Semantic relatedness in L2 vocabulary learning: Does it really matter?

by

Brody Dingel

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Major: Teaching English as a Second Language/Applied Linguistics

(Corpus and Computational Linguistics)

Program of Study Committee:
Evgeny Chukharev-Hudilainen, Major Professor
Gary Ockey
Charles Nagle

Iowa State University

Ames, Iowa

2020

DEDICATION

To my wife, Rachel, who has been a constant source of support, care, and counsel.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

ACKNOWLEDGMENTS

First and foremost, I would like to thank my committee chair, Dr. Evgeny Chukharev-Hudilainen, for his unmeasurable support, guidance, and friendship throughout the years. He has always challenged me to think critically and approach problems in a "clever" manner, and so much of my ability to reason derives from his mentorship. I am also grateful for his refreshing balance of wit and seriousness: our conversations often sway between hearty laughs and furrowed brows. His mentorship has resulted in some of my most cherished moments at Iowa State University.

Next, I would like to extend my gratitude to my committee members, Dr. Charlie Nagle and Dr. Gary Ockey, for their guidance and support throughout my research. Their generous feedback and critical eyes have been a great source of support in the writing of this thesis.

Además, quisiera agradecerle a la profesora Marta Vessoni-de-Lence por inculcarme una pasión por la lengua española, a la profesora Julia Domínguez por introducirme al mundo de la lingüística, a la profesora Cristina Pardo-Ballester por darme tanta guía y oportunidades como estudiante e investigador, y al resto del departamento de lenguas por cultivar en mí una apreciación profunda por la lengua. También estoy muy agradecido por la ayuda de los alumnos del programa de español que participaron en este estudio y a sus profesores que generosamente ofrecieron tiempo en sus clases para la recaudación de datos.

Finally, I would also like to thank my friends, colleagues, and professors in the English Department and in the Department of World Languages and Cultures for a rewarding experience throughout my years at Iowa State University. It is here that I began my journey as a scholar, and I am thankful for the unending support provided to me.

ABSTRACT

Second language (L2) textbooks often organize new vocabulary in lists of semantically related words under a common superordinate concept, such as food or family members. However, research on this topic has shown mixed results, with some studies suggesting that related lists facilitate learning, and others showing inhibiting effects. Importantly, all studies to date have been carried out in a laboratory or strictly controlled classroom setting where individual differences among students are often controlled for. Given that these differences may result in different learning gains in the authentic classroom environment compared to a controlled setting, the potential effects of semantic relatedness on vocabulary acquisition may similarly manifest differently when students are left to their own devices. This thesis reports on the first empirical study (to the author’s knowledge) to test the effects of semantic relatedness on vocabulary learning in a truly authentic classroom environment. Two hundred and twelve students in beginner- and intermediate-level Spanish classes at Iowa State University were tested on their ability to translate items from one related list and one unrelated list from their course textbooks near the end of their respective units. Data were analyzed using mixed-effects logistic regression models under strict and sensitive scoring protocols.

Results indicated no evidence for a significant difference between scores on related and unrelated lists. Further regression analysis indicated a significant effect of individual lexical items on the learning outcomes, and item analyses suggested that some control over item-level characteristics may be needed to facilitate research even in the authentic classroom environment. Implications for teachers, materials developers, and researchers are discussed.

CHAPTER 1. INTRODUCTION

The well-established and growing field of second language acquisition (SLA) is often critiqued for leaving a gap between research and practice: despite advances in SLA, teachers are often left without firm answers from research to guide their instruction (Ellis, 2010). This gap, however, is not unique to the field of SLA but can also be found in other disciplines, such as engineering and medicine (Ellis, 2010; Long, 2011). For example, a healthcare provider cannot possibly diagnose a patient with absolute certainty, even with years of research done in the field, as results from all past studies will rarely converge perfectly. However, neither can they withhold all suggestions from the patient for fear of a potentially suboptimal diagnosis or treatment. Rather, they must make an educated decision based on both a holistic view of the research in the field and their own practical experience. With both theory and practice as their combined toolbelt, the healthcare provider should be able to offer a well-informed diagnosis and treatment plan for the patient.

Language teaching is hardly different. The field of SLA rarely comes to a full consensus on a given topic, and yet instructors are expected to base their pedagogical decisions on research findings. However, three issues arise with the translation of these findings into practice. First, it has been argued that “research knowledge per se does not articulate easily and cogently into classroom practice” (Freeman & Johnson, 1998, p. 411), suggesting that deriving pedagogical implications from research findings is an important task that needs to be carried out deliberately . Second, many teachers either lack the technical knowledge and scientific training to understand research articles, or simply do not have time to read them. Third, any given implication from findings may or may not be applicable to a given teacher’s context (Ellis, 2010).

In an attempt to address the first issue, researchers often include pedagogical implications sections at the end of their articles (Ellis, 2010). The utility of such sections has been disputed, as can be seen in an exchange between researchers in *TESOL Quarterly* 41.2 (2007). This exchange was initiated by Han's (2007) critique of pedagogical implications sections in *TESOL Quarterly* (*TQ*) that "ostentatiously link the research to practice" (p. 387). Han points out that not all research in SLA is related to language teaching and suggests that researchers take more care when considering implications for pedagogy rather than assuming that their findings must certainly translate into the classroom. Han is met with both support and criticism from other authors. Cargill (2007) and Magnan (2007) agree with Han and provide concrete suggestions to their fellow authors and editors to guard against undue implications for teachers (e.g., detailing the setting of the study, linguistic hedging, and proper interpretations of statistical significance); meanwhile, Belcher (2007) and Chapelle (2007) are more hesitant to accept Han's call and argue for the importance of implications sections in *TQ* articles given practitioners' need for empirically based suggestions for pedagogy.

Regardless, including an implications section does not resolve the second issue: that teachers may not have the time to read and the capacity to understand research articles. Neither does it solve the third issue, which Ellis (2010) describes as the following:

> All research – including research based on an experimental design and the use of inferential statistics intended to ensure generalizability – is necessarily conducted in a specific research site (not always a classroom), which may or may not share characteristics with the instructional site in which an individual teacher operates. It does not follow then that the implications drawn from a single study are of any relevance to the individual teacher. (p. 186)

In an attempt to solve all three issues, Ellis offers concise principles for teaching SLA concepts to teachers, in a similar vein as the ten principles for language pedagogy that he offered in Ellis (2005). Similarly, Long (2011) offered advice for language teachers with his methodological principles for language pedagogy, and Folse (2004) broke down a series of myths about vocabulary learning to help teachers understand what research has found about the topic. These relatively teacher-friendly, yet fully research-backed, summaries provided by Ellis, Long, and Folse work toward a solution for all three issues: (1) they deliberately derive pedagogical implications from research findings, (2) they are concise and easy to read, and (3) they apply to the language classroom generally.

Given this proposed solution to the issue of bridging the gap between research and practice, the discussion now turns to one of Folse's (2004) myths in particular, myth #3: "vocabulary should be presented in semantic sets" (p. 4). This myth stems from a decades-long search for an answer to the question of whether presenting vocabulary in semantically related sets facilitates or hinders learning, with many (e.g., Folse, 2004; Papathanasiou, 2009; Tinkham, 1993, 1997; Waring, 1997) being convinced of the negative effects of related sets. However, as will be discussed in chapter 2, one critical issue possibly undermines this claim: to the best of the author's knowledge, no study to date has investigated this issue in the authentic language classroom. As will be discussed below, laboratory studies and tightly controlled classroom studies (i.e., those that control for characteristics that are allowed to vary in the business-as-usual routine, such as length and quantity of study sessions, means of studying, etc.) do not generalize to the authentic classroom.

This thesis first seeks to contribute novel findings to the field on the use of semantically related lists and their effects on L2 vocabulary acquisition in the authentic classroom environment. Secondly, and perhaps more importantly, it seeks to contribute to the discussion on a potentially major issue with the question of bridging the worlds of researchers and teachers: What are practitioners and materials developers to do when differences between the worlds of research and practice result in a lack of generalizability of research findings to the classroom?

To investigate the issue of semantic relatedness in the authentic classroom environment, this thesis will first discuss the relevant literature in chapter 2, which will be followed by an introduction of the research questions. In chapter 3, the methodology of the present study will be outlined, and the results will be presented in chapter 4. Finally, chapter 5 will discuss the results and provide conclusions and implications for teachers, materials developers, and researchers.

CHAPTER 2. LITERATURE REVIEW

L2 learners need to acquire several thousand vocabulary items in order to function in their L2 (Schmitt, 2008), many of which are presented in list format in a textbook. Researchers (e.g., Tinkham, 1993, 1997; Waring, 1997) have noted that one common trend in many textbooks is for these lists to follow a dichotomous presentation scheme: in one type of list, items are all semantically related to one another (i.e., the meaning of one given item from the list will be somehow similar to the meaning of the remaining items; for example, all words might be hyponyms of a single hypernym); in the other type of list, items are not related semantically. For example, in a single, semantically related word list one might find the words *aunt, uncle, mother, father,* etc. (where all words are hyponyms of *family member*). An example of a semantically unrelated list would be one containing the words *wedding, birthday, celebrate, surprise*, etc. (Note that, in this case, the words are still related thematically: they may all be useful in a conversation about holidays or parties. However, this thematic relationship among words is not the same as semantic relatedness.)

