UNIVERSITY OF CALIFORNIA,
IRVINE

Exploring Corpus Use in Second Language Vocabulary Learning: Toward the Establishment of a Data-Driven Learning Model

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Hansol Lee

Dissertation Committee:
Professor Mark Warschauer, Chair
Professor Robin C. Scarcella
Professor Dorothy M. Chun

2018

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisor and committee chair, Dr. Mark Warschauer. His perpetual enthusiasm as a researcher continually inspired and encouraged me to navigate my doctoral journey with confidence and trust; he guided me through the journey from beginning to end.

My sincere gratitude also goes to my committee members, Dr. Robin C. Scarcella and Dr. Dorothy M. Chun. Dr. Scarcella, who is my co-advisor, clearly demonstrated her expertise in second language (L2) research and passion for supporting L2 learners. Dr. Chun, who thankfully agreed to serve as a committee member out of her busy schedule, gave me thoughtful and detailed feedback on my dissertation.

It has been my honor to have Dr. Warschauer, Dr. Scarcella, and Dr. Chun, who are nation's leading researchers on L2 studies, as my committee members. Without their guidance and help, this dissertation would not have been possible.

I also would like to thank my dearest friend, Dr. Jang Ho Lee, who has greatly influenced my academic journey as an applied linguist. He has always been there for me to give me advice, help, laugh, and encouragement for this journey.

Lastly and most importantly, I cannot thank enough my wife. Words cannot express my gratitude and love for her.

I thank the Oxford University Press for permission to include Chapter 1 of my dissertation, which was published in *Applied Linguistics*, Mark Warschauer and Jang Ho Lee for permission to include Chapter 2 of my dissertation, which was published in *Language Learning & Technology*, and the Cambridge University Press for permission to include Chapter 3 of my dissertation, which was published in *ReCALL*.

Financial support was provided by the University of California, Irvine, Republic of Korea Army Headquarters, and Phi Beta Kappa Alumni Association.

CURRICULUM VITAE

Hansol Lee

2018	Ph.D. in Education, University of California, Irvine Specialization: Language, Literacy, and Technology (Applied Linguistics) Advisor and dissertation chair: Dr. Mark Warschauer
2016	M.A. in Education, University of California, Irvine
2009	M.A. in English Language and Literature, Seoul National University, South Korea
2005	B.A. in English, Korea Military Academy, South Korea

FIELD OF STUDY

Applied Linguistics

ABSTRACT OF THE DISSERTATION

Exploring Corpus Use in Second Language Vocabulary Learning:
Toward the Establishment of a Data-Driven Learning Model

By

Hansol Lee

Doctor of Philosophy in Education

University of California, Irvine, 2018

Professor Mark Warschauer, Chair

This dissertation aims to understand and further explore the effects of corpus use on second language (L2) vocabulary learning in classroom contexts using new methodologies and theoretical perspectives. Corpora (the singular is corpus), which are rich collections of authentic language data, have been widely used for language learning purposes because they provide an ample size of authentic L2 exposures to learners, usually hard to attain in real life. Along with the development of the concordance program, which displays multiple example contexts of selected target vocabulary items in a visually salient way, researchers in the field of L2 studies have demonstrated a continuous empirical effort to prove the positive effect of corpus use. However, given the short history of corpus linguistics, little is known about the overall effect of corpus use across different learning contexts. To address the paucity, I conducted four studies as following. In Chapter 1, I systemically reviewed the overall effect of corpus use on L2 vocabulary learning using a multilevel meta-analysis approach. In Chapter 2, by conducting an experiment I examined the effect of the confirmation process for corpus use as glossary information using a linear

regression analysis. In Chapter 3, I re-analyzed the data collected to unearth hidden groups of learners based on their vocabulary learning across different learning conditions using a data mining approach. In Chapter 4, I conducted another experiment to investigate the role of learner factors in corpus-based L2 vocabulary learning by using a mixed-method approach and examined if the collected data fit the hypothesized model of Data-Driven Learning (DDL; Johns, 1991). I expect that my dissertation will lead to a deeper understanding of corpus use in L2 vocabulary learning. Considering that the use of technology in an L2 classroom has become a norm, the findings of the dissertation provide researchers and teachers with a few guidelines they should consider when they adopt technologies in L2 vocabulary learning contexts.

