentences appeared not to cause competition in the same way. This finding suggests that using definitions to train homonyms is not advisable.

One of the exploratory aims of this study was to examine whether the use of generation was more or less effective with context sentences or definitions. In the free recall data we did not find any evidence that these variables interacted. However, the meaning production results show a significant interaction of training condition and training material, such that generation effects

were strongest for words trained with definitions. A similar result emerged from the sentence completion results: generation effects were strongest for words trained with definitions.

The second exploratory aim of this study was to investigate whether individual differences in WM and inhibitory control impacted the effectiveness of context sentences vs. definitions. Across the three outcome measures there were no interactions of individual differences in WM and training materials. Therefore, the current results show no relationship between individual differences and training novel vocabulary words with context sentences vs. definitions.

Finally, one result reported in this experiment sheds light on the ways in which various individual differences may interact. In the meaning production data, we report a significant interaction of Waters total span and Simon score, such that the best performance on the meaning production task was observed for participants who were both higher in inhibitory control and had greater WM. This replicates findings from Michael et al. (2011), who reported the same interaction, albeit using different measures of inhibitory control and WM, in a translation production task with moderately-proficient bilinguals. This finding emphasizes the fact that simply having a greater WM span alone may not be helpful if learners do not also have the ability to inhibit task-irrelevant activation, such as from word meanings that are not contextually appropriate.

Taken in sum, several patterns emerge from these findings. First, it does appear that generation effects manifest differently for different type of unambiguous and ambiguous words. However, the patterns are mixed across tasks, and do not always follow the dynamics predicted by the SSD model of semantic ambiguity (Armstrong, 2012). In free recall, we would typically expect to see an ambiguity *advantage* (Degani et al., 2014; Ekves, 2014), and indeed we do see this, although it is restricted to participants with a higher WM span, whereas participants with a lower

WM span exhibit a polysemy disadvantage. This may be because the differences between polysemes meanings are often subtle, and to understand how one meaning differs from another a learner must maintain both representations in WM simultaneously in order to compare and contrast them.

When participants are required to produce a meaning in the meaning production task, we do not see strong differences in how generation effects manifest across word types. However, when participants are asked to complete a sentence, rather than produce a meaning, at least some of the dynamics predicted by the SSD do emerge: generation effects appear only for unambiguous words and polysemes trained with definitions, whereas homonyms are not helped by generation. This may reflect differences in the time span of these tasks, as well as the task demands. The meaning production task was self-timed, and participants could take as long as they needed to produce a response. In contrast, the sentence completion task was time-restricted and participants were instructed to make a response quickly. The SSD states that semantic processing begins when a word is read, and excitatory activation gradually increases over time as the meaning(s) of a word are activated. Early on in processing polysemes obtain activation quickly, because their multiple meanings cooperate, whereas unambiguous words have less activation from having only one meaning, and homonyms receive the least early activation because their multiple meanings compete. Therefore, in a time-limited task, differences between word types would be most apparent. However, as the time course of processing is extended, such as during a task without a time limit, these dynamics change and eventually activation for all word types reaches approximately the same level as processing wraps up. This may explain why we see differences in how different types of words respond to generation in the meaning production and sentence completion tasks.

Finally, there is the question of how much the congruence or incongruence of the training and testing task demands may have impacted the results of this study, especially with regard to the use of context sentences vs. definitions. Transfer-appropriate processing models (Morris et al., 1977) have demonstrated that information is best recalled when the processes that are engaged during initial learning are also engaged during recall. According to the instance-based framework (Bolger et al., 2008), learning words with context sentences and learning words with definitions engage difference processes: whereas learning from context sentences allows learners to build contextualized knowledge of how words are used, learning words from definitions allows learners to acquire a decontextualized and abstract representation of the core meaning of a word. To examine if our results might be influenced by these differences, we included one outcome measure that examines a learner's ability to produce abstract core definitions of words (meaning production) and one task that measures a learner's ability to use words in context (sentence completion). If our results are substantially influenced by transfer-appropriate processing, we would expect an advantage for words trained with definitions on the meaning production task, and a contrasting advantage for words trained with context sentences on the sentence completion task. However, this is not what we found—although we found an advantage for words trained with definitions on the meaning production task, we also found an advantage for words trained with definitions in the sentence completion task (although both of these effects were qualified by an interaction with generation, such that words trained with generation and definitions performed best). Therefore, incongruencies in training and testing demands is not a sufficient explanation for our results.

4.5.1 Future directions

We now turn to a discussion of some limitations of the current study and interesting future directions. In comparison to Experiments 1 and 2, Experiment 3 was based on the least amount of prior research, and was more exploratory than the previous studies. As such, the work suggests a number of exciting future directions, but admittedly has a number of limitations to acknowledge.

We begin with the limitations of this study. First, although the study design was complex, there were a number of other interesting training manipulations that might be more effective than the present design. For example, Bolger et al. (2008) and Fischer (1994) reported that a combination of both definitions and context sentences might be the most effective way to train novel vocabulary words, because this allows learners to acquire both a sense of an abstract definition and how a word is used in context. However, three of the four conditions in the present study learners were given training materials that used only context sentences, or only definitions. In only one condition, in which learners were given definitions in training and then later asked to generate their own context sentence, did they have experience with both contextualized and decontextualized representations. And interestingly, this condition showed the highest level of correct responses in both the meaning production and the sentence completion tasks. Future research should continue to investigate the efficacy of the combination of definitions and context sentences with various types of ambiguous and unambiguous words.

An additional limitation of the current study was that there are a large number of other psycholinguistic variables for which this study was unable to control, given the already large number of predictors in our models. These variables include meaning relatedness, meaning dominance, word frequency, and word concreteness. The last two variables are especially

challenging given the rarity of the words in question—frequency and concreteness information was not available in existing databases for most of the words in the present stimulus set.

There were two exploratory aims of the current study, and the results of these suggest several future directions. First, we demonstrated for the first time that the efficacy of generation depends on the type of training materials used. Whereas generation was effective for words trained with definitions, sentence reading was more effective for words trained with context sentences. Future research should make use of generation in combination with definition training. Additionally, a second exploratory aim of the current study was to investigate the role of individual differences on the effectiveness of context sentences vs. definitions for training novel vocabulary items. We report no effects of individual differences on the effectiveness of context sentences vs. definitions, and therefore future research in this area may not be fruitful.

4.5.2 Conclusions

The present study investigated the use of context sentences vs. definitions for training unambiguous and ambiguous words using generation. We examined this question using a variety of outcome measures that examined multiple facets of word learning. Overall, results revealed complex patterns of interaction between these variables, demonstrating that past research concerning the use of the generation effect for vocabulary learning may has overlooked a number of critical variables that moderate generation effects. One clear conclusion emerged from the results: words trained with both definitions and generation performed best across a variety of tasks. Additional work is needed to determine exactly how best to train novel vocabulary words given these results.

5.0 General discussion

We now turn to a general discussion of the results of all three of the experiments described above. Our aims in this section are to discuss commonalities and differences between the findings, to discuss implications of these findings for models of semantic ambiguity and generation effects, to outline limitations, and finally to provide directions for future research.

In three experiments, we examined whether strengthening meaning representations via the generation of semantically-related material during learning mitigates difficulties associated with the learning of translation-ambiguous words and semantically-ambiguous words. In Experiment 1, we examined whether generation ameliorated the translation-ambiguity disadvantage, and furthermore whether it did so for both homonyms and polysemes. Although we predicted that generation would offset the translation-ambiguity disadvantage for polysemes but increase the disadvantage for homonyms, we ultimately found that generation was beneficial for both polysemes, and either beneficial (for free recall) or not harmful (for translation production) for homonyms (although recall that accuracy for full word-pair recall was very low for free recall, and so we interpret these findings tentatively). We also reported that individual differences in WM and inhibitory control both impacted word learning in different ways for unambiguous words, homonyms, and polysemes. Higher WM scores predicted better performance for homonyms than polysemes and unambiguous words, whereas higher inhibitory control abilities predicted better performance for polysemes than homonyms and unambiguous words.

In Experiment 2, we predicted that generation would offset the ambiguity disadvantage for polysemes, but increase learning difficulties for homonyms. Overall, we report that the effects of generation were weaker in Experiment 2 than Experiment 1, and individual differences in

inhibitory control were necessary for generation effects to emerge. In short, generation neither helped nor harmed ambiguous words, but for participants lower in inhibitory control it was helpful for unambiguous words, and for participants higher in inhibitory control it was detrimental for unambiguous words. In Experiment 3, we examined the same questions as in Experiment 2, but additionally tested whether training words with context sentences or definitions might increase the effects of generation for learning ambiguous and unambiguous words. We predicted that definitions would further enhance meaning representations, thereby increasing the power of the generation effect to offset ambiguity disadvantages for polysemes, and further harming the performance of homonyms. We found an overall benefit of generation that was enhanced by the use of definition training. We also found that generation successfully offset the ambiguity disadvantage for learners with lower WM, and furthermore reported that the best performance was observed for learners with both higher WM and higher inhibitory control.