It has been further noted that L2 textbooks often favor semantic relatedness (e.g., Erten & Tekin, 2008; Finkbeiner & Nicol, 2003; Folse, 2004; Tinkham, 1993, 1997; Waring, 1997). This means of organization may seem intuitive to materials developers and practitioners since these words “go together.” However, when years of learning vocabulary are at stake, it is important to use evidence rather than intuition to determine whether one type of list is better than the other, and to what extent this difference affects the efficiency of learning new L2 vocabulary.

The large body of prior research on this topic has yielded inconclusive results. Some studies have shown the benefit of related lists (e.g., Hashemi & Gowdasiaei, 2005; Hoshino, 2010), while others indicate the benefit of unrelated lists (e.g., Finkbeiner & Nicol, 2003;

Tinkham, 1993, 1997; Waring, 1997), and yet others show that there is no difference between the two list types (e.g., Ishii, 2013, 2015, 2017). These studies will be discussed in detail below.

Importantly, however, almost all studies cited above are controlled laboratory experiments, that is, they were conducted in settings radically different from naturalistic classrooms. While a few studies have been carried out in the classroom context, these have been far from truly authentic; in other words, the typical classroom flow was interrupted and multiple variables were controlled for, creating an environment similar to the laboratory. The purpose of this thesis, therefore, is to conduct an authentic classroom-based study to investigate the relative benefits of related and unrelated sets of vocabulary items for students of L2 Spanish.

Evidence in support of and against semantic organization comes from both theoretical and empirical research. Therefore, both types of evidence will be considered in turn.

Theoretical Evidence

Theoretical evidence exists both for and against related sets in L2 vocabulary learning. First, the mental lexicon (i.e., the representation of words in the long-term memory of a language speaker) is organized semantically: mental representations of words that are semantically related seem to be interconnected (Meara, 2009; Nation, 2000). This organization has been modeled as “semantic networks” (Meara, 2009) and “semantic fields” (Lehrer, 1974). The semantic organization of mental lexicons has been confirmed in L2 learners, albeit to a lesser extent than in native speakers (Meara, 2009). If a native speaker’s lexicon can be assumed a model for the L2 learner, then the argument is that learning materials should be presented in ways that are congruent to the target mental representations. Further support for related sets is found in schema theory (Stoller & Grabe, 1993), which suggests that related lists can provide an “anchor” of sorts that allows new knowledge to be connected to existing knowledge, thereby providing a means other than rote memorization for learners to hook onto.

On the other hand, interference theory (Baddeley, 1997) suggests that concurrent introduction to multiple similar concepts may make it harder for the mind to distinguish between them, thereby hindering learning. Semantically related words, by definition, capture similar concepts. The author can illustrate this with his own anecdotal experience: when teaching Spanish as an L2, he found his students repeatedly confusing names of family members during their unit on family relations. In a similar vein, the distinctiveness hypothesis (Hunt & Elliot, 1980) posits that dissimilar concepts may promote learning: every item in memory is distinguished from other items by many semantic features, and the distinctiveness of an item directly promotes its retention. This suggests that "increasing the non-similarity of information increases its ease of learning, and as such, vocabulary should be presented in a nonrelated fashion so that the mind is presented with information organized in a way that is conducive for learning" (Wilcox & Medina, 2013, p. 1058).

Yet another perspective is provided by the desirable difficulty framework (Bjork, 1999) which suggests that the difficulties associated with processing related items might, in the long term, benefit the learner: "the act of retrieval is assumed…to be a potent learning event, but the increments in storage strength (and retrieval strength) are assumed to be greater, the more difficult or involved the act of retrieval" (p. 442). In addition, as Nakata and Suzuki (2019) point out, the challenges of learning semantically related words may push students to apply more efforts or engage with the content in ways that an unrelated word list may not necessitate.

In sum, theoretical models provide evidence, on various grounds, both in favor and against learning vocabulary in semantically related lists. We turn now to reviewing empirical research that directly tested the effects of related vs. unrelated vocabulary presentation.

Empirical Evidence

A number of empirical studies have directly compared the effectiveness of related and unrelated vocabulary presentation. In some of the earliest research on this topic, Tinkham (1993, 1997) tested English-speaking adults on their ability to learn English translations of pseudowords in related and unrelated sets. Results in both studies indicated that participants learned unrelated words better than their related counterparts. Similarly, Waring (1997) sought to replicate Tinkham's (1993) study and concluded with similar recommendations for teachers and materials developers to stray from semantically related lists in L2 vocabulary teaching. However, all three studies took place in the laboratory and were tightly controlled. Tinkham (1997) notes the following point in discussing the limitation of his study:

> Also calling for further research is the limited generalizability of the current research: limited generalizability to an expanded stimulus base (more word sets within a particular condition); limited generalizability to evaluation criteria (long-term rather than short-term evaluation); and limited generalizability to other instructional contexts (context-based rather than rote-based learning). (p. 161)

In a similar vein, Waring (1997) provides this word of caution in the interpretation of his results:

> The experimental design of these studies had its problems and was tightly controlled to benefit the researcher, not the learner, and thus it dilutes the real-world application of the results found. While there are benefits to doing tightly controlled studies, we should be aware that the more tightly controlled it is, there is a possibility that the results it generates might not fully apply to the dynamic classroom. (p. 271)

It is clear, therefore, that researchers have long been aware of the potential difficulties that may arise when data obtained in the laboratory setting are used to make recommendations for the authentic classroom environment.

Shortly after the above studies, Schneider, Healy, and Bourne (1998) conducted two experiments testing college-age students on learning French-English word pairs on a computer at a rate of two seconds per pair over a series of trials. They conclude that semantic organization "facilitated initial acquisition but either hindered or had no effect on retention" (p. 88). These findings were corroborated in Schneider, Healy, and Bourne (2002), which had a similar experimental design. In another experiment, Finkbeiner and Nicol (2003) tested undergraduate students on their ability to learn pseudoword-English word pairs. Here, participants heard a recording of the L2 word, subsequently saw the word and a corresponding picture for 500 milliseconds, heard a second recording of the word, and finally repeated the word aloud twice. This training phase was followed by a recognition task and finally translation tasks. Once again, results indicated a hindering effect of related lists on learning. In all of these studies, however, the rapid pace of learning makes the findings not immediately generalizable to authentic classroom settings.

In another laboratory study, Wilcox and Medina (2013) tested whether phonological relatedness (alongside semantic relatedness) benefits vocabulary learning. The authors argue that "grouping vocabulary either randomly or phonologically could better facilitate long-term retention than presenting words exclusively clustered semantically" (p. 1065). Their results are as follows: words that are semantically related but phonologically unrelated are learned significantly worse than (a) semantically unrelated but phonologically related words, (b) words that are related both semantically and phonologically, and (c) words that are not related either semantically or phonologically. However, for phonologically related words, semantic relatedness did not significantly affect learning gains. Therefore, it appears from their findings that semantic relatedness may have a negative impact on acquisition only if words do not share phonological similarity. This would imply that materials developers should control for both semantic relatedness and phonological similarity when creating word lists, an aspect that may arguably be difficult to apply in practice.

One final series of laboratory studies to note explored the relationship between physical relatedness, semantic relatedness, and learning gains. Ishii (2013) sought to determine whether any difference occurs in the learning of physically related words (e.g., denoting long and thin objects), semantically related words, and unrelated words. Results indicated no significant difference between learning gains of related and unrelated words, while physically related words were significantly harder to learn. Ishii concludes that it may not be semantic relatedness that should be controlled for in vocabulary learning, but rather physical relatedness. However, he notes the limitations of small sample size and lack of randomization of materials. These limitations were addressed in Ishii (2015) with a replication study, coming to the same conclusion. These findings were further corroborated in Ishii (2017). Similar to other laboratory studies reviewed above, Ishii notes the limitations of his research with pseudowords and tight time constraints: “[i]f a similar study is conducted in a genuine classroom setting, where students learn words they perceive are important, with abundant time to review the target words, different results might be obtained” (2017, p. 28).

We will now proceed to review prior research that has been conducted in controlled classroom settings. Hashemi and Howdasiaei (2005) tested Iranian EFL students in their normal classrooms on their learning of 100 words presented either in sets of related words or in random order. In their study, students were provided with each word in a sentence context along with its definition, and the authors noted that students in the related condition were able to guess the meanings of words more easily than students in the random (i.e., unrelated) condition. Results indicated that students in the related condition had greater learning gains than students in the unrelated condition, suggesting that semantic organization actually facilitates learning. However, this study notably deviated from the usual classroom practices in controlling the learning strategy, time for learning, and target words.