INTRODUCTION

Vocabulary learning has been recognized for its central role in improving literacy (e.g., Freebody & Anderson, 1983; Snow, Burns, & Griffin, 1998). Closing the vocabulary gap is particularly important in helping struggling students improve their reading comprehension (e.g., Lawrence, Crosson, Paré-Blagoev, & Snow, 2015; Mancilla-Martinez & Lesaux, 2010). Similarly, in the field of second language (L2) learning, vocabulary learning is crucial for developing literacy (e.g., Bernhardt, 1991; Nation, 2001; Read, 2000). In her model of L2 reading, for example, Bernhardt (1991) suggests that vocabulary serves as a critical component; this framework has been supported by empirical studies that demonstrate that vocabulary is a strong predictor of students' reading comprehension (Guo, 2008; Shin & Kim, 2012).

However, L2 learners generally have fewer opportunities to be exposed to L2 language input than to their first language (L1). Moreover, considering that an understanding of multi-faceted disciplinary vocabulary is required for them to succeed, L2 learners' lexical knowledge needs more attention (e.g., Graves, 2006; Laufer & Yano, 2001; Stahl & Nagy, 2006). As a response to the unique learning environment faced by L2 learners, some researchers in the field of L2 vocabulary focused on using corpora, which can be defined as structured collections of natural language data (Biber, Conrad, & Reppen, 1998; the singular is corpus), as one way to provide learners with more opportunities for L2 exposure.

Although the use of corpora has long seemed pedagogically appropriate, it has become available only with the development of computer technology since the 1960s (e.g., Godwin-Jones, 2001a; Reppen, 2010; Sinclair, 1997). That is, a computer program called a

concordancer, which is an essential tool for analyzing how a target word or phrase is used in different contexts, allows learners to gain easy and useful access to corpora for a pedagogical purpose (Cheng, 2012). The programs display the typed item in the center of multiple example sentences, a format called "Key Word In Context" (KWIC), which exposes target language items more frequently (Ellis, 2002), makes the target language items salient input (Chapelle, 2003), and thus increases the possibility of learner' notice as well as acquisition of the target language items (Schmidt, 2001). The field has seen an increasing number of empirical studies on the effects of corpora use on L2 vocabulary learning; however, it should be noted that these studies have shown large variations in methodological features (e.g., learning tasks, instructions, participants, and assessments) and some intrinsic limitations of the use of corpora in L2 vocabulary learning (e.g., the language input should be comprehensible as well as informative; see the following section for a detailed review).

In view of the current state of research, I conducted four studies (Chapters 1-4) in my dissertation to extend our understanding of Data-Driven Learning (DDL; Johns, 1991) by not only synthesizing the overall findings of empirical studies on corpus use in L2 vocabulary learning but also seeking a more reliable and comprehensive conclusion with new theoretical perspectives and statistical techniques. In Chapter 1, I conducted a meta-analysis to estimate the overall effect of corpus use on L2 vocabulary learning and to find out how much moderating variables influence the estimated effect of corpus use. This approach differs from previous meta-analyses on the effect of corpus use, such that I used the multilevel effect size (ES) calculation (i.e., multiple ESs for each study; ESs nested in studies), specifically focused on sub-dimensions of L2 vocabulary knowledge, and used

regression analyses not only to statistically compare and contrast mean ESs across moderators (i.e., elements of population, publication, and treatment data of the collected studies) but also to provide adjusted mean ESs for each value of the moderators (Woltman, Feldstain, MacKay, & Rocchi, 2012).