5.1 Implications for models of semantic ambiguity

The three experiments described above were designed to examine whether the processing dynamics described by the SSD hypothesis (Armstrong, 2012) can be used as a framework to understand ambiguous word learning in L1 and L2. It is important to note that the SSD hypothesis was intended to be a model of word recognition and processing, and was not originally intended to describe vocabulary learning. Therefore, our findings should be interpreted as evidence for and against the extension of the SSD hypothesis to a new domain, and not as evidence for or against the original hypothesis.

As reviewed in more depth in the Introduction, the SSD hypothesis described a pattern of semantic settling dynamics that occur from the time a reader encounters a word to the time that they settle on a contextually-appropriate meaning. Excitatory activation increases as the various meanings of words become activated, pushing a word towards a recognition threshold. Polysemes benefit from early excitatory activation because their multiple related meaning features are related and provide cooperative activation. In contrast, homonyms elicit competitive dynamics from their multiple unrelated meanings, which necessitates later inhibitory feedback to suppress contextually-inappropriate meanings. Unambiguous words do not benefit as much as polysemes from early excitatory activation due to their more limited number of active meaning features, but also do not require later inhibitory feedback as do homonyms, and so they are recognized more quickly.

We hypothesized that these dynamics might also be involved during novel word learning, such that polysemes would benefit from cooperative activation as learners encounter multiple related polyseme senses during word learning, but homonyms would be harmed because of the competitive dynamics that arise when multiple unrelated meanings of a word are encountered. We furthermore hypothesized that these dynamics might be heightened by the use of a sentence generation task that increases semantic processing and semantic activation during learning.

Overall, our results show that generation is particularly beneficial for polysemes, but neither beneficial nor harmful for homonyms. To what extent do these findings support or not support the extension of the SSD hypothesis to novel vocabulary learning? The finding that generation is beneficial for polysemes is in line with the predictions of the SSD, but the lack of evidence for an increased homonymy disadvantage when generation was used is unexpected. We consider this issue next.

There are a number of possible explanations for this finding. First, it is important to note that even the original work that proposed the SSD (Armstrong, 2012) did not always consistently produce the dynamics the SSD model predicted, in that a homonymy disadvantage was inconsistently observed in the behavioral results. For example, Armstrong's Experiment 1 replicated a polysemy advantage but did not find a homonymy disadvantage in lexical decision, although in Experiment 2 there was a homonymy disadvantage when the response time latency was increased due to stimulus degradation. This is consistent with a number of other reports of failures to find a homonymy disadvantage (Hino & Lupker, 1996; Hino et al., 2002). Thus, it appears that although it is generally easy to observe the cooperative dynamics that promote polyseme processing, the dynamics surrounding homonym processing and learning are less clear, and furthermore vary across tasks.

We next examine how task requirements may explain some of our findings. We report that in Experiment 1 generation benefitted both homonyms and polysemes in free recall. Free recall required that learners retrieve a word form, but it does not require meaning access nor does it require learners to select a contextually-appropriate meaning. Selecting a contextually-appropriate meaning is one of the key drivers of the dynamics proposed by the SSD. When meaning selection is required, homonyms alone require inhibitory feedback to suppress irrelevant meanings, and it is this inhibitory feedback that slows homonym processing and results in a homonymy disadvantage. Interestingly, in the translation production results, generation benefitted polysemes, but neither benefitted nor harmed homonyms. Translation production both requires meaning access and meaning selection, and so this task may have been better suited to elicit the dynamics proposed by the SSD hypothesis than free recall tasks.

One additional possibility for why polysemes and homonyms performed similarly is that participants may not have been exposed to the words enough times to be able to distinguish between unrelated meanings and related senses. Overall, participants only had three total encounters with the words: two encounters with the word (or word pair in the case of Experiment 1) and its definition, and one encounter with the word during the sentence generation/sentence reading condition. This was done because the sentence generation/sentence reading portion of the training is longer than most typical training tasks, but the outcome may have been that participants simply did not have enough encounters with a word to notice subtle differences between meanings and decide if the meanings were similar or not. Therefore, it is possible that to the participants, all ambiguous words appeared to be functionally the same. However, Hulme, Barsky, and Rodd (2019) demonstrated that only two encounters with ambiguous words in context is sufficient for a large portion of learners (close to 40%) to establish accurate meaning representations, and gains in accuracy from increasing the number of encounters would be modest.

There are important implications of these finding for models of semantic ambiguity. A large debate in the literature has been the degree to which meaning relatedness matters for processing ambiguous words (for a review, see Eddington & Tokowicz, 2015). The present study provides evidence that meaning relatedness is not as important a factor for vocabulary learning as it is for word processing. Rather it seems that number of meanings/senses matters more, at least when the words in question are unambiguous words with one meaning and ambiguous words with two meanings/senses. Future research should investigate whether increasing the number of meanings/senses may make ambiguous words even harder to learn.

Another interesting consideration is whether there are models of word learning that may better describe ambiguous word learning than the SSD hypothesis. For example, the instance-

based framework (Bolger et al., 2008; described in the Introduction to Experiment 3), may be well-suited for this purpose. This model described novel word learning as an incremental process in which the meaning of a word is built up over multiple encounters with the word in context. According to Bolger et al., a resonance mechanism such as that proposed by Myers and O'Brien (1998) may be involved in the process of deriving word meanings from context. According to this view, when a word is encountered in context, semantically-related words in the lexicon become activated. This activation spreads as a function of how much featural overlap there is between a concept and its representation in context. This co-activation of a target word and associated words is encoded in episodic memory traces, and this pattern of activation becomes part of the contextual meaning representation of the word.

The original resonance model of Myers and O'Brien (1998) was a model of sentence comprehension. The model outlined how new information has to be integrated with previous information during sentence processing, which may require reactivation of related, previously-encountered concepts. This parallels processes that occur during ambiguous word learning. When learners encounter a new meaning/sense for a homonymous or polysemous word, they attempt to access the previous meaning and integrate the new meaning/sense with the existing meaning/sense (Fang & Perfetti, 2017, 2019; Maciejewski, Rodd, Mon-Williams, & Klepousniotou, 2019). This type of model may explain some of the results observed in Experiments 1-3, particularly the ways in which we observed differential effects for different types of words. For example, in a resonance-based model, resonance can either activate information that facilitates word comprehension (such as activation of the related senses of polysemes), or it can activate information that is disruptive to comprehension (such as the activation of unrelated or contextually-inappropriate meanings of

homonyms). The resonance model has been applied to anaphor resolution, but not yet to semantic ambiguity resolution.

A resonance model could also account for the facilitative effects of sentence generation observed in these experiments. Generation can be thought of as a type of semantic elaboration, and the process of generating semantically-meaningful sentences during training would increase overall resonance by activating additional words and their associations. In contrast to depth-of-processing views (Craik & Lockhart, 1972), which expect any semantic elaboration to be helpful, resonance from sentence generation would be helpful only insofar as the generated material was actually semantically related to the target word. Generation of unrelated material would result in activation of material that would interfere rather than facilitate comprehension.

5.2 Implications for models of generation effects

In the Introduction, we review two theories of generation effects: *the two-factor theory* (Hirshman & Bjork, 1988) and the *enhanced semantic processing hypothesis* (McElroy, 1987). To briefly review, the two-factor theory proposed that generation strengthens both form-form connections and meaning representations, whereas the *enhanced semantic processing hypothesis* proposed that generation effects simply strengthen meaning representations, but not necessarily form-form connections. We proposed that if we found generation effects in tasks that require strong connections between word forms, such as translation production and free recall, then we can interpret this as evidence to support the two-factor theory. In contrast, if we found generation effects in tasks that require meaning access (such as meaning production), but not in tasks that require only form representations (such as free recall) or form-form connections (such as

translation production), this would support the enhanced semantic processing theory but not the two-factor theory.

Evidence supporting this from the results of Experiments 1-3 is mixed. One clear pattern is that we found evidence of generation effects in all three experiments in free recall, whereas we only found evidence of generation effects in translation production in Experiment 1, and found no evidence of generation effects in meaning production in either Experiment 2 or 3. This suggests that generation was highly effective for tasks the require strong form representations and form-form connections, but may not have been as effective for enhancing meaning representations. This offers support for the *two-factor* theory of generation effects.

The possibility that sentence generation did not strongly enhance meaning representations in Experiments 2 and 3 has important implications for these studies. We proposed that only under conditions of enhanced semantic activation would the dynamics predicted by the SSD emerge. Because we did not find generation effects in the meaning production results, it is not surprising that we also did not observe differences between word types. It may be that the sentence generation task simply did not cause participants to engage with the meaning representation of the word as much as we had hoped. The generation task required participants to write a meaningful sentence that contained the word they were trying to learn, but one important thing to note is that the participants were able to read the definition of the word when they were writing the sentences. Therefore, they did not have to retrieve the meaning of the word, and this might have resulted in less engagement with meaning representations as a result. Correspondingly, this may have required more engagement with form representations than we anticipated.