Erten and Tekin (2008) tested fourth-grade EFL students in their classroom on a picture-matching task with related and unrelated lists. Students spent 40 minutes participating in teacher-led flashcard exercises for each list of 20 words and then took an immediate post-test followed by a delayed post-test one week later. Results showed higher learning gains on the unrelated lists, providing more evidence for the hindering effect of semantic relatedness.

In Papathanasiou (2009), the teacher led students (beginner adults and intermediate children) for ten minutes in the creation of their own flashcards for either semantically related or unrelated words, depending on the condition, saying the words aloud as they went. After this, students spent fifteen minutes practicing retrieval with their flashcards individually, followed by a time of teacher-led group practice with flashcards. Finally, students completed two different exercises for 20 minutes to practice “generation” of the new vocabulary, though these exercises were not elaborated on in her study (p. 317). At the end of the learning phase, students took a receptive translation test over the new items. Results indicated that adults performed significantly worse on the related test than the unrelated test, though the children showed no significant difference between the two list types. Papathanasiou notes the relatively naturalistic setting of this experiment in comparison to previous studies such that results “might apply to natural L2 learners” (2009, p. 319). However, the study lacked any productive assessment (which is common in the language classroom) and the limited amount of time participants were able to study.

Finally, the only study (to the author’s knowledge) that approaches a truly authentic classroom environment is Hoshino (2010), wherein university-level students were given 3-4 days outside of their normal classroom to learn 10- and 20-item word lists, each being classified as synonyms, antonyms, categorical (i.e., semantically related), thematic, or unrelated. Students were allowed to use whatever means they wished to study the materials in preparation for L1 translations (Japanese) of the L2 English words. Tests were administered in class and lasted either two or four minutes (for 10-item and 20-item tests, respectively). Students repeated this process for each of the 15 word lists (two 10-item lists and one 20-item list for each of the five categories). Answers were counted as correct as long as the correct meaning was given, regardless of word class. Results showed that students performed significantly better on tests over categorical word lists than any other list type, and no other comparisons were significant. Hoshino (2010) interpreted this finding as showing, among others, that a list of related items is easier to learn than a list of unrelated items.

It is important, however, to note the following limitations of Hoshino’s (2010) study. First, students were given only two or four minutes for each 10- or 20-item test, a time span that may be arguably short for some students to complete an L2 vocabulary test. The concern here is that some students may have struggled more on a particular test and therefore would have benefited from more time to take the assessment. Perhaps what would have served the study better would have been to allow each testing session a sufficient amount of time for all students to complete the test and indicate as such, as is often done in the L2 classroom. Second, participants in Hoshino (2010) were tasked with learning and testing over 15 discrete word lists, each list spanning a period of 3-4 days for learning and one day for testing. This procedure could arguably have become tedious for learners and caused them to wane in their performance, especially as they approached the end of the experiment. Third, testing participants only on their receptive translation ability may not have been the most effective way of assessing L2 word knowledge. As productive language ability is generally a major goal of L2 learning, testing ought to reflect this construct. Finally, word lists in Hoshino (2010) were presumably imposed upon the classroom curriculum, thereby further distancing the study from a purely authentic context.

In sum, the results of past empirical research seem to be largely inconclusive, with some studies showing positive effects of semantic relatedness, others showing negative effects, and yet others showing no significant difference, along with differential results across age groups. Throughout these experiments, a number of variables have been manipulated in the experimental designs, leading Nakata and Suzuki to critique these methodological differences and their potential influence on “the inconsistent results of previous studies” (2019, p. 290). Specifically, they note the means of vocabulary knowledge assessment, learning stimuli, participant age and proficiency, duration of treatment, use of posttest, and item difficulty as variables often manipulated in such studies. A summary of the above-cited experiments and their methodological differences may be seen in Table 1 and Table 2 below. Table 1 compares each study with regard to general experimental design, while Table 2 compares the same studies with regard to the stimuli used therein.

Table 1. *Summary of Empirical Evidence*

Study	Target L2	Participant age/level	Environment	Participant motivation	Instructional conditions		Testing conditions		List type with higher gains
					Type	Time	Type	Time	
Tinkham (1993)	Pseudowords	Teenagers, adults	Lab	Volunteer	Teacher-led oral translation to L2	Several minutes (?)	Oral translation to L2	Several minutes (?)	Unrelated
Tinkham (1997)	Pseudowords	University	Lab	Course requirement	Teacher-led oral/written translation (both ways)	Several minutes (?)	Oral/written translation (both ways)	Several minutes (?)	Unrelated
Waring (1997)	Pseudowords	Adults	Lab	Volunteer	Teacher-led oral translation to L2	40+ min	Oral translation to L2	N/A	Unrelated
Schneider et al. (1998, 2002)	French	University	Lab	Course credit	Computer-led word pair memorization	2 sec/pair	Translation to L1 (1998), both ways (2002)	N/A	Either hinders or has no effect
Finkbeiner and Nicol (2003)	Pseudowords	Undergraduate students	Lab	Course credit	Computer-led picture matching with audio	2 45-min sessions	Oral translation (both ways)	N/A	Unrelated
Hashemi and Howdasiaei (2005)	English	Adults	Class	N/A	Guess meaning from sentence context	4 45-min sessions	Vocabulary Knowledge Scale	2 hours	Related
Erten and Tekin (2008)	English	4th grade	Class	Part of class	Teacher-led flashcards and picture matching	8 40-min sessions	Picture matching	No limit	Unrelated
Papathanasiou (2009)	English	Intermediate children, novice adults	Class	Certification (adults), N/A (children)	Teacher-led flashcards and oral translations	6 45-min sessions	Written translation to L1	45 minutes	Unrelated (adults), No difference (children)
Hoshino (2010)	English	University	Class	N/A	Autonomous study of word pairs	3-4 days	Written translation to L1	2-4 minutes	Related
Wilcox and Medina (2013)	Spanish	University	Lab	N/A	Computer-led translation to L2	20 min	Translation to L2	20 minutes	Unrelated

Table 1 Continued

Study	Target L2	Participant age/level	Environment	Participant motivation	Instructional conditions		Testing conditions		List type with higher gains
					Type	Time	Type	Time	
Ishii (2013, 2015)	Pseudowords	University	Lab	N/A	Computer-led word pair memorization	45 sec/list	Translation to L1	N/A	No difference for semantically related/unrelated; physically related worst
Ishii (2017)	Pseudowords	University	Lab	N/A	Computer-led word pair memorization	40 sec/list	Translation to L1	N/A	No difference

Table 2. *Summary of Stimuli in Empirical Evidence*

Study	List categories	No. items / category	Lexical characteristics	Concreteness of items
Tinkham (1993)	Clothes, fruits, 2 unrelated	3 or 6	2 syllables, varying stress, varying vowel/consonant combinations	Concrete
Tinkham (1997)	Semantic (dishes, clothes, metals, fruits), unrelated, thematic (beach, library, frogs, caves), unassociated	3 or 6	In top 5k frequency, phonological variation, pseudowords similar to 1993 study	Semantic/unrelated: concrete Thematic/unassociated: mixed
Waring (1997)	Clothes, fruits, and 2 unrelated	3 or 6	2 syllables, varying stress, varying vowel/consonant combinations	Concrete
Schneider et al. (1998, 2002)	Body parts, vehicles, silverware, foods, clothes, and 5 unrelated	5	N/A	Concrete
Finkbeiner and Nicol (2003)	Animals, kitchen utensils, furniture, body parts (mixed for unrelated)	8	1-2 syllables, followed English phonotactics	Concrete
Hashemi and Howdasiaei (2005)	13 categories (unspecified)	~7	Above Level 4 in difficulty according to JACET word list	N/A
Erten and Tekin (2008)	Animals, foods, and two unrelated	20	~4 letters, ~1.5 syllables	Related: all concrete Unrelated: mostly concrete
Papathanasiou (2009)	Food, nature, crime, synonyms, antonyms, homonyms, and 6 unrelated	10	N/A	Some concrete
Hoshino (2010)	Synonym, antonym, categorical, thematic, unrelated	10 or 20	N/A	N/A
Wilcox and Medina (2013)	[+S–P] tools [–S–P] unrelated [–S+P] initial *t* [+S+P] torment, initial *m*	5	~3 syllables, controlled initial consonant	Mixed

Table 2 Continued

Study	List categories	No. items / category	Lexical characteristics	Concreteness of items
Ishii (2013, 2015)	Semantically related: animals, vegetables, utensils Physically related: round, long and thin, rectangular Unrelated: 3 lists	6	Pseudowords generated using software that conforms to English spelling rules	Concrete
Ishii (2017)	Personality traits, feelings, talking, crime, and 4 unrelated lists	5	Pseudowords generated using software that conforms to English spelling rules; length and phonological pattern controlled	Abstract

Notably, many variables that are generally controlled in previous research are related to learners' individual differences, such as motivation, learning strategy, and anxiety, among others (Ellis, 2015). These three in particular seem particularly relevant to this discussion, because they seem likely to have some variance in the laboratory setting that may differ greatly from how they would generally manifest in the classroom setting.