In Chapter 2, to examine the value of concordance lines as effective glossary information for L2 learners' acquisition of word meaning, I tested the effects of two different types of e-glosses, with the first type providing the concordance lines of target vocabulary, and the second type providing the concordance lines plus the definition of target vocabulary, under a repeated-measure design (i.e., within-subject). I also analyzed log data related to the participants' interactions with e-glosses in order to gauge the extent to which they consulted the glossed items and comprehended concordance glossary information. Along with results from the experimental phase, interview data with a subset of the participants and the record of their implementation of e-glosses aided in understanding of the learners' complex interactions with e-glosses.

In Chapter 3, I used a data mining technique to unearth possible different learner types from the previously collected experimental data (Chapter 2). In doing so, I hypothesized that there may be different learner types who show different learning patterns to maximize their L2 vocabulary learning. Such differential learning patterns would deviate from previous variable-centred findings on the correlation between the amount of glossary information provided and L2 vocabulary knowledge gains at the group level (Chapter 2).

In Chapter 4, I investigated the role of learner factors in corpus-based L2 vocabulary learning, with the aim of establishing a model of DDL. To this end, a mixed-method

investigation, which consisted of qualitative (e.g., analyzing participants' strategy use in DDL activities) and quantitative (e.g., structural equation modeling) components, was conducted to explore types of lexical inferencing strategies, the roles of L2 proficiency, strategy use, and working memory, and the relationships of these factors to corpus-based L2 vocabulary acquisition and retention.

It should be noted here that this dissertation incorporates my own materials and work that were published during my doctoral journey. Chapter 1 is based on Lee, Warschauer, and Lee (2018b), Chapter 2 is based on Lee, Warschauer, and Lee (2017), and Chapter 3 is based on Lee, Warschauer, and Lee (2018a). Also, some parts of these papers are interpolated into this Introduction section. I obtained necessary copyright permissions to reproduce materials from the related original sources.

CHAPTER 1: Literature Review and Meta-Analysis[1]

Effects of Corpus Use on Second Language Vocabulary Learning

1. Introduction

Large collections of structured language data (Sinclair, 2004), also known as corpora (singular: *corpus*), have been widely used in the field of L2 learning (e.g., Biber, Conrad, & Reppen. 1998; Johns, 1991; Sinclair, 1991, 2004) and have served several pedagogical purposes and applications, such as for creating learning materials, designing hands-on activities, and building textbooks, workbooks, and dictionaries (see Flowerdew, 1993). In particular, corpus use has been applied to instruction in L2 learning of grammatical and lexical items by providing multiple contextual examples of target items. With such examples, a learner is induced to discover linguistic features by exploring the language data that provide authentic linguistic information necessary for learning (Johns, 1991), a process called DDL. More recently, corpus tools (e.g., concordance software) equipped with useful pedagogical features have become widely available, offering students opportunities for hands-on analysis of corpora in classrooms (Godwin-Jones, 2001b; Reppen, 2010).

Though corpus use has many purposes in instructed L2 learning, there has been a particular pedagogical and research focus on its effects on vocabulary learning (see studies with an asterisk in the Reference section), by drawing on the notion of DDL. These corpus-

[1] The text of this chapter is a reprint of the material as it appears in Lee, H., Warschauer, M., & Lee, J. H. (2018b). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*. Advance online publication. https://doi.org/1.1093/applin/amy012. I was the primary investigator and author of this paper, and the co-authors directed and supervised research which forms the basis for the paper.

based vocabulary studies have covered a wide range of methodologies (e.g., characteristics of a given task, pedagogical contexts, and types of measurement) and topics (e.g., suitability of language data, mastery of corpus consultation skills; see the following sections for a detailed review). A synthesis of this rich and varied research can thus inform pedagogical guidelines for the corpus-based approach to L2 vocabulary teaching and learning.