Additionally, although we verified that the sentences that were generated contained the correct target word, and manually inspected a small number of responses for each participant to

ensure the participant had followed directions and attempted to write meaningful sentences, we did not comprehensively assess if sentences were an accurate representation of the meaning of the target word. It is possible that participants generated sentences that did not capture the meaning of the target word, and so could not promote enhanced meaning representations. Future research should examine this possibility. An additional possibility is that a different generation task, such as definition generation, may have been better suited to promoting meaning access.

Overall, there is a noticeable pattern in which the predicted results emerged for L2 vocabulary learning, but largely did not emerge for L1 vocabulary learning, demonstrating that generation was more helpful for L2 vocabulary than L1 vocabulary. There are a number of reasons why this might be the case. First, the L1-L2 word pairs were much more common than the rare L1 words in Experiments 2 and 3. In the L2 experiment (Experiment 1), participants were learning to map an L2 word form to existing L1 form and meaning representations. Because of these pre-existing meaning representations, generation may have been more effective at increasing semantic activation. In contrast, in the L1 experiments (Experiments 2 and 3) the words were rare enough that it was likely that they did not have strong pre-existing meaning representations (although see Experiment 2 for discussion of how rare word synonyms may have contributed to weak existing meaning representations). Past research has demonstrated that generation is of limited effectiveness when applied to novel material (Gardiner & Rowley, 1984; Lutz et al., 2003; McElroy & Slamecka, 1982; Nairne et al., 1985; Payne et al., 1986). This might be why we did not find consistent effects of generation in Experiments 2 and 3: the words might have been too unfamiliar and the absence of strong pre-existing meaning representations may have limited the benefit of enhanced semantic activation. Future studies could examine this possibility by

increasing the number of encounters learners have with rare words or training words across multiple sessions.

5.3 The role of individual differences

The results of Experiments 1-3 highlight the important and varied roles that individual differences in WM and inhibitory control play in L1 and L2 vocabulary learning. In Experiment 1, both WM and inhibitory control abilities were important. In English free recall, participants with higher WM were better able to recall homonym meanings. However, when generation was considered, inhibitory control and not WM was important: participants with greater inhibitory control benefitted more from the use of generation than did participants with higher WM. When participants were required to recall known English words they had recently viewed, higher WM allowed them to recall more words, but the ability to suppress task-irrelevant information was crucial for generation to benefit performance. However, in German free recall and translation production there was no effect of WM, and instead inhibitory control interacted with polyseme recall. Participants with stronger inhibitory control performed better than participants with weaker inhibitory control, but this was most apparent for polysemes: whereas participants with weaker inhibitory control performed most poorly on polyseme recall, participants with stronger inhibitory control performed significantly better on polyseme recall. This is a challenging finding to interpret, because for polysemes the SSD does not predict competitive activation that would necessitate the activation of inhibitory control. However, we also find this same pattern in the translation production results, and so it bears consideration. One possibility is that learners treated homonyms as one-to-one mappings (i.e., they mapped German words to a unique meaning representation

because homonym meanings are distinct) but treated polysemes as one-to-two mappings (i.e., they viewed the multiple senses of polysemes as part of the same core meaning, and mapped two word forms to one unified meaning). If this were the case, participants who were better able to suppress one of the word forms might have an advantage over participants who were unable to suppress one of the word meanings.

Again in Experiment 2, both WM and inhibitory control had separate and distinct effects. In free recall, only Simon score affected performance. Here we observed that participants with higher inhibitory control performed significantly better on homonyms than participants with lower inhibitory control, but only when words were trained with generation. This suggests that participants who were better able to suppress interference from irrelevant homonym meanings performed better, but this interference suppression was only necessary when generation enhanced semantic processing. In meaning production, participants with higher WM performed better overall, as did participants with better inhibitory control, but these variables did not interact with each other, with generation, or with word type.

In Experiment 3, both WM and inhibitory control were again important for understanding outcome measures. In free recall, individuals with higher WM performed significantly worse on polysemes when they were trained without generation than did individuals lower in WM. This finding implies that participants who were better able to remember that a word had multiple senses, but did not have the benefit of generation, performed most poorly. This difference disappeared when polysemes were trained with generation. In meaning production, an interaction of WM and inhibitory control emerged such that participants who were higher in both WM and inhibitory control performed best. This interaction is consistent with past findings by Michael et al. (2011). There were no effects of individual differences in the forced choice sentence recognition task.

Overall, both WM and inhibitory control played important roles in vocabulary learning with ambiguous words. Future studies in this area should consider the potential effects of both variables.

5.4 Future directions

There are a wealth of future directions suggested by this work. These fall into two main categories: 1) future uses of generation, 2) methodological refinements, and 3) models of ambiguous word learning.

First, with regard to future applications of generation, the most clear and exciting possibility is the use of generation to offset the translation-ambiguity disadvantage. In Experiment 1, we present evidence that at least for polysemes, generation successfully offset the translation-ambiguity disadvantage. Whereas past research into alleviating this disadvantage has primarily focused on establishing the appropriate form-meaning connections (e.g., Degani et al., 2014), we provide evidence that training methods that aim to strengthen meaning representations may be a fruitful line of inquiry. Additionally, we provide evidence that generation is particularly effective for word learning when it is used in combination with definition training. Future research investigating how best to achieve this pairing may also yield interesting results.

Second, there are a number of methodological refinements that would improve the results presented above. First, an open question is the extent to which the sentence generation task elicited responses that were semantically appropriate and helpful for learning target words. Future research may benefit from developing methods of assessing the semantic fit of learner-generated sentences.

This could be done manually (by having human raters score sentences), or via an automated procedure such as the MESA algorithm described by Frishkoff et al. (2008). Additionally, an important future direction is replicating or extending the research described in Experiments 1-3 with the goal of improving learning and accuracy. Across all three experiments accuracy was generally low, which, as discussed above, may be partially due to the relatively low number of encounters with words. As reported by Beck, Perfetti, and McKeown (1982), high levels of word knowledge are difficult to obtain, even with many encounters with a word. However, either increasing the number of encounters or increasing training so that it occurs on multiple days may improve performance. This would allow us, for example, to be able to examine full word pairs in free recall in Experiment 1. Finally, as described above in more detail, there are a number of psycholinguistic variables that we did not address in the present study. Now that we have demonstrated that word type and generation interact during novel word learning, future research should investigate whether the efficacy of generation is affected by factors such as word frequency, meaning relatedness, or concreteness.

Third, one additional variable which is related to language learning outcomes is vocabulary knowledge, which was not directly measured in this study. The number of both L1 (e.g., Ouelette, 2006) and L2 (e.g., Elgort & Warren, 2014; Ferrel Tekmen & Daloğlu, 2006) words that a learner knows predicts acquisition of novel words from context. Although we did not directly measure individual differences in word knowledge, we selected a working memory measure (i.e., Waters Reading span) that partially captures linguistic knowledge, but is also important for unambiguous word learning. Future studies may wish to investigate whether word knowledge is as important for learning ambiguous words as it is known to be for unambiguous words.

Lastly, the present research investigated whether a model of semantic ambiguity resolution, the SSD hypothesis, could be extended from word recognition and processing to vocabulary learning. The SSD hypothesis proved useful for understanding some of the dynamics described in the Results above, but ultimately its applicability to vocabulary learning is limited by the fact that it was designed to examine much earlier time windows of lexical and semantic processing than are typically of interest for vocabulary learning. This research highlights a need for further development of models of ambiguous word learning, and suggests that the instance-based framework (i.e., Bolger et al., 2008), and resonance models (i.e., Myers & O'Brien, 1998) may provide useful models for building a better theoretical understanding of translation-ambiguous and semantically-ambiguous word learning. Furthermore, the SSD hypothesis has typically examined knowledge of well-known words after consolidation of word meaning has occurred. In contrast, in the present study we examined word knowledge immediately after learning, and again a week later. It is possible that because we investigated the early stages of word learning we were primarily examining words for which learners only had episodic, contextualized knowledge. In contrast the predictions of the SSD were derived from studies in which learners had decontextualized, consolidated word representations. Future research should investigate this possibility.

5.5 Conclusions

In conclusion, the results of these experiments provided support for the two-factor theory of generation, the instance-based framework for word learning, and limited support for the extension of the SSD hypothesis of semantic ambiguity resolution to ambiguous word learning. We showed that generation was an effective tool for offsetting the translation-ambiguity

disadvantage, and furthermore that generation affected unambiguous words, polysemes, and homonyms in different ways. We highlighted the important roles that WM and inhibitory control play in these processes, and showed that the use of both generation and definitions was optimal. These findings emphasize the need to build new models of ambiguous word learning, not just processing, and provide some direction for productive avenues for future research.