First, motivation has been termed "a critical determinant of success in language learning" in general (Tseng & Schmitt, 2008, p. 358), and therefore it can be expected to play a role in vocabulary acquisition specifically. While much theoretical work has been done on the topic of motivation in language learning, one model developed by Noels, Pelletier, Clément, and Vallerand (2000) and based on self-determination theory (Deci & Ryan, 1985) lends the concepts of intrinsic and extrinsic motivation which seem relevant for this discussion. In Noels et al., intrinsic motivation is defined as "motivation to engage in an activity because that activity is enjoyable and satisfying to do", while "extrinsically motivated behaviors are those actions carried out to achieve some instrumental end, such as earning a reward or avoiding a punishment" (p. 61). While both types of motivation may be broken down into multiple subtypes (Noels et al., 2000), any of which might be found in the laboratory or classroom settings, these overarching concepts of intrinsic and extrinsic motivation will serve the present purposes.

These types of motivation can manifest differently in the classroom versus laboratory setting, both in terms of quantity and quality. In the laboratory, the participant might be motivated by an extra credit opportunity or simply by a requirement to participate, where their performance in the experiment may not have any impact on their course grade, thereby providing a form of purely extrinsic motivation for the student. Meanwhile, a student in the classroom might be motivated by a number of factors, both intrinsic, such as if learning the language is fun or exciting for the student, and extrinsic, such as a desire to earn a high grade or the need to learn the language for career purposes. In addition to this variance in the quality of motivation, the quantity could also vary considerably between the two settings: while a requirement to participate in an experiment may generate little motivation to perform well, an upcoming exam may drive the student's performance higher.

Second, participant learning strategies may also differ in how they manifest in the laboratory versus in the classroom. Learning strategies have been defined as "behaviors or actions which learners use to make language learning more successful, self-directed, and enjoyable" (Oxford, 1989, p. 235). Importantly, there are many strategies available for learners, many of which can be classified based on O'Malley and Chamot's (1990) typology of metacognitive strategies (e.g., selective attention), cognitive strategies (e.g., inferencing), and social/affective strategies (e.g., asking questions).

In the laboratory setting, learning strategies permitted are often controlled, while real-world students have freedom to learn how they prefer in the classroom setting and certainly outside the classroom. For example, a laboratory study might require participants to use double-sided flashcards or repeat words aloud for learning purposes, regardless of whether the participant would choose those particular strategies or not. In these cases, participants may not have a chance to learn the material as thoroughly as they might on their own.

Third, the laboratory environment and the classroom environment may trigger different levels of anxiety for each participant that may have an effect on their learning. Indeed, language anxiety, stemming from emotional responses as a result of experiences in a particular learning environment, has been called "one of the key affective factors that has been shown to impact on L2 learning", and therefore it is critical to take into account (Ellis, 2015, p. 55).

In the laboratory setting, for example, a participant may feel anxiety from having little experience being part of an experiment and may perform differently than they would normally in the classroom, where they are accustomed to the learning environment. Meanwhile, it may be the case that another student experiences large levels of anxiety from the need to participate in class, while they may not mind participating in an experiment. Differential experiences of this sort are discussed in Horwitz (2001), concluding that “in almost all cases, any task that was judged ‘comfortable’ by some learners was also judged ‘stressful’ by others” (p. 118). This is important because researchers generally agree that “high levels of anxiety impede learning” (Ellis, 2015, p. 56). If students experience different levels of anxiety in the laboratory than they would in the classroom, then once again, inference of experimental results to the classroom setting may not be as valid as one might hope.

In sum, due to the controlled nature of the laboratory setting, it is possible that a participant would perform differently when learning vocabulary compared to the classroom environment. In addition to that, students will exhibit individual differences within the classroom environment itself. For example, in terms of motivation, some students may be driven strongly by a desire for a high grade, while others would be satisfied with average performance. The same principle holds for both anxiety and learning strategies. Therefore, while students will exhibit their individual differences in the authentic language-learning setting, such differences might be removed by experimental controls in the laboratory. This suggests that empirical research may need to be done in authentic classrooms in order to confidently generalize results to other real-life contexts.

The Present Study

To summarize the discussion thus far, research in the field of L2 vocabulary learning has not yet provided a conclusive answer regarding whether semantically related word lists promote, hinder, or have no effect on learning gains as compared to unrelated word lists, nor the extent to which any positive or negative impact may exist. Furthermore, it is clear that the learning process may be radically different in the language classroom than in the laboratory, thereby raising concerns regarding whether findings from strictly controlled laboratory studies may be sufficiently generalizable to the real-world classroom settings. To date, no truly authentic classroom study on the learning of L2 vocabulary in related and unrelated lists has been carried out (to the author's knowledge). The present study, therefore, seeks to fill this gap. This is accomplished by investigating the learning of L2 Spanish vocabulary that is presented naturally (i.e., in existing real-world textbooks) in related and unrelated lists.

The research question guiding this study is as follows:

RQ. Do students perform better on quizzes of semantically related or semantically unrelated vocabulary lists learned in the authentic classroom environment?

CHAPTER 3. METHODS

This chapter will provide a description of the participants of the present study, the materials used, and the procedure carried out both prior to and on the day of data collection, as well as the procedure used for scoring the data. An overview of the data analysis will also be provided.

Participants

Two hundred and twelve participants were recruited from one upper-elementary and two intermediate Spanish classes (Spanish 102, 201, and 202, respectively) at Iowa State University during the fall semester of 2018. The three classes consisted of a total of eight sections and were taught by a total of four different instructors, as shown in Table 3. The author of this thesis was not one of the instructors. Students were not compensated for participating in this study as they were required, as part of their course curriculum, to take vocabulary quizzes developed for the present study. However, students were permitted to opt out of the study (and have their data dropped) by checking a box at the end of each quiz. Forty-eight participants' data were discarded due to having chosen to opt out or being contaminated, and these were spread out across the classes. The final sample consisted of 164 students, with an average of 20.5 students per section.

Table 3. *Number of Participants*

Variable	Spanish 102		Spanish 201					Spanish 202
Section	1	2	1	2	3	4	5	1
Instructor	A		A	B			C	D
No. students	19	15	19	18	24	21	24	24

No further data regarding the demographics of the participants were collected through the quizzes in order to maintain the typical classroom atmosphere. However, these courses are most often taken to fulfill a general education requirement. Therefore, overall university enrollment statistics can be used to approximate the demographic characteristics of this sample. In the semester that data were collected, the undergraduate student population at Iowa State University had 57% male students and 5% international students (Iowa State University, 2019a). Additionally, the university's Office of Admissions reported that less than 7% of undergraduate students were 25 years of age or older (Iowa State University, 2019b). Given these statistics, it is safe to assume a relatively equal mix of male and female students with a majority of native English speakers and students in their early 20s in the present sample. All participants, regardless of their L1, were assumed to be proficient in English because all non-native speakers of English must fulfill the university's English proficiency requirement in order to enroll.

Materials

To maintain a fully authentic classroom environment, it was important to only use materials that were normally a part of the class. For this reason, one related list and one unrelated list were identified from each of the two textbooks used across the three courses, as shown in Table 4. The two lists for each course were selected based on the extent to which they intuitively seemed to differ in semantic relatedness among the items in each list; such intuition-based classification is common in the literature (e.g., Finkbeiner & Nicol, 2003; Tinkham, 1993; Waring, 1997). For example, chapter 10 of the textbook used in Spanish 102 (Blanco, 2016) focused on the names of human body parts, which were clearly semantically related, while the lexical items in chapter 9 centered around a party theme but did not all share common semantic features, and therefore were considered unrelated. The lists for the lower-intermediate class (Pérez-Gironés & Adán-Lifante, 2014) were similar: the related list consisted of words for family members, while the unrelated list centered around a theme of pastimes. The lists that were selected for the upper-intermediate course (Pérez-Gironés & Adán-Lifante, 2014), however, did not follow such an obvious trend. The items in the related list did not share a common hypernym, yet their meanings were all centered around the environment (e.g., Earth, river, forest, species). Meanwhile, the unrelated list contained abstract items that were clearly not semantically related. All the selected lists were independently rated by a faculty member in applied linguistics whose designations of "related" and "unrelated" matched the author's in all instances.