There have been several recent meta-analyses (e.g., Cobb & Boulton, 2015; Mizumoto & Chujo, 2015; Boulton & Cobb, 2017) reporting that corpus use had overall positive effects on L2 learning. To further this synthesizing work, in this chapter, I conducted a meta-analysis to not only estimate the overall effect of corpus use on L2 vocabulary learning, but also find out how much moderating variables influence the estimated effect of corpus use. Chapter 1 differs from previous meta-analyses on the effect of corpus use, such that in this study I (a) use the multilevel effect size (ES) calculation (i.e., multiple ESs for each study; ESs nested in studies), (b) specifically focus on sub-dimensions of L2 vocabulary knowledge, and (c) use regression analyses not only to statistically compare and contrast mean ESs across moderators (i.e., elements of population, publication, and treatment data of the collected studies) but also to provide adjusted mean ESs for each value of the moderators (Woltman, Feldstain, MacKay, & Rocchi, 2012; see the data analysis plan section for details). In the following sections, I review the theoretical background of corpus use as well as previous studies on corpus-based L2 vocabulary learning. Next, I describe the statistical procedures for multilevel meta-analysis, including the literature search, inclusion criteria, coding procedure, multilevel ES calculation, and data analysis plan. Then, I present the results of the meta-analysis, along with an in-depth

discussion of the findings. Last, I discuss the limitations of the chapter and the implications of the findings.

2. Background

2.1 Theoretical background for corpus use in L2 vocabulary learning

The core of corpus use lies in its encouraging learners to construct their L2 knowledge independently by exploring the linguistic data compiled from corpora, such as concordances that provide multiple sentential examples of how a target linguistic item is used (Johns, 1994). The inductive process in this definition, then, must involve complex cognitive engagement by learners, including inferencing and hypothesizing on language items (Flowerdew, 2015). Given that the process "presumes that learners build knowledge actively, largely through inductive processes" (Collentine, 2000, p. 45), constructivism can provide theoretical support for corpus use (Cobb, 1999). In principle, the constructivist approach rejects the idea of transferring knowledge to students (Collentine, 2000), but embraces a notion that learners actively participate in learning, reflecting a paradigm shift in education from teacher-centered to student-centered learning (Kaufman, 2004).

Figure 1.1. A Snapshot of Concordance Lines from the Corpus of Contemporary American English presented by Brigham Young University (Davies, 2008-).

Note. Shown are concordance lines (sample sentences) of "in the vicinity of." These sentences are sorted and aligned by the target vocabulary item. In addition, the target expression is highlighted and is thus visually salient.

The other unique feature of corpus use is 'Key Word In Context' (KWIC), which is the most common way to use corpora as learning materials. By means of the KWIC function, which is installed in most corpus-based computer programs (e.g., a concordancer), a student can search for a target linguistic item by typing it, and the programs will display multiple concordance lines of the typed item. For example, Figure 1.1 is a screenshot of concordance lines of an idiom, 'in the vicinity of,' excerpted from the Corpus of Contemporary American English presented by Brigham Young University (Davies, 2008-). As the concordance lines expose a learner to multiple incidences of target language items (i.e., frequency effect: Ellis, 2002), and the target item is highlighted, the KWIC format, shown in Figure 1.1, does not only make the concordance lines salient learning input (i.e., input enhancement: Chapelle, 2003; Wong, 2005) but also increases the possibility that learners will notice target items and be able to acquire them (i.e., noticing hypothesis: Schmidt, 2001; Lai & Zhao, 2005). Engaging with a large compilation of language data inductively also requires great learner involvement, which makes learning target lexical items easier according to Laufer and Hulstijn's (2001) involvement load hypothesis. Taken together, corpus use can be theoretically supported by constructivism, the inductive approach, and several major second language acquisition (SLA) frameworks.