Appendix A Experiment 1 stimulus characteristics

Table 24. Experiment 1 stimulus characteristics

English Word	German Word	Word Type	Dominance	English Length	German Length	English Concreteness	English Frequency
drill	Bohrer	homonym	dom	5	6	4.40	2.85
drill	Übung	homonym	sub	5	5	4.40	2.85
match	Streichholz	homonym	dom	5	11	4.14	3.40
match	Gegenstück	homonym	sub	5	10	4.14	3.40
mold	Schmimmel	homonym	dom	4	9	4.85	2.34
mold	Abdruck	homonym	sub	4	7	4.85	2.34
pitcher	Krug	homonym	sub	7	4	4.93	2.22
pitcher	Werfer	homonym	dom	7	7	4.93	2.22
present	Geschenk	homonym	dom	7	8	3.39	3.66
present	Gegenwart	homonym	sub	7	9	3.39	3.66
pupil	Sehloch	homonym	dom	5	7	4.55	2.21
pupil	Schulkind	homonym	sub	5	9	4.55	2.21
root	Wurzel	homonym	dom	4	7	4.34	2.73
root	Ursprung	homonym	sub	4	8	4.34	2.73
scale	Waage	homonym	dom	5	5	4.39	2.69
scale	Schuppe	homonym	sub	5	7	4.39	2.69
toast	Röstbrot	homonym	dom	5	8	4.93	3.23
toast	Trinkspruch	homonym	sub	5	11	4.93	3.23
trunk	Kofferraum	homonym	dom	5	10	4.71	3.00
trunk	Rüssell	homonym	sub	5	7	4.71	3.00
arena	Kampfbahn	polyseme	dom	5	9	4.83	2.27
arena	Schauplatz	polyseme	sub	5	10	4.83	2.27
atmosphere	Lufthülle	polyseme	dom	10	9	3.04	2.69
atmosphere	Stimmung	polyseme	sub	10	8	3.04	2.69
bottle	Flasche	polyseme	dom	6	7	4.91	3.41
bottle	Schoppen	polyseme	sub	6	8	4.91	3.41
cotton	Baumwolle	polyseme	dom	6	9	4.97	2.86
cotton	Wette	polyseme	sub	6	5	4.97	2.86
doll	Puppe	polyseme	dom	4	5	5.00	3.10
doll	Schatz	polyseme	sub	4	6	5.00	3.10
mouth	Mund	polyseme	dom	5	4	4.74	3.73
mouth	Öffnung	polyseme	sub	5	7	4.74	3.73

pipe	Pfeife	polyseme	dom	4	6	4.88	3.00
pipe	Rohr	polyseme	sub	4	4	4.88	3.00
sheet	Laken	polyseme	dom	5	5	4.93	2.77
sheet	Blatt	polyseme	sub	5	5	4.93	2.77
shower	Brause	polyseme	dom	6	6	4.89	3.32
shower	Regenfall	polyseme	sub	6	9	4.89	3.32
sign	Zeichen	polyseme	sub	4	7	4.62	3.83
sign	Schild	polyseme	dom	4	6	4.62	3.83
arrow	Pfeil	single		5	5	4.97	2.60
art	Kunst	single		3	5	4.17	3.56
bird	Vogel	single		4	5	5.00	3.37
bone	Knochen	single		4	7	4.90	3.12
boot	Stiefel	single		4	7	4.96	2.76
candle	Kerze	single		6	5	4.86	2.61
chain	Kette	single		5	5	4.55	3.03
cloud	Wolke	single		5	5	4.54	2.78
color	Farbe	single		5	5	4.08	3.30
coward	Feigling	single		6	8	2.93	2.87
example	Beispiel	single		7	8	3.03	3.18
face	Gesicht	single		4	7	4.87	4.17
funeral	Beerdigung	single		7	10	3.83	3.23
head	Kopf	single		4	4	4.75	4.28
juice	Saft	single		5	4	4.89	3.14
knight	Ritter	single		6	7	4.79	3.14
meat	Fleisch	single		4	7	4.90	3.35
mirror	Spiegel	single		6	7	4.97	3.09
monkey	Affe	single		6	4	4.90	3.23
recovery	Erholung	single		8	8	2.68	2.67
river	Fluss	single		5	5	4.89	3.45
road	Strasse	single		4	7	4.75	3.76
roof	Dach	single		4	4	4.79	3.26
scar	Narbe	single		4	5	4.74	2.64
spine	Rückrat	single		5	7	4.88	2.47
task	Aufgabe	single		4	7	2.84	2.81
tension	Spannung	single		7	8	2.60	2.64
trash	Müll	single		5	4	4.70	3.06
voice	Stimme	single		5	6	4.13	3.64
wing	Flügel	single		4	7	4.86	3.01

Appendix B Sentence norming for Experiment 1

We collected normative ratings for sentences containing target words from Experiment 1. Context sentences were generated for each word pair (two sentences per meaning).

B.1 Methods

B.1.1 Participants

Participants were 10 native English-speaking undergraduate students from the University of Pittsburgh, were 18 years or older, and had normal or corrected-to-normal vision, and were right-handed. Participants were recruited from the Introduction to Psychology Subject Pool and were compensated with class credit for their participation. Participation took approximately one hour.

B.1.2 Procedure

Sentence ratings were collected using Qualtrics. Participants were presented with a short English sentence with one word missing, and asked to type in the first English word that came to mind to complete the sentence. In total participants rated 140 sentences (2 sentences per meaning). The mean proportion of responses that included the intended English target word was calculated for each sentence. For the final sentence set used in Experiment 1, we selected the one sentence

from each pair that had the higher proportion of correct responses (with a few exceptions that were made to ensure the average proportion correct was matched across word type). These ratings and the final stimuli set appear below in Table 25.

Table 25. Experiment 1 stimulus definitions and sentences

English Word	German Word	Definition	Sentence	Proportion correct
drill	Bohrer	a shaft-like object with cutting edges for making holes in firm materials	He used an electric _____ to screw the shelves into the wall.	0.7
drill	Übung	any strict, methodical or repetitive training	His coach pushed him through the demanding training _____ for practice.	0
match	Streichholz	a slender piece of flammable material tipped with a chemical substance that produces fire	The _____ was used to light a candle.	0.6
match	Gegenstück	a person or thing that equals or resembles another in some respect	Because they were a pair the two shoes were a perfect _____ for each other.	0.7
mold	Schmimmel	a growth of minute fungi	Pasta sauce will grow _____ if left out for a few days.	0.5
mold	Abdruck	a hollow form or matrix for giving a particular shape to something	She had ordered a plaster _____ to cast the exact shape she wanted.	0.8
pitcher	Krug	a container, usually with a handle and spout or lip, for liquids	She filled the metal _____ up with water to refill glasses.	0.7
pitcher	Werfer	the player who throws the ball to the opposing batter	The batter struck out because the _____ was the best in the league.	0.9
present	Geschenk	a thing presented as a gift	The child looks forward unwrapping her birthday _____ every year.	0.8
present	Gegenwart	being, existing, or occurring at this time	During roll call, one student was not _____ due to illness.	0.8

pupil	Sehloch	a person learning under the close supervision of a teacher	The _____ asked his tutor a question about the assignment.	0
pupil	Schulkind	the expanding and contracting opening in the iris of the eye	His _____ shrunk when he stepped outdoors and into the bright light.	0.7
root	Wurzel	the part of a plant that anchors it in the ground	The weed had a strong _____, she struggled to pull it out.	0.8
root	Ursprung	the place where something starts, where it springs into being	The _____ of the problem was that she had stolen from him.	0.5
scale	Waage	an instrument or device for weighing	She placed the produce on the _____ to weigh out two pounds.	1
scale	Schuppe	one of the thin, flat plates forming the covering of certain animals	The salmon was not prepared carefully; he found a shiny _____ in his dish.	0.8
toast	Röstbrot	sliced bread that has been browned by dry heat	She grabbed a slice of buttered _____ for breakfast before rushing to the bus.	0.3
toast	Trinkspruch	a drink in honor of or to the health of a person or event	They made a champagne _____ to him on his 50th birthday.	0.9
trunk	Kofferraum	compartment in an automobile that carries luggage	The _____ of the car was too small to fit all of the suitcases.	1
trunk	Rüssell	a flexible snout of a large mammal	The elephant's _____ is used to wash itself and to pick leaves from trees.	0.9
arena	Kampfbahn	a central stage or ring used for sports or other forms of entertainment	The gladiators entered the _____ to do battle before the crowd.	0.4
arena	Schauplatz	a sphere of conflict or intense activity	She decided to enter the political _____ to fight for her beliefs.	0
atmosphere	Lufthülle	the gaseous envelope surrounding the earth	The Earth's _____ has five layers including the exosphere.	0.6
atmosphere	Stimmung	a general pervasive feeling	The _____ at the Superbowl is exhilarating because of the excitement of the crowds.	0.2
bottle	Flasche	a portable container for holding liquids, made of glass or plastic	He bought a reusable _____ for water to reduce plastic waste.	0.8