Table 4. *Materials Overview*

	Related lists		Unrelated lists	
Course and textbook	Chapter	No. items	Chapter	No. items
102 (Blanco, 2016)	10	19	9	19
201 (Pérez-Gironés & Adán-Lifante, 2014)	3	16	6	21
202 (Pérez-Gironés & Adán-Lifante, 2014)	8	26	12	16

All textbooks presented new Spanish vocabulary on the left and their English translations on the right, so from these items a quiz was created for each word list to test participants' ability to translate from English to Spanish. Translation was determined to be the best format for this study to closely match tasks that were used in previous studies in the field where translation was highly used (e.g., Schneider et al., 2002, 1998; Wilcox & Medina, 2013). In addition, translation exercises were common in the normal curricular activities (both in the classroom and in the online platform where students were introduced to new material and completed assignments to practice the vocabulary and grammar they had learned), and it provided for straightforward scoring of quizzes. Forward (L1 to L2) translation was chosen for two reasons. First, production of the target language was one of the primary objectives of each of the three courses, as demonstrated by the classroom assessments (such as chapter exams consisting primarily of L2 production). Second, forward translation is a more challenging task (Schneider et al., 2002) and therefore might provide for better discrimination.

The items and their English translations appeared on the quizzes exactly as they appeared in the participants' textbooks. However, some items appeared in the textbooks under a heading of "cognates" or "review" and did not have accompanying translations; these items were excluded from the quizzes so that participants would only be tested on those items that they could be assumed to have little or no prior knowledge of. In addition, the items *nieto/a* ("grandson/granddaughter") and *bisnieto/a* ("great-grandson/great-granddaughter") were both incorrectly translated as "grandson/granddaughter" in one textbook, and therefore both items were removed from the lower-intermediate related list. Finally, all items were randomized for each course; in other words, all students of a given course received the same quiz that had been randomized relative to the textbook order. All quizzes can be found in APPENDIX A. .

Procedure

Students were initially introduced to the vocabulary items between 9 and 22 days prior to the day of data collection. This discrepancy in time of exposure to the material is due to the different lengths of the chapters covered in each course: some chapters are simply covered more rapidly than others. Importantly, each chapter is concluded with an online quiz over the new vocabulary and grammar learned, and the day of data collection was scheduled as close to this usual end-of-chapter assessment as possible. Students were informed by their instructor of the exact nature of the quiz (i.e., forward translation over a particular vocabulary list) at least two days ahead of time. Therefore, despite the variation in the duration of learners' prior exposure to vocabulary across different quizzes, the schedule of data collection accurately reflected the authentic course schedule.

Pre-tests were not part of the present study design for two reasons. First, students were assumed to have little or no prior knowledge of the vocabulary items given that they were presented as new content in their course. Second, pre-tests are not normally a part of the business-as-usual classroom instruction and would therefore diminish the authenticity of the environment.

Students received no guidance on how or to what extent they should study the vocabulary items for each quiz. Each student, therefore, was expected to utilize different methods of studying and to spend different amounts of time in preparation for the quiz. Much unlike the laboratory setting, this variable was intentionally left uncontrolled in order to maintain the authenticity of the learning environment.

Paper copies of each quiz were given to the instructor prior to the start of each class and were administered by the instructor during the first few minutes of class time. The researcher, therefore, had no direct contact with the participants at any point during the study. Before starting the quiz, the instructor reminded the students that spelling and diacritics did count in assessing their responses, as was also the case in their online learning platform, and that the quizzes were closed-book (i.e., to be taken without the assistance of books or notes), as was common practice in their classroom assessments. Students were not informed of the exact purpose of the study, nor whether a list was deemed related or unrelated, but they were allowed the opportunity to opt out of the study by checking a box at the end of the quiz. Students were allowed as much time as needed to take the quiz, though instructors reported that the quizzes generally took no more than ten to fifteen minutes. Instructors then returned the quizzes to the researcher after class for scoring.

Scoring

Quizzes were scored by the researcher according to a scale of 0-2, after which they were returned to the instructor to provide feedback to their students:

2: for items that had been perfectly translated, including spelling, diacritics, and part of speech, though omission of an article was not considered erroneous;

1: for items that had been translated mostly correctly but with a minor mistake (e.g., incorrect grammatical gender, minor spelling or diacritic mistake, wrong part of speech, etc.);

0: for those items that were translated incorrectly, contained a major spelling mistake such that the meaning was changed, or no response was provided.

In order to ascertain the reliability of the researcher's scoring, a graduate student in applied linguistics was asked to independently score a random sample of 10% of all quizzes. The second rater was trained using one graded quiz from each list and was provided instructions for scoring similar to those reported above. The interval metric was deemed appropriate for quantifying agreement between the two raters for two reasons: (1) the scale of 0-2 included three points, and each point was given a different meaning, and (2) the distances between 0-1 and between 1-2 were assumed to be equal. Krippendorff's α was used to measure inter-rater reliability; α_{interval} = .97 suggested that the scoring was highly reliable.

Data Analysis

For each participant and each quiz, a quiz score was calculated as the sum of scores (on the scale 0-2) assigned to each item on that quiz. A diagnostic histogram was plotted to assess the distribution of quiz scores. The distribution of quiz scores looked sufficiently close to the normal distribution, so the assumption of normality was met (see Figure 1). However, one outlier was identified in a boxplot analysis plotting total quiz scores: one participant in Spanish 202 had scored a 100% on the “unrelated” quiz and just a 27% on the “related” quiz (see Figure 2).

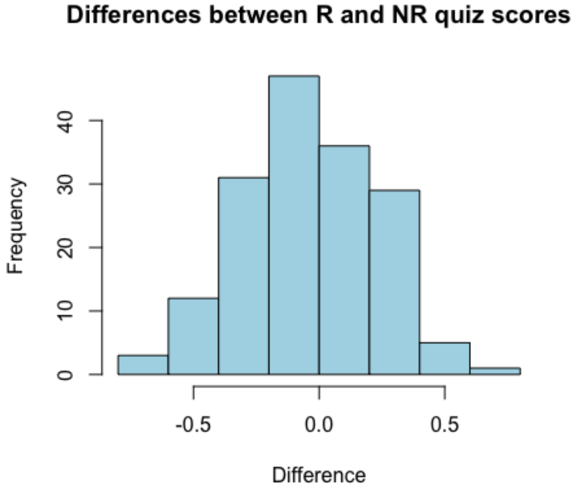

Figure 1. *Histogram of Quiz Scores*

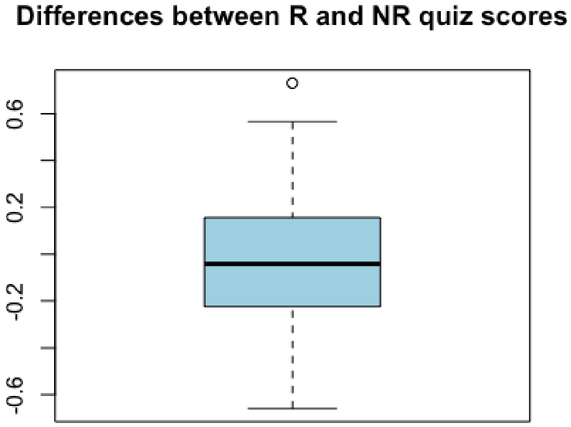

Figure 2. *Boxplot of Quiz Scores*

To get an initial look at the differences between the related and unrelated lists, a paired samples *t*-test was conducted on the total scores to determine whether participants tended to score higher on one list type than the other. Then, for a more detailed analysis, logistic mixed-effects regression (MER) models were fit to predict the probability of a correct response occurring for each item to provide a more detailed model of the data. Unlike the *t*-test, MER models were run on individual items rather than aggregated by quiz. Because logistic MER only permits binary outcome variables, the data were conflated according to two scoring methods: "strict" and "sensitive" (Nakata, 2015). Under the "strict" scoring method, all item-level scores of 1 were converted to 0, and the scores of 2 were left intact. Under the "sensitive" scoring method, all item-level scores of 1 were converted to 2, and the scores of 0 were left intact. These two scoring methods enabled analysis of the data under two simulated conditions: (1) where the instructor grades strictly and awards the point only for a perfectly correct answer, and (2) where the instructor grades leniently and awards the point even if a minor mistake may be present. In addition, such scoring protocols mirror practices established in the language-learning classroom, where the online learning platform automatically grades answers strictly, while the instructor may be more lenient in their grading. Lenience in grading with regard to vocabulary is typical for Spanish language instructors at Iowa State University, as the courses take a communicative approach to language learning and place less emphasis on accuracy of grammar and vocabulary than communicative ability. This is reflected in the rubrics used for regular assessment in the courses, where exams are graded based on three categories: (1) communicative competence for 50% of the grade, (2) language use (i.e., accuracy of vocabulary, grammar, etc.) for 25%, and (3) content for 25%.