2.2 Empirical evidence and limitations of corpus use in L2 vocabulary learning

The abovementioned theoretical frameworks have attracted the attention of the empirical researchers who are investigating the effectiveness of corpus use for improving different dimensions of L2 vocabulary knowledge. For example, Cobb (1997, 1999) conducted an empirical study with two different vocabulary learning conditions: (1) using corpus tools for vocabulary learning (experimental condition) and (2) using dictionaries

and word lists (control condition—traditional vocabulary learning). The results of Cobb's study with adult Omani L2 learners showed that using corpus tools yielded more gains in terms of the learners' L2 vocabulary knowledge, such as their definitional knowledge and their ability to transfer their word knowledge to other contexts. In a study by Chan and Liou (2005), adult Taiwanese L2 learners completed web-based practice units using a bilingual corpus tool, and the results showed significant improvement in the participants' knowledge of collocation. In addition, Frankenberg-Garcia's studies (2012, 2014) confirmed the positive effects of using concordance lines excerpted from corpus data on L2 vocabulary learning. The results indicated that these examples were effective for L2 learners in terms of understanding target vocabulary (2012, 2014), correcting typical L2 mistakes (2012), and writing sentences using target vocabulary (2014).

However, corpus use for L2 vocabulary learning is not without limitations. Previously, Boulton (2010) suggested *barrier* as an umbrella term for the limitations of corpus use in language learning, including (1) new material (e.g., KWIC format), (2) technology (e.g., concordancer), and (3) learning approaches (e.g., inductive learning), all of which appear to be germane to the suitability of language data and corpus consultation skills. For example, it has been previously suggested that learners' linguistic inferences from contexts (e.g., concordance lines) could be fallacious (Schmitt, 2008) and that sometimes learners retain these inaccurate inferences in their lexicons (Mondria, 2003). Thus, concordance lines given in corpus use should be comprehensible to learners and should provide enough contextual clues for learners' processing of target lexical items for their linguistic inquiries. In other words, they must be suitable for target learners.

Moreover, learners might need to understand how to effectively use concordance lines or corpus tools for their L2 vocabulary learning practice.

2.3 Considerations about corpus use in L2 vocabulary learning

When the suitability of language data is evaluated, one may observe that a list of concordance lines automatically retrieved from corpus data has an arbitrary proportion of sentences that give strong, relevant, or useful information about target vocabulary (i.e., high-constraint sentences; Adlof, Frishkoff, Dandy, & Perfetti, 2016) and/or sentences that are comprehensible to learners. For this reason, learners may be cognitively burdened if the language in the sentences is either low-constraint or beyond the learners' current L2 proficiency level (Allan, 2009).

With this in mind, several researchers in the field of L2 vocabulary learning have emphasized the importance of carefully selecting corpus data for teaching L2 vocabulary. For example, Frankenberg-Garcia (2012, 2014) identified sample sentences that contain enough contextual clues about the target lexical items and confirmed that the use of such materials resulted in positive learning outcomes. Allan (2009) recommended that teachers use corpus-informed materials in the form of graded readers, which are more finely tuned to learners' L2 proficiency levels, rather than using pre-established reference corpora, such as the Corpus of Contemporary American English (COCA), British National Corpus (BNC), Brown Corpus, and Open American National Corpus (OANC). On similar grounds, Lee, Lee, and Cert (2015) developed a corpus tool designed to assist teachers to easily upload language data in order to build suitable corpus-informed materials for their students. Taken together, the suitability of language data in terms of adapting corpus use for L2 vocabulary learning depends on the learners' L2 proficiency levels, the type of concordance

lines (i.e., carefully selected for comprehension or automatically retrieved), and type of learning material (i.e., corpus-informed materials or reference corpora).

To evaluate the learners' mastery of corpus consultation skills, it has been argued that the KWIC format can hinder effective implementation of corpus use for learners (see Sinclair, 2003) because the presentation of "unfinished sentences" (Johns, 1986, p. 157) in this format (see Figure 1.1) could be unfamiliar to learners. Indeed, concordance lines are not designed for traditional text-reading strategy (e.g., linear reading; Boulton & Cobb