bottle	Schoppen	bottled milk or substitute mixtures given to infants	He fed the hungry baby a _____ of milk.	1
cotton	Baumwolle	a natural type of cloth or thread	He made a shirt out of soft, white _____ cloth.	0.7
cotton	Wette	a plant with soft, white, downy substance attached to the seeds	_____ was a key crop in the American South during the Civil War era.	0.9
doll	Puppe	a small figure representing a baby or other human being	The child dressed her _____ and pushed it around in a small stroller.	1
doll	Schatz	a generous or helpful person	The old lady asked her to be a _____ and help with the groceries.	0.6
mouth	Mund	the opening through which many animals take in food and issue vocal sounds	She knew the answer and opened her _____ to speak.	1
mouth	Öffnung	the opening of or place leading into a cave, tunnel, volcano, etc.	They walked cautiously into the dark _____ of the cave.	0.3
pipe	Pfeife	tube with a small bowl at one end, used for smoking	Popeye ate spinach and smoked tobacco from a large _____.	0.8
pipe	Rohr	a hollow cylinder of metal, wood, or other material	An old _____ had burst and flooded the bathroom.	0.9
sheet	Laken	a large rectangular piece of fabric generally one of a pair used as inner bed clothes	For his Roman costume he draped a white _____ around him like a toga.	0.8
sheet	Blatt	a piece of printed paper to be folded into a section	She asked to borrow a _____ of paper because she'd forgotten her notebook.	0
shower	Brause	a room or booth containing a plumbing fixture that sprays water over you	His old clawfoot tub did not have a modern _____ attachment.	0.5
shower	Regenfall	a brief period of rain, hail, sleet, or snow	The one day she didn't bring her umbrella there was a quick _____.	0.1
sign	Zeichen	a perceptible indication of something not immediately apparent	She gave him no warning _____ before tackling him to the ground.	0.7

sign	Schild	a public display of a message	The _____ on the door announced the meeting was moved to a different room.	0.9
arrow	Pfeil	a slender and straight weapon made to be shot	She shot the _____ from her bow right into the bullseye.	1
art	Kunst	the products of human creativity	Both Monet and Van Gogh created extraordinary works of _____.	0.9
bird	Vogel	any warm-blooded vertebrate, having a body covered with feathers, has wings, scaly legs, and a beak	A _____ could be heard chirping on the branch outside her window.	1
bone	Knochen	one of the structures composing the skeleton	She fell off her bike, broke a _____, and had to wear a cast.	0.4
boot	Stiefel	a covering of leather, rubber, or the like, for the foot and all or part of the leg	Her old _____ had a hole in the bottom that let in the rain.	0
candle	Kerze	a long, usually slender piece of tallow or wax with an embedded wick that is burned	He lit the _____ when the power went out.	0.8
chain	Kette	a series of objects connected one after the other	The delicate pearls were beaded through a thin metal _____.	0.3
cloud	Wolke	a visible collection of particles of water or ice suspended in the air	The _____ blocked the sun, and threatened to rain out the event.	0.1
color	Farbe	the quality of an object or substance with respect to light reflected by an object	The shirt was a bright _____ that matched his shoes perfectly.	0.8
coward	Feigling	a person who lacks courage in facing danger, difficulty, opposition, or pain	The soldier ran away from battle and was branded a _____.	0.6
example	Beispiel	one of a number of things, or a part of something, taken to show the character of the whole	To clarify what she meant she gave him a _____.	0.2
face	Gesicht	the front part of the head, from the forehead to the chin	He was worried; she could see it on his _____.	1
funeral	Beerdigung	the ceremonies for a deceased person prior to burial or cremation	Many sad friends sent flowers to the old man's _____.	0.4

head	Kopf	the upper part of the body in humans, joined to the trunk by the neck	He turned his _____ towards the speaker at the front.	0.5
juice	Saft	the natural fluid content, or liquid part the can be extracted from a plant or one of its parts	She made her own orange _____, the store-bought ones had too much sugar.	1
knight	Ritter	a mounted soldier serving under a feudal superior in the Middle Ages	The _____ in armor knelt before the queen.	1
meat	Fleisch	the edible part of anything	He said he wasn't vegan because he liked _____ too much.	0.8
mirror	Spiegel	a surface, such as polished metal or glass coated with a metal film, that reflects light	She checked her outfit in the _____ before leaving.	1
monkey	Affe	any primate except man	The playful _____ hooked its tail around the tree and swung back and forth.	0.9
recovery	Erholung	restoration to a former or better condition	He had a slow _____ after his car accident left him injured.	0.7
river	Fluss	fresh water flowing along a definite course	The _____ had dried up due to the drought and water was rationed.	0.1
road	Strasse	a long, narrow stretch with a smoothed or paved surface, made for traveling	They drove down a long, winding gravel _____ in the countryside.	1
roof	Dach	the external upper covering of a house or other building	During the heavy rain, the _____ on the house sprung a leak.	0.6
scar	Narbe	a mark left after skin is damaged	The old _____ on his finger reminded him to be careful with knives.	0.9
spine	Rückrat	the series of vertebrae forming the axis of the skeleton and protecting the spinal cord	The X-ray showed the nerve damage on the lower vertebrae of her _____.	1
task	Aufgabe	any piece of work that is undertaken or attempted	She checked the completed _____ off of her list.	0.8
tension	Spannung	the act of stretching or straining	The extra weight added too much _____ to the rope swing and it broke	0.7

trash	Müll	worthless material that is to be disposed of	He threw the candy wrapper into the _____.	0.3
voice	Stimme	the distinctive quality or pitch or condition of a person's speech	She called out in a loud _____ in order to attract his attention.	0.5
wing	Flügel	a moveable organ for flight	The bird gently shook its injured _____ before flying away slowly .	0.6

Appendix C Language history questionnaire data for Experiment 1

Table 26. Language history questionnaire data for Experiment 1

	Experiment 1	
	M	SD
Proportion female	0.42	-
L1 reading proficiency	9.58	1.58
L2 reading proficiency	4.56	2.55
L1 writing proficiency	9.85	0.37
L2 writing proficiency	3.80	2.50
L1 conversational fluency	9.96	0.20
L2 conversational fluency	4.56	2.72
L1 speech comprehension	9.96	0.20
L2 speech comprehension	4.96	2.88

L1 and L2 proficiency, fluency, and comprehension scores were self-reported on a scale of 1-10, where 1 indicated very low and 10 indicated very high.

Appendix D Sentence norming procedures for Experiments 2 and 3

We collected normative ratings for sentences containing the 40 target words (20 unambiguous, 10 homonyms, 10 polysemes) used in Experiments 2 and 3. Context sentences were generated for each word (two sentences per meaning), for a total of 120 sentences.

D.1 Methods

D.1.1 Participants

Participants were 20 native English-speaking undergraduate students from the University of Pittsburgh, 18 years or older, and had normal or corrected-to-normal vision, and were right-handed. Participants were recruited from the Introduction to Psychology Subject Pool and were compensated with class credit for their participation. Participation took approximately one hour.

D.1.2 Procedure

Sentence ratings were collected using Qualtrics. Participants were presented with a target word and its definition, followed by a short English sentence containing the target word. They were asked to read the word, definition, and sentence, and rate how well the sentence captured the meaning of the word on a Likert scale from 1 (not well at all) to 7 (extremely well). Data from one participant were removed due to incompleteness, leaving data from 19 participants for analyses.

In total, each participant rated 120 sentences (2 sentences per meaning). Ratings were averaged across participants to create an average rating for each sentence. For the final stimulus set used in training in Experiments 2 and 3, we selected the sentence for each word meaning that had the higher average rating of the two available (with a few exceptions that were made to ensure the average rating was matched across word type). Sentence ratings were matched across word types for both sentence 1: $F(37) = 0.44$, $p = .65$, and sentence 2: $F(17) = 1.84$, $p = .19$. These ratings and the final stimuli set appear below in Table 27.

Table 27. Stimulus and sentence ratings for Experiments 2 and 3

Word	POS	Type	Definition1	Sentence 1	Sentence 1 rating	Definition2	Sentence 2	Sentence 2 rating
nutant	ADJ	H	drooping; nodding; used in the context of describing botany	Because of the drought, sunflowers became nutant; they drooped over.	6.0	peevish; irritable; cranky	Babies will start getting nutant if they are not fed.	6.2
aphotic	ADJ	H	characterized by or growing in the absence of light	Some plants are aphotic; they can grow without sunlight.	6.5	relating to the region of a body of water that is not reached by sunlight	The submarine went down into the aphotic zone of the ocean, where it was completely dark.	6.3
plangent	ADJ	H	resounding loudly, especially with a plaintive sound, as a bell	The bell rang with a resounding plangent sound.	5.9	uncultivated, wild	After years of neglect, the yard had grown wild and plangent.	5.7
arrect	ADJ	H	(of animals' ears) pricked up	While grazing, the deer's ears suddenly became arrect to a sound behind it.	6.0	attentive	The beaver stood arrect near its den to alert the others of approaching predators.	5.6
discinct	ADJ	H	without a belt	Some jeans might look wrong when they are discinct because of the empty belt loops.	5.7	loosely dressed	His clothes were discinct, with baggy cargo pants and a long, loose shirt.	5.6
asperity	N	H	Harshness of manner; ill temper or irritability	Marsha was upset by the asperity in her daughter's retort.	5.3	A pressing or urgent situation	In a moment of asperity, Sarah was forced to beg for aid from her parents.	5.3
vicissitude	N	H	The quality of being changeable; mutability	The work was subject to the vicissitude of the weather, which slowed progress considerably	4.8	Sincere remorse for wrongdoing; repentance	George‚Äôs vicissitude was apparent when he spoke tearfully of his use of bribery and blackmail.	5.0
corrigibility	N	H	Capable of being corrected, reformed, or improved	It is not yet safe to enter the building, but the firefighters assured us of the problem’s corrigibility.	4.5	Overbearing pride or presumption; arrogance	After winning the championship, he became so full of corrigibility that no one could stand him.	5.5
indemnity	N	H	Security against damage, loss, or injury	It would be wise to have some indemnity in case something should happen to you.	5.4	A deceptive stratagem or device	The clever indemnity won him the card game but upset his opponents.	4.9