The data were formatted as a data frame where each row represented one participant's score on one item, and every participant received a unique identifier. Under each scoring method, then, a series of nested logistic MER models were fit to the data to predict item-level scores. All models included the random intercepts for Participant, Section, Instructor, and Item. First, an intercept-only model was fit to the data (model M_0). Then, ListType (a factor with two levels: “related” and “unrelated”) was added as a fixed effect to the model, yielding model M_1. Then, a fixed effect for Course (a factor with three levels: “102”, “201”, and “202”, corresponding to the three Spanish courses that took part in the study) was added to the model, yielding model M_2. Finally, an interaction between ListType and Course was added, yielding the full model, M_3. Gains in goodness of fit of successive models were evaluated by a likelihood ratio test. All statistical tests were run using R (R Core Team, 2019) (see https://github.com/brodyd795/btdingel-thesis for R code).

CHAPTER 4. RESULTS

This chapter presents the results of the study. It presents descriptive statistics, the results from the paired samples *t*-test, and the significance levels from the logistic MERs. Finally, post-hoc MER analyses and item-level analyses are presented.

Descriptive Statistics and *t*-test Results

Descriptive statistics showed that participants scored higher on quizzes over related words than quizzes over unrelated words, regardless of course level. Means, SDs, and *t*-test results (with α=0.05) are shown in Table 5. Weak evidence was found for a significant difference between scores on related and unrelated quizzes across all participants, with participants scoring significantly higher overall on related quizzes.

Table 5. *Result of* t*-test*

List type	n	Mean	SD	*t*	df	95% CI	*p*-value
Related	163	0.612	0.278	-2	162	[-0.081, -0.001]	0.04*
Unrelated	163	0.570	0.204				

Note: * denotes significance at α=0.05.

Results from Logistic MER Analyses

Logistic MER models were fit to predict the probability of a correct score on a particular item for an individual participant. No model resulted in significantly better fit than the previous: M_1, $\chi_2(1)=2.28$, $p=0.13$ for strict scoring and $\chi_2(1)=0.14$, $p=0.71$ for sensitive scoring; M_2, $\chi_2(2)=1.34$, $p=0.51$ for strict scoring and $\chi_2(2)=1.09$, $p=0.58$ for sensitive scoring; and M_3, $\chi_2(2)=1.28$, $p=0.53$ for strict scoring and $\chi_2(2)=0.78$, $p=0.68$ for sensitive scoring. Therefore, the MER models yielded no evidence for a significant difference between scores on related and unrelated lists for the full dataset.

Given the difference between the results of the *t*-test and those of the MER models, post-hoc analysis with one new MER model was carried out to investigate the differences in significance levels. It was hypothesized that Item may play a significant role in determining score, thus M_4 was built with the same predictors as M_3 but without the random effect for Item. M_4 resulted in a significantly worse fit to the data than M_3 under both scoring protocols: $\chi_2(1)=1524$, $p<0.001$ for strict scoring and $\chi_2(1)=1165$, $p<0.001$ for sensitive scoring. This finding indicates that Item plays a significant role in determining score.

Given this finding, further analyses were conducted to determine Item Facility (IF) and Item Discrimination (ID) of each item. IF is a measure widely used in language assessment. It is defined on a scale from 0 to 1 as the proportion of test-takers who correctly answered the item. IF, therefore, measures how easy or difficult the item is. According to Carr (2011), values greater than 0.7 and less than 0.3 indicate that an item is overly easy or overly difficult, respectively. ID is defined as a correlation between the score on the item and the total score on the test. ID, thus, ranges from -1 to +1 and reflects the ability of an item to discriminate between high-scoring students and low-scoring students. Carr (2011) suggests that ID values above 0.3 indicate an item's ability to successfully discriminate between these two groups of students, while values between 0 and 0.3 do not discriminate well (i.e., are of little use to the test), and values below 0 suggest that an item is actively hurting the test's ability to discriminate between high- and low-scoring students, as these values are often answered correctly by low-scoring students but incorrectly by high-scoring students.

The results of the IF/ID analysis in terms of the courses are presented in Table 6 and Table 7. IF measures with respect to the three courses indicate that around half of the items that were too easy (i.e., greater than 0.7) were in the Spanish 202 course, and the number of items that were too easy under sensitive scoring was over double the number under strict scoring (66 and 31, respectively). Meanwhile, there were relatively fewer items that were too difficult (i.e., IF less than 0.3) under strict scoring, and nearly all of these became at an appropriate difficulty level under sensitive scoring (19 and 4, respectively). Additionally, the number of items that had unacceptable IF under both scoring protocols was relatively well distributed among courses, as can be seen in the final column under each scoring protocol. As for ID (see Table 7), it is notable that very few of the 117 total items were bad (i.e., greater than 0 and less than 0.3) at discriminating high- from low-scoring participants, and no items were actively hurting the quizzes' ability to discriminate (i.e., less than 0). Around half of the poorly discriminating items were found in the 102 lists, and there were considerably more poorly discriminating items under sensitive scoring than under strict scoring.

Table 6. *Item Facility Values by Course*

		IF strict			IF sensitive		
Course	# Items	>0.7	<0.3	>0.7 or <0.3	>0.7	<0.3	>0.7 or <0.3
102	38	9	9	18	20	2	22
201	37	7	8	15	15	2	17
202	42	15	2	17	31	0	31
Total	117	31	19	50	66	4	70

Table 7. *Item Discrimination Values by Course*

		ID strict	ID sensitive
Course	# Items	<0.3	<0.3
102	38	5	9
201	37	3	4
202	42	2	4
Total	117	10	17

Note. These numbers include items that had 100% IF and therefore undefined ID values.

The results of the IF/ID analysis in terms of related and unrelated lists are presented in Table 8 and Table 9. First, the overly easy items were split between related and unrelated lists under both scoring protocols, as can be seen in Table 8, while overly difficult items were largely found in unrelated lists. This sheds further light on the previous finding that the majority of the overly difficult items under strict scoring came from lists in Spanish 102 and Spanish 201. In addition, many of the overly difficult items were made to be of appropriate difficulty under sensitive scoring. It is also notable that 66 out of all 117 items (56%) were overly easy under sensitive scoring, and roughly half of all items were either too easy or too difficult under both scoring protocols. Finally, ID measures show that most poorly discriminating items were found in unrelated lists. This suggests that higher-performing learners may utilize certain strategies that allow them to learn related lists better than lower-performing learners, while unrelated lists don't allow higher-performing learners to utilize such strategies and the unrelated items therefore discriminate more poorly.

Table 8. *Item Facility Values by List Type*

		IF strict			IF sensitive		
Course	# Items	>0.7	<0.3	>0.7 or <0.3	>0.7	<0.3	>0.7 or <0.3
Related	61	16	3	19	39	1	40
Unrelated	56	15	16	31	27	3	30
Total	117	31	19	50	66	4	70

Table 9. *Item Discrimination Values by List Type*

		ID strict	ID sensitive
Course	# Items	<0.3	<0.3
Related	61	3	4
Unrelated	56	7	13
Total	117	10	17

Note. These numbers include items that had 100% IF and therefore undefined ID values.

In sum, item analyses indicated several major findings: (1) around half of all 117 items were either overly easy or overly difficult for participants, with more items being overly easy than overly difficult; (2) overly easy or difficult items were not constrained to just one course or list type, but rather they were spread out between courses and list types, for the most part; (3) no items were actively hurting the quizzes’ ability to discriminate among participants; and (4) a number of items did not help in discriminating among high- and low-scoring participants.

CHAPTER 5. DISCUSSION AND CONCLUSION

This study investigated the effects of semantic relatedness in L2 vocabulary lists on learning in an authentic classroom environment by testing students' vocabulary knowledge of one semantically related list and one unrelated list at the end of their respective units in their courses. This chapter will summarize and discuss the findings of this experiment followed by limitations and implications for the field.

Results of the paired samples *t*-test indicated that students scored significantly higher on quizzes over related lists than unrelated lists. In a controlled experiment, this would suggest that related lists are better for language learners than unrelated lists. However, the present study was designed as purely observational and was conducted in an authentic learning environment. Control of any variables that would hurt authenticity of the study was intentionally avoided. Thus, the finding of the *t*-test might, in fact, be related to an effect other than the list type, such as course level or item characteristics.

MER analyses, therefore, were conducted to account for the different possible predictor variables, which included, in addition to the fixed effect of list type, also a fixed effect of course and random effects of participants and items. The MER models failed to find a significant main effect of list type, and course, and their interaction on test scores under both strict and sensitive scoring protocols. Specifically, the addition of list type to the intercept-only model did not explain the data better, suggesting that list type does not play a significant role in predicting test scores. Further, the addition of course into the model did not result in a significant gain in model fit, which indicates that course does not play a role in predicting test scores either. Finally, the addition of an interaction between list type and course showed no significant gain in model fit, which suggests that students in the different courses reacted similarly to related and unrelated lists.