levity	N	H	Lightness of manner or speech, especially when inappropriate	His unrestrained levity is in stark contrast to the sternness of his father.	5.3	Foolhardy disregard of danger; recklessness	The officer scolded the teenagers for their levity, cautioning them to drive with care.	5.6
dumose	ADJ	P	filled with bushes	The landscape was dumose, dense with bushes and shrubs.	6.3	having a bushlike manner of growing	The children's book shows that old Saint Nicholas has a dumose beard.	3.7
recondite	ADJ	P	difficult or impossible for one of ordinary understanding or knowledge to comprehend	Theoretical physics was too recondite for her to understand with her basic physics knowledge.	5.8	hidden from sight; concealed	It was important that the recondite mission wasn't ever discovered by the enemy.	5.0
saturnine	ADJ	P	sluggish in temperament; gloomy; taciturn	After receiving a bad grade, his mood turned saturnine.	4.6	suffering from lead poisoning	Children are at greater risk to be saturnine, because they put lead-contained objects in mouths.	5.5
otiose	ADJ	P	being at lesiure; idle; indolent	Some people find it harder to be otiose than working; they enjoy being busy.	6.1	ineffective or futile	Their attempts were impressive but ultimately otiose; nothing has changed.	6.2
ringent	ADJ	P	having the mouth wide open; gaping	After hearing the shocking news, she stood there with her mouth ringent.	6.0	In biology, having the lips separated by a distinct gap	Some flowers' petals have a distinct gap separating each one, and are characterized as ringent.	6.3
apotheosis	N	P	An exalted or glorified example	The founder of the McDonald's franchise is an apotheosis of the American entrepreneurial success.	5.4	Elevation to the status of a god	After his death, the Romans had to adjust to the apotheosis of Caesar	3.7
ignominy	N	P	Great personal dishonor or humilition	She was unable to avoid the ignominy of having failed.	5.1	Behavior or a quality that merits disgrace or dishonor	The incumbent party experienced the ignominy of defeat in the last election	4.3
lacuna	N	P	An empty space or a missing part; a gap	The lacuna in her argument merits consideration and could ultimately be devastating to her case.	4.4	In anatomy, a small pit or cavity	The ancient skull had a shallow lacuna from an old wound.	5.6

riposte	N	P	A retaliatory action, maneuver, or retort	The mayor felt it necessary to make a riposte to the reporter's untruthful comments	5.4	In fencing, a quick return thrust made after parrying a lunge by one's opponent	The fencing student made a quick riposte and scored a point	
spate	N	P	A sudden flood, rush, or outpouring	The author received a spate of fan mail after she published a highly acclaimed book.	6.2	A large number or quantity	After ordering takeout all week, Sarah had a spate of empty Tupperware containers	5.6
cautelous	ADJ	U	crafty or cunning	Foxes are seen as cautelous animals because of their crafty nature.	5.7			
vagient	ADJ	U	crying like a child	After the tragic movie ending, she was vagient and searched for a tissue to wipe her tears.	5.2			
ennomic	ADJ	U	lawful; legal	Robbery and murder are clearly not ennomic.	5.1			
esculent	ADJ	U	fit to be eaten; edible	While foraging in the forest, one must remember certain berries are not esculent.	5.7			
macilent	ADJ	U	lean; thin; emaciated	The activist uses pictures of macilent children to demonstrate the problem of world hunger.	5.9			
condign	ADJ	U	well-deserved; fitting; adequate	Most people think that severe punishments are condign for murderers.	5.6			
pinguid	ADJ	U	fat; oily	The salad dressing was so pinguid that I thought I was swallowing oil.	5.6			

priscan	ADJ	U	dealing with or existing in ancient times	The relic is priscan; it has lasted from a remote period.	5.2
hiemal	ADJ	U	of or pertaining to winter, wintry	The hiemal winds swept over the tundra, burying the land in snow.	4.9
thrasonic	ADJ	U	bragging; boastful	After winning he became very thrasonic, describing his exploits to anyone who would listen.	6.0
canard	N	U	Great personal dishonor or humilition	Surely no one would believe such a ridiculous canard published in that disreputable magazine.	3.9
comity	N	U	An atmosphere of social harmony	There are many group activities that promote comity among the students in contrast with debates and arguing.	6.0
encomium	N	U	A formal expression of praise; a tribute	The president gave an encomium fit for the national hero when they laid him to rest.	5.6
exculpate	N	U	To clear of guilt or blame	The jury wanted to exculpate the man of any wrongdoing.	5.4
evanescence	N	U	To dissipate or disappear like vapor	Joy and sorrow are characterized by evanescence, coming and going over time.	4.3
marplot	N	U	An officious meddler whose interference compromises the success of an undertaking.	The plan never got off the ground because of a vicious marplot who took all the funding.	4.9

probity	N	U	Complete and confirmed integrity; uprightness	She strongly believed in the probity of the firm and never questioned their judgment.	5.5
rapparee	N	U	A bandit or robber	The police caught the rapparee as he tried to escape after holding up the bank.	6.0
sagacity	N	U	The quality of being sound in judgement; wisdom	It will require sagacity to choose the appropriate course of action among so many ideas.	5.1
solecism	N	U	A violation of etiquette	Children who are not taught manners often commit solecisms and act improperly in formal situations.	5.5

Bibliography

Armstrong, B. C., Beekhuizen, B., Rice, C., Milic, S., & Stevenson, S. (2018, July). *How are ambiguous word meanings learned, represented, and processed? Insights from computational modeling*. Paper presented at the Joint Meeting of the Canadian Society for Brain, Behavior, and Cognitive Science and the Experimental Psychology Society, St. John's, Newfoundland.

Armstrong, B. C., & Plaut, D. C. (2016). Disparate semantic ambiguity effects from semantic processing dynamics rather than qualitative task differences. *Language, Cognition and Neuroscience*, *31*(7), 940–966. https://doi.org/10.1080/23273798.2016.1171366

Azuma, T., & van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*, 484–504.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Badgett, B. (2003). *Vocabulary acquisition and the generation effect* (Unpublished Master's thesis). University of Nevada, Las Vegas.

Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly*, *43*(1), 79–103. https://doi.org/10.1002/j.1545-7249.2009.tb00228.x

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). https://doi.org/10.1016/j.jml.2012.11.001

Basi, R. K., Thomas, M. H., & Wang, A. Y. (1997). Bilingual generation effect: variations in participant bilingual type and list type. *The Journal of General Psychology*, *124*(2), 216–222. https://doi.org/10.1080/00221309709595519

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4 (R Package version 1.1-7).

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford Publications.

Beck, I. L., McKeown, M. G., & McCaslin, E. S. (1983). Vocabulary development: All contexts are not created equal. *The Elementary School Journal*, *83*(3), 177–181. https://doi.org/10.1086/461307

Beck, I. L., Perfetti, C. A., & McKeown, M. G. (1982). Effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, *74*(4), 506–521. https://doi.org/10.1037/0022-0663.74.4.506

Begg, I., Vinski, E., Frankovich, L., & Holgate, B. (1991). Generating makes words memorable, but so does effective reading. *Memory & Cognition*, *19*(5), 487–497. https://doi.org/10.3758/BF03199571

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Brain Research. Cognitive Brain Research*, *24*(1), 57–65. https://doi.org/10.1016/j.cogbrainres.2004.12.006

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

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. https://doi.org/10.1080/01638530701792826

Bornkessel, I. D., Fiebach, C. J., & Friederici, A. D. (2004). On the cost of syntactic ambiguity in human language comprehension: an individual differences approach. *Brain Research. Cognitive Brain Research*, *21*, 11–21. https://doi.org/10.1016/j.cogbrainres.2004.05.007

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 63–85. https://doi.org/10.1037//0278-7393.22.1.63

Christensen, R. H. B. (2019). Ordinal - regression models for ordinal data. (2019.4-25). The R Project for Statistical Computing.

Coltheart, M., & Ulicheva, A. (2018). Why is nonword reading so variable in adult skilled readers? *PeerJ*, *6*, e4879. https://doi.org/10.7717/peerj.4879

Commission on Language Learning. (2016). *The State of Languages in the U.S.: A Statistical Portrait*. Cambridge, MA: American Academy of Arts & Sciences Comission on Language Learning.

Coomber, J. E., Ramstad, D. A., & Sheets, D. R. (1986). Elaboration in vocabulary learning: A comparison of three rehearsal methods. *Research in the Teaching of English*, *20*(3), 281–293.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. https://doi.org/10.3758/BF03214546

de Groot, A. M. B., & van Hell, J. G. (2005). The learning of foreign language vocabulary. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 9–29). Oxford University Press.

Degani, T., Prior, A., Eddington, C. M., Arêas da Luz Fontes, A. B., & Tokowicz, N. (2016). Determinants of translation ambiguity: A within and cross-language comparison. *Linguistic Approaches to Bilingualism*, *6*(3), 290–307. https://doi.org/10.1075/lab.14013.deg

Degani, T., Tseng, A., & Tokowicz, N. (2014). Together or apart: Learning of translation-ambiguous words. *Bilingualism: Language and Cognition*, *17*(04), 749–765. https://doi.org/10.1017/S1366728913000837

DeLosh, E. L., & McDaniel, M. A. (1996). The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1136–1146. https://doi.org/10.1037/0278-7393.22.5.1136

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*(2), 162–170. https://doi.org/10.1037/0022-0663.79.2.162

Doherty, M. J. (2004). Children's difficulty in learning homonyms. *Journal of Child Language*, *31*(1), 203–214. https://doi.org/10.1017/S030500090300583X

Drum, P. A., & Konopak, B. C. (1987). Learning word meanings from written context. In M. McKeown & M. Curtis (Eds.), *The nature of vocabulary development* (pp. 73–87). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Durso, F. T., & Shore, W. J. (1991). Partial knowledge of word meanings. *Journal of Experimental Psychology: General*, *120*(2), 190–202. https://doi.org/10.1037/0096-3445.120.2.190

Eddington, C. M. (2015). *Effects of within- and cross-language semantic ambiguity on learning and processing* (Upublished doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Eddington, C. M., Martin, K. I., & Tokowicz, N. (2012). *How meaning-based strategies and the generation effect influence German vocabulary learning*. Presented at the UIC Bilforum, Chicago, IL.

Eddington, C. M., & Tokowicz, N. (2013). Examining English–German translation ambiguity using primed translation recognition. *Bilingualism: Language and Cognition*, *16*(02), 442–457. https://doi.org/10.1017/S1366728912000387

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review*, *22*(1), 13–37. https://doi.org/10.3758/s13423-014-0665-7

Ekves, Z. (2014). *Effects of word concreteness on translation-ambiguous word learning* (Unpublished undergraduate thesis). University of Pittsburgh, Pittsburgh, PA.

Elgort, I., Perfetti, C. A., Rickles, B. & Stafura, J. Z. (2015) Contextual learning of L2 word meanings: second language proficiency modulates behavioural and event-related brain potential (ERP) indicators of learning, *Language, Cognition and Neuroscience*, 30(5), 506-528, doi: 10.1080/23273798.2014.942673

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. Language Learning, 64, 365–414. doi:10.1111/lang.12052

Elley, W. B. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, *24*(2), 174. https://doi.org/10.2307/747863

Fang, X., & Perfetti, C. A. (2017). Perturbation of old knowledge precedes integration of new knowledge. *Neuropsychologia*, *99*, 270–278. https://doi.org/10.1016/j.neuropsychologia.2017.03.015

Fang, X., & Perfetti, C. A. (2019). Learning new meanings for known words: Perturbation of original meanings and retention of new meanings. *Memory & Cognition*, *47*(1), 130–144. https://doi.org/10.3758/s13421-018-0855-z

Ferrel Tekmen, E. A., & Daloğlu, A. (2006). An investigation of incidental vocabulary acquisition in relation to learner proficiency level and word frequency. Foreign Language Annals, 39, 220–243. doi:10.1111/j.1944-9720.2006.tb0 2263.x

Fischer, U. (1994). Learning words from context and dictionaries: An experimental comparison. *Applied Psycholinguistics*, *15*(4), 551–574. https://doi.org/10.1017/S0142716400006901

Fox, J., & Weisberg, S. (2019). An R companion to applied regression (version 3). Thousand Oaks, CA: The R Project for Statistical Computing.

Frank, A. (2011). Diagnosing collinearity in mixed models from lme4. Retrieved November 11, 2019, from https://hlplab.wordpress.com/2011/02/24/diagnosing-collinearity-in-lme4/.

Frishkoff, G. A., Collins-Thompson, K., Perfetti, C. A., & Callan, J. (2008). Measuring incremental changes in word knowledge: experimental validation and implications for learning and assessment. *Behavior Research Methods*, *40*(4), 907–925. https://doi.org/10.3758/BRM.40.4.907

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement (0.84.1). The R Project for Statistical Computing.

Gardiner, J. M., & Rowley, J. M. C. (1984). A generation effect with numbers rather than words. *Memory & Cognition*, *12*(5), 443–445. https://doi.org/10.3758/BF03198305

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Gipe, J. P. (1980). Use of a relevant context helps kids learn new word meanings. *The Reading Teacher*, *33*(4), 398–402.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, *28*(1), 99–108. https://doi.org/10.1037/0012-1649.28.1.99

Gunter, T. C., Wagner, S., & Friederici, A. D. (2003). Working memory and lexical ambiguity resolution as revealed by ERPs: a difficult case for activation theories. *Journal of Cognitive Neuroscience*, *15*(5), 643–657. https://doi.org/10.1162/089892903322307366

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*.

Hino, Y., Lupker, S. J., & Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(4), 686–713. https://doi.org/10.1037//0278-7393.28.4.686

Hino, Y., Lupker, S. J., Sears, C. R., & Ogawa, T. (1998). The effects of polysemy for Japanese katakana words. *Reading and Writing*.

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language*, *55*(2), 247–273. https://doi.org/10.1016/j.jml.2006.04.001

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 484–494. https://doi.org/10.1037/0278-7393.14.3.484

Hulme, R. C., Barsky, D., & Rodd, J. M. (2019). Incidental learning and long-term retention of new word meanings from stories: The effect of number of exposures. *Language Learning*, *69*(1), 18–43. https://doi.org/10.1111/lang.12313

Jager, B., & Cleland, A. A. (2016). Polysemy advantage with abstract but not concrete words. *Journal of Psycholinguistic Research*, *45*(1), 143–156. https://doi.org/10.1007/s10936-014-9337-z

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

Jenkins, J. R., Matlock, B., & Slocum, T. A. (1989). Two Approaches to vocabulary instruction: The teaching of individual word meanings and practice in deriving word meaning from context. *Reading Research Quarterly*, *24*(2), 215. https://doi.org/10.2307/747865

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal*, *21*(4), 767–787. https://doi.org/10.3102/00028312021004767

Johns, E. E., & Swanson, L. G. (1988). The generation effect with nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 180–190. https://doi.org/10.1037/0278-7393.14.1.180

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), 601–609.

Kinoshita, S. (1989). Generation enhances semantic processing? The role of distinctiveness in the generation effect. *Memory & Cognition*, *17*(5), 563–571. https://doi.org/10.3758/BF03197079

Klein, D. E., & Murphy, G. L. (2001). Paper has been my ruin: conceptual relations of polysemous senses. *Journal of Memory and Language*, *47*, 548–570.

Klepousniotou, E. (2002). The processing of lexical ambiguity: homonymy and polysemy in the mental lexicon. *Brain and Language*, *81*(1–3), 205–223. https://doi.org/10.1006/brln.2001.2518

Kroll, J. F., & Tokowicz, N. (2001). The development of conceptual representation for words in a second language. In J. L. Nicol (Ed.), *One mind, two languages: Bilingual language processing* (pp. 49-71). Malden, MA: Blackwell.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Laxén, J., & Lavaur, J.-M. (2010). The role of semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition*, *13*(02), 157. https://doi.org/10.1017/S1366728909990472

Lev-Ari, S., & Keysar, B. (2014). Executive control influences linguistic representations. *Memory & Cognition*, *42*(2), 247–263. https://doi.org/10.3758/s13421-013-0352-3

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: a meta-analysis. *Psychonomic Bulletin & Review*, *21*(4), 861–883. https://doi.org/10.3758/s13423-013-0565-2

Loess, H. B., & Duncan, C. P. (1952). Human discrimination learning with simultaneous and successive presentation of stimuli. *Journal of Experimental Psychology*, *44*(3), 215–221. https://doi.org/10.1037/h0061719

Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning*, *48*(1), 31–69. https://doi.org/10.1111/1467-9922.00032

Lutz, J., Briggs, A., & Cain, K. (2003). An examination of the value of the generation effect for learning new material. *The Journal of General Psychology*, *130*(2), 171–188. https://doi.org/10.1080/00221300309601283

Maciejewski, G., Rodd, J. M., Mon-Williams, M., & Klepousniotou, E. (2019). The cost of learning new meanings for familiar words. *Language, Cognition and Neuroscience*, 1–23. https://doi.org/10.1080/23273798.2019.1642500

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in l2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*(3), 379–413. https://doi.org/10.1017/S0272263112000125

Mazzocco, M. M. (1997). Children's interpretations of homonyms: a developmental study. *Journal of Child Language*, *24*(2), 441–467. https://doi.org/10.1017/s0305000997003103

McElroy, L. A. (1987). The generation effect with homographs: evidence for postgeneration processing. *Memory & Cognition*, *15*(2), 148–153. https://doi.org/10.3758/BF03197026

McElroy, L. A., & Slamecka, N. J. (1982a). Memorial consequences of generating nonwords: Implications for semantic-memory interpretations of the generation effect. *Journal of Verbal Learning and Verbal Behavior*, *21*(3), 249–259. https://doi.org/10.1016/S0022-5371(82)90593-X

McElroy, L. A., & Slamecka, N. J. (1982b). Memorial consequences of generating nonwords: Implications for semantic-memory interpretations of the generation effect - Semantic Scholar.