Post-hoc analyses, however, revealed that the random effect of item was significant for predicting test scores. This means that lexical idiosyncrasy, that is the various individual features associated with each lexical item, was the only significant factor determining learning outcomes in the study. While the results of the *t*-test seem to contradict those of the MER models, these findings are not, in fact, contradictory. Rather, the *t*-test found a significant difference between list types because the random effects were not taken into account as they were in the MER models. Thus, no evidence was found for the benefit of either the related or the unrelated list type. This is in line with previous findings on this topic, which have been mixed, as discussed above.

To further describe the significant effect of item on the results, item facility (IF) and item discrimination (ID) measures were calculated for each item to see which items were too easy or difficult and which were not useful for discriminating between participants. IF and ID results indicated that many (around half) of the 117 items were overly easy (e.g., *eye* and *to swim*) or overly difficult (e.g., *to play chess* and *to surprise*), and a number of items were not useful in discriminating among participants (e.g., *nose* and *Christmas*). This corroborates the previous finding that lexical idiosyncrasy had an effect on participants' performance. Furthermore, the finding that ID is better in related lists suggests that higher-performing students may utilize certain strategies that allow them to cope with related lists better, while those strategies may not manifest themselves in unrelated lists. Thus, presenting students with related lists may allow higher-performing students to utilize those learning strategies.

Importantly, while the IF and ID measures are used in language assessment for improving tests (i.e., to remove items that perform sub-optimally), there were no grounds for removing poorly performing items in the present study, as this would result in *ex post facto* control over the experimental conditions and diminish the authenticity of the study. In other words, the authenticity of the present study makes it “noisy” by nature, and control over the items used would necessarily eliminate some of that (desirable) noise and thus would take away the core design characteristic of the study. Instead, to mitigate the influence of item on score, future research might control for item-level characteristics statistically, such as by including IF values in the regression models as a covariate. This would allow for control over item-level characteristics that might be statistically relevant (e.g., word length, frequency, cognates, part of speech) without invalidating the ecological validity of the study. Furthermore, it is also important to note that the inclusion of item as a predictor in the regression models is critical to accurately interpret the data. Exclusion of item from the models would assume that all items are of equal difficulty, when certainly no such argument may be made.

It is important to consider the potential root causes of the lexically idiosyncratic effect on learning gains. On one hand, it is possible that some lexical items are inherently more difficult than others. For example, the most difficult item (by average score) in the dataset, *to surprise*, may be more difficult than the easiest item, *eye*, regardless of other factors. On the other hand, items could be easier or harder precisely due to their relationships to items presented simultaneously. In other words, perhaps items are rendered easier when they are alongside other semantically related items. Of course, it may also be the case that a combination of these two explanations, or some other unknown factor, lies at the root cause of item effect. Unfortunately, it is not possible to determine this root cause with the present data, as there is no way to control for item difficulty.

Limitations and Future Research

Before concluding, the limitations of this study should be discussed. In order to maintain the authentic nature of this study, many factors were intentionally left uncontrolled in order to most accurately answer the research question at hand, and therefore the research design lends itself to more critique than the typical controlled experiment. However, the benefits of the use of this uncontrolled design have been argued above. The design decisions that can also be perceived as limitations included naturalistic sampling of participants and items and the lack of pretest.

Some factors, however, could be controlled in future studies without much impact on their authenticity. Quiz duration, for example, could be controlled for, and the use of forward translation could be supplemented with additional means of assessment to measure other aspects of the many-faceted vocabulary knowledge (Nation, 1990).

It also important to note that the nonsignificant results from the model comparisons may have two possible causes: (1) true lack of difference between list types, or (2) lack of statistical power due to small sample size. Future studies should attempt to increase the sample size or carry out Bayesian analysis (as is becoming increasingly popular in applied linguistics; see Norouzian, de Miranda, & Plonsky, 2018) to better interpret the present nonsignificant findings.

Next, the amount and type of studying, arguably two of the most important factors governing performance on the quizzes, were left uncontrolled. Leaving participants to their own devices with regard to these factors was one of the key aspects of the experimental design, and though it differs rather starkly from much other empirical research in the field, this aspect of the study allowed the investigation of the present research question and should not be considered a limitation. However, in future research, participants could be asked about their studying habits, and this information could be used to control the findings statistically.

The question of high heterogeneity across experimental designs in the field of vocabulary learning is also worth mention. It is possible that the high degree of differences across methodologies, as noted by Nakata and Suzuki (2019) and as shown in Table 1 and Table 2 above, may be partially responsible for the difficulty in determining why one study does or does not align with another. Therefore, in order to more easily compare results across studies, future researchers should consider adopting a more standardized experimental design, such as a common set of items or a common means of vocabulary assessment.

Finally, as hinted at in the discussion above, the lack of control over item characteristics in this study may have rendered a potentially significant effect of semantic relatedness hidden. Future studies should therefore control for item-level characteristics statistically, such as by including IF measures as a covariate in the regression models, in order to mitigate these concerns without reducing ecological validity of the study.

Implications

The present results have two major implications. First, the question in the title of this thesis may be addressed: does semantic relatedness really matter in L2 vocabulary learning? While some researchers, such as Folse (2004), urge teachers and materials developers to revise all of their semantically related lists to be thematically related or unrelated, the findings presented in this study suggest that such definitive recommendation may not be fully justifiable. Rather, some factors (possibly related to lexical idiosyncrasy) may result in related items being not only *not* inferior but even possibly superior, in some cases, to unrelated words for learning. Furthermore, this study was purely observational: no manipulations such as re-writing related lists were imposed on the classroom setting, and the materials used in this study had been developed by textbook authors based on usual pedagogical considerations. These usual considerations seem to have worked equally well for both related and unrelated lists. Therefore, teachers and materials developers may take the following recommendation from this study: when the material calls for a related list, use a related list, and when it calls for an unrelated list, use an unrelated list.

Secondly, this study is the first (to the author's knowledge) to investigate the effects of semantic relatedness on L2 vocabulary learning in a truly authentic classroom environment. As past researchers have noted (e.g., Ishii, 2017; Papathanasiou, 2009; Waring, 1997), investigations that are generalizable to the classroom are critical for teachers to implement findings into pedagogy. While some previous studies have attempted to carry out investigations in the authentic classroom (e.g., Hoshino, 2010), they have in many respects failed to capture true authenticity. As such, this study may serve as a model for future studies.

In conclusion, this thesis has hopefully made a step toward closing the gap between research and practice discussed at the beginning of this thesis. The research carried out here in an authentic classroom environment may contribute to the field in the effort to one day provide generalizable research findings that would be useful for teachers.

REFERENCES

Baddeley, A. (1997). *Human memory: Theory and practice*. East Sussex, UK: Psychology Press.

Belcher, D. (2007). A bridge too far? *TESOL Quarterly*, *41*(2), 396–399.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Blanco, J. (2016). *Portales*. Vista Higher Learning.

Cargill, M. (2007). The research/pedagogy interface in a 21st-century publication context. *TESOL Quarterly*, *41*(2), 394–396. https://doi.org/10.1002/j.1545-7249.2007.tb00065.x

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Chapelle, C. (2007). Pedagogical implications in TESOL Quarterly? Yes, please! *TESOL Quarterly*, *41*(2), 404–406. https://doi.org/10.1002/j.1545-7249.2007.tb00068.x

Deci, E., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Ellis, R. (2005). Principles of instructed language learning. *System*, *33*, 209–224. https://doi.org/10.1016/j.system.2004.12.006

Ellis, R. (2010). Second language acquisition, teacher education and language pedagogy. *Language Teaching*, *43*(2), 182–201. https://doi.org/10.1017/S0261444809990139

Ellis, R. (2015). *Understanding second language acquisition* (2nd ed.). Oxford University Press.

Erten, I. H., & Tekin, M. (2008). Effects on vocabulary acquisition of presenting new words in semantic sets versus semantically unrelated sets. *System*, *36*, 407–422. https://doi.org/10.1016/j.system.2008.02.005

Finkbeiner, M., & Nicol, J. (2003). Semantic category effects in second language word learning. *Applied Psycholinguistics*, *23*, 369–383. https://doi.org/10.1017/S0142716403000195

Folse, K. S. (2004). Myths about teaching and learning second language vocabulary: What recent research says. *TESL Reporter*, *37*(2), 1–13.

Freeman, D., & Johnson, K. E. (1998). Reconceptualizing the knowledge-base of language teacher education. *TESOL Quarterly*, *32*(3), 397–417. https://doi.org/10.4324/9780203835654

Han, Z. (2007). Pedagogical implications: Genuine or pretentious? *TESOL Quarterly*, *41*(2), 387–393.

Hashemi, M. R., & Gowdasiaei, F. (2005). An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC Journal*, *36*(3), 341–361. https://doi.org/10.1177/0033688205060054

Horwitz, E. K. (2001). Language Anxiety and Achievement. *Annual Review of Applied Linguistics*, *21*, 112–126.