McFarland, C. E., Warren, L. R., & Crockard, J. (1985). Memory for self-generated stimuli in young and old adults. *Journal of Gerontology*, *40*(2), 205–207.

https://doi.org/10.1093/geronj/40.2.205

McNamara, D. S., & Healy, A. F. (2000). A procedural explanation of the generation effect for simple and difficult multiplication problems and answers. *Journal of Memory and Language*, *43*(4), 652–679. https://doi.org/10.1006/jmla.2000.2720

Michael, E. B., Tokowicz, N., Degani, T., & Smith, C. J. (2011). Individual differences in the ability to resolve translation ambiguity across languages. *Vigo International Journal of Applied Linguistics*, *8*, 79–97.

Miyake, A., Just, M. A., & Carpenter, P. A. (1994). Working memory constraints on the resolution of lexical ambiguity: maintaining multiple interpretations in neutral contexts. *Journal of Memory and Language*, *33*(2), 175–202. https://doi.org/10.1006/jmla.1994.1009

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Mulligan, N. W. (2002). The generation effect: dissociating enhanced item memory and disrupted order memory. *Memory & Cognition*, *30*(6), 850–861. https://doi.org/10.3758/bf03195771

Myers, J. L., & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, *26*(2–3), 131–157. https://doi.org/10.1080/01638539809545042

Nagy, W. E. (1995). *On the role of context in first- and second-language vocabulary learning* (Center for the Study of Reading Technical Report No. 627). Champaign, Ill. : University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Nagy, W. E., Anderson, R. C., & Herman, P. A. (1987). Learning word meanings from context during normal reading. *American Educational Research Journal*, *24*(2), 237–270. https://doi.org/10.3102/00028312024002237

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233. https://doi.org/10.2307/747758

Nairne, J. S., Pusen, C., & Widner, R. L. (1985). Representation in the mental lexicon: Implications for theories of the generation effect. *Memory & Cognition*, *13*(2), 183–191. https://doi.org/10.3758/BF03197011

Nairne, J. S., & Widner, Jr., R. L. (1987). Generation effects with nonwords : the role of test appropriateness. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *13*(1), 164–171.

Nash, H., & Snowling, M. (2006). Teaching new words to children with poor existing vocabulary knowledge: a controlled evaluation of the definition and context methods. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists*, *41*(3), 335–354. https://doi.org/10.1080/13682820600602295

O'Neill, W., Roy, L., & Tremblay, R. (1993). A translation-based generation effect in bilingual recall and recognition. *Memory & Cognition*, *21*(4), 488–495. https://doi.org/10.3758/BF03197180

Ouellette, G. P. (2006). What's meaning got to do with it: The role of vocabulary in word reading

and reading comprehension. Journal of Educational Psychology, 98, 554–566. doi:10.1037/0022-0663.98.3.554

Payne, D. G., Neely, J. H., & Burns, D. J. (1986). The generation effect: further tests of the lexical activation hypothesis. *Memory & Cognition*, *14*(3), 246–252. https://doi.org/10.3758/bf03197700

Perfetti, C., & Hart, L. (2002). The lexical quality hypothesis. In L. Vehoven, C. Elbro, & P. Reitsma (Eds.), *Precursors of functional literacy* (pp. 189–213). Amsterdam/Philadelphia: John Benjamins.

Pesta, B. J., Sanders, R. E., & Murphy, M. D. (1999). A beautiful day in the neighborhood: What factors determine the generation effect for simple multiplication problems? *Memory & Cognition*, *27*(1), 106–115. https://doi.org/10.3758/BF03201217

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115. https://doi.org/10.1037/0033-295X.103.1.56

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science*, *330*(6002), 335. https://doi.org/10.1126/science.1191465

Reichle, E. D., & Perfetti, C. A. (2003). Morphology in word identification: a word-experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading*, *7*(3), 219–237. https://doi.org/10.1207/S1532799XSSR0703_2

Rice, C. A., Ekves, Z., & Tokowicz, N. (2017, November). *Mitigating the translation-ambiguity disadvantage: The effect of presenting multiple translations simultaneously*. Poster presented at the Psychonomic Society Annual Meeting, Vancouver, British Columbia, Canada.

Rice, C. A., & Tokowicz, N. (2019). A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition*. Advance online publication.

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*, 1–17. https://doi.org/10.3758/s13421-019-00934-4

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. https://doi.org/10.1006/jmla.2001.2810

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Schindler, J., Schindler, S., & Reinhard, M.-A. (2019). Effectiveness of self-generation during learning is dependent on individual differences in need for cognition. *Frontline Learning Research*, *7*(2), 23–39. https://doi.org/10.14786/flr.v7i2.407

Shore, W. J., & Durso, F. T. (1990). Partial knowledge in vocabulary acquisition: General constraints and specific detail. *Journal of Educational Psychology*, *82*(2), 315–318. https://doi.org/10.1037/0022-0663.82.2.315

Simon, J. R., & Wolf, J. D. (1963). Choice reaction time as a function of angular stimulus-response correspondence and age. *Ergonomics*, *6*(1), 99–105. https://doi.org/10.1080/00140136308930679

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology. Human Learning and Memory*, *4*, 592–604.

Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displaced rehearsal. *Journal of Memory and Language*, *26*(6), 589–607. https://doi.org/10.1016/0749-596X(87)90104-5

Swanborn, M. S. L., & de Glopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, *69*(3), 261–285. https://doi.org/10.3102/00346543069003261

Takashima, A., Bakker-Marshall, I., van Hell, J. G., McQueen, J. M., & Janzen, G. (2019). Neural correlates of word learning in children. *Developmental Cognitive Neuroscience*, *37*, 100649. https://doi.org/10.1016/j.dcn.2019.100649

Tan, L. H., Spinks, J. A., Feng, C.-M., Siok, W. T., Perfetti, C. A., Xiong, J., … Gao, J.-H. (2003). Neural systems of second language reading are shaped by native language. *Human Brain Mapping*, *18*(3), 158–166. https://doi.org/10.1002/hbm.10089

Tokowicz, & Jarbo, K. J. (2009, November). *The generation effect applied to second language vocabulary learning*. Poster presented at the Fiftieth Annual Meeting of the Psychonomic Society, Boston, MA.

Tokowicz, N. (2014). Translation ambiguity affects language processing, learning, and representation. In M. Ryan T, K. I. Martin, C. M. Eddington, A. Henery, N. Marcos Miguel, A. M. Tseng, … D. Walter (Eds.), *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*. Somerville, MA: Cascadilla Proceedings Project.

Tokowicz, N., & Kroll, J. F. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes*, *22*(5), 727–779. https://doi.org/10.1080/01690960601057068

Tokowicz, N., Michael, E. B., & Kroll, J. F. (2004). The roles of study-abroad experience and working-memory capacity in the types of errors made during translation. *Bilingualism: Language and Cognition*, *7*(3), 255–272. https://doi.org/10.1017/S1366728904001634

Tokowicz, N., Rice, C., & Terrazas-Duarte, G. (2018, June). *Words with multiple translations across languages: Alleviating the learning disadvantage.* Oral presentation presented at the 6th International Workshop on Advanced Learning Sciences: Perspectives on the learner: Cognition, brain, and education, Pittsburgh, PA.

Tseng, A. M., Doppelt, M. C., & Tokowicz, N. (2018). The effects of transliterations, thematic organization, and working memory on adult L2 vocabulary learning. *Journal of the Society of Laparoendoscopic Surgeons*, *1*(1), 141–165. https://doi.org/10.1075/jsls.17018.tse

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading comprehension. *The Quarterly Journal of Experimental Psychology. A,*

Human Experimental Psychology, *49*(1), 51–75. https://doi.org/10.1080/713755607

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*, 33-52.

Wilkinson, K. S., & Houston-Price, C. (2013). Once upon a time, there was a pulchritudinous princess . . .: The role of word definitions and multiple story contexts in children's learning of difficult vocabulary. *Applied Psycholinguistics*, *34*(3), 591–613. https://doi.org/10.1017/S0142716411000889

Yudes, C., Macizo, P., & Bajo, T. (2011). The influence of expertise in simultaneous interpreting on non-verbal executive processes. *Frontiers in Psychology*, *2*, 309. https://doi.org/10.3389/fpsyg.2011.00309