Hoshino, Y. (2010). The categorical facilitation effects on L2 vocabulary learning in a classroom setting. *RELC Journal*, *41*(3), 301–312. https://doi.org/10.1177/0033688210380558

Hunt, R. R., & Elliot, J. M. (1980). The role of nonsemantic information in memory: Orthographic distinctiveness effects on retention. *Journal of Experimental Psychology: General*, *109*(1), 49–74. https://doi.org/10.1037/0096-3445.109.1.49

Iowa State University. (2019). *Fall semester 2019 enrollment* [Data set]. https://www.registrar.iastate.edu/enrollment/sex-ethnicity-residence

Iowa State University. (2019, October 4). *Nontraditional students*. https://www.admissions.iastate.edu/nontrad/index.php

Ishii, T. (2013). Reexamining semantic clustering: Insight from memory models. *Vocabulary Learning and Instruction*, *2*(1), 1–7. https://doi.org/dx.doi.org/10.7820/vli.v02.1.ishii

Ishii, T. (2015). Semantic connection or visual connection: Investigating the true source of confusion. *Language Teaching Research*, *19*(6), 712–722. https://doi.org/10.1177/1362168814559799

Ishii, T. (2017). The impact of semantic clustering on the learning of abstract words. *Vocabulary Learning and Instruction*, *6*(1), 21–31. https://doi.org/10.7820/vli.v06.1.Ishii

Lehrer, A. (1974). *Semantic fields and lexical structure*. Amsterdam: North-Holland.

Long, M. H. (2011). Methodological principles for language teaching. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 373–394). John Wiley & Sons.

Magnan, S. S. (2007). Gauging the scholarly value of connecting research to teaching. *TESOL Quarterly*, *41*(2), 400–404. https://doi.org/10.1002/j.1545-7249.2007.tb00067.x

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins.

Nakata, T. (2015). Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition*, *37*(4), 677–711. https://doi.org/10.1017/S0272263114000825

Nakata, T., & Suzuki, Y. (2019). Effects of massing and spacing on the learning of semantically related and unrelated words. *Studies in Second Language Acquisition*, *41*(2), 287–311. https://doi.org/10.1017/S0272263118000219

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2000). Learning vocabulary in lexical sets: Dangers and guidelines. *TESOL Journal*, *9*(2), 6–10. https://doi.org/10.1002/j.1949-3533.2000.tb00239.x

Noels, K. A., Pelletier, L. G., Clément, R., & Vallerand, R. J. (2000). Why are you learning a second language? Motivational orientations and self-determination theory. *Language Learning*, *50*(1), 57–85. https://doi.org/10.1111/0023-8333.00111

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*(4), 1032–1075.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. New York: Cambridge University Press.

Oxford, R. L. (1989). Use of language learning strategies: A synthesis of studies with implications for strategy training. *System*, *17*(2), 235–247.

Papathanasiou, E. (2009). An investigation of two ways of presenting vocabulary. *ELT Journal*, *63*(4), 313–322. https://doi.org/10.1093/elt/ccp014

Pérez-Gironés, A. M., & Adán-Lifante, V. (2014). *Más* (2nd ed.). New York: McGraw Hill.

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363. https://doi.org/10.1177/1362168808089921

Schneider, V. I., Healy, A. F., & Bourne, L. E. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E. Bourne (Eds.), *Foreign Language Learning: Psycholinguistic Studies on Training and Retention* (pp. 77–90). Mahwah, NJ: Lawrence Erlbaum Associates.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What Is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440. https://doi.org/10.1006/jmla.2001.2813

Stoller, F., & Grabe, W. (1993). Implications for L2 vocabulary acquisition and instruction from L1 vocabulary research. In T. Huckin, M. Haynes, & J. Coady (Eds.), *Second language reading and vocabulary learning* (pp. 24–45). Norwood, NJ: Ablex Publishing Corporation.

Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21*(3), 371–380.

Tinkham, T. (1997). The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Language Research*, *13*(2), 138–163.

Tseng, W.-T., & Schmitt, N. (2008). Toward a model of motivated vocabulary learning: A structural equation modeling approach. *Language Learning*, *58*(2), 357–400. https://doi.org/10.1111/j.1467-9922.2008.00444.x

Waring, R. (1997). The negative effects of learning words in semantic sets: A replication. *System*, *25*(2), 261–274.

Wilcox, A., & Medina, A. (2013). Effects of semantic and phonological clustering on L2 vocabulary acquisition among novice learners. *System*, *41*(4), 1056–1069. https://doi.org/10.1016/j.system.2013.10.012

APPENDIX A. QUIZZES

Table A1. *Related Quizzes by Course*

Spanish 102	Spanish 201	Spanish 202
Eye	Nephew/Niece	Earth
(Sense of) hearing; inner ear	Brother-in-law/Sister-in-law	To cultivate
Toe	Kinship	To waste
Neck	Stepbrother/Stepsister	Forest
Knee	Daughter-in-law/Son-in-law	Drought
Head	Father-in-law/Mother-in-law	To sow
Ankle	Stepmother/Stepfather	Species
Leg	In-laws	Environmental
Stomach	Husband/Wife	To create
Foot	Paternal (on the father's side)	Soil/Land
Finger	Great grandfather/grandmother	Savings
Bone	Half-brother/Half-sister	Ground
Body	Godmother/Godfather	Harvest/Crop
(Outer) ear	Maternal (on the mother's side)	Ozone layer
Mouth	Godson/Goddaughter	To collect/pick up
Throat	Stepson/Stepdaughter	To save
Heart		Flood
Arm		To protect
Nose		Wood
		To exploit
		Sky/Heaven
		River
		Resource
		Jungle/Tropical rain forest
		Waste
		Agricultural

Table A2. *Unrelated Quizzes by Course*

Spanish 102	Spanish 201	Spanish 202
To surprise	To stay up all night	Summit
(Wedding) anniversary	To do a crossword puzzle	To benefit
To give (a gift)	To dance	Tax/Levy
Surprise	Fair	Peace
To toast (drink)	To go to a concert	To unite
Christmas	To swim	To promote
To have fun	To stroll	To separate
To have a good/bad time	To play cards	Exchange
Young woman celebrating her fifteenth birthday	To chat/converse	To agree to
To celebrate	Beach	Tie
To smile	Swimming pool	Conference/Lecture
Party	To have a barbeque	Income
To laugh	Stroll	Commerce
Guest	To go to the theater	To harm
Wedding	Square	Organization/Body
Holiday	To tell a joke	To strengthen
To relax	To play dominoes	
To invite	Dance	
Birthday	To go to the movies	
	To play chess	
	Street	

APPENDIX B. IRB APPROVAL

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
2420 Lincoln Way, Suite 202
Ames, Iowa 50014
515 294-4566

Date: 03/26/2018

To: Evg Chukharev-Khudilaynen

From: Office for Responsible Research

Title: **Effects of semantic relatedness on learning vocabulary lists in the second language classroom**

IRB ID: **18-080**

Submission Type: Initial Submission **Exemption Date:** 03/26/2018

The project referenced above has been declared exempt from the requirements of the human subject protections regulations as described in 45 CFR 46.101(b) because it meets the following federal requirements for exemption:

1: Research conducted in an established or commonly accepted educational setting; involving normal educational practices, such as (i) Research on regular and special education instructional strategies, or (ii) Research on the effectiveness or the comparison among instructional techniques, curricula, or classroom management methods.

The determination of exemption means that:

□ **You do not need to submit an application for annual continuing review.**

□ **You must carry out the research as described in the IRB application.** Review by IRB staff is required prior to implementing modifications that may change the exempt status of the research. In general, review is required for any *modifications to the research procedures* (e.g., method of data collection, nature or scope of information to be collected, changes in confidentiality measures, etc.), modifications that result in the *inclusion of participants from vulnerable populations*, and/or any *change that may increase the risk or discomfort to participants*. *Changes to key personnel* must also be approved. The purpose of review is to determine if the project still meets the federal criteria for exemption.

Non-exempt research is subject to many regulatory requirements that must be addressed prior to implementation of the study. Conducting non-exempt research without IRB review and approval may constitute non-compliance with federal regulations and/or academic misconduct according to ISU policy.

Detailed information about requirements for submission of modifications can be found on the Exempt Study Modification Form. A Personnel Change Form may be submitted when the only modification involves changes in study staff. If it is determined that exemption is no longer warranted, then an Application for Approval of Research Involving Humans Form will need to be submitted and approved before proceeding with data collection.

Please note that you must submit all research involving human participants for review. **Only the IRB or its designees may make the determination of exemption**, even if you conduct a study in the future that is exactly like this study.

Please be aware that **approval from other entities may also be needed**. For example, access to data from private records (e.g., student, medical, or employment records, etc.) that are protected by FERPA, HIPAA or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **An IRB determination of exemption in no way implies or guarantees that permission from these other entities will be granted**.

Please be advised that your research study may be subject to **post-approval monitoring by Iowa State University's Office for Responsible Research**. In some cases, it may also be subject to formal audit or inspection by federal agencies and study sponsors.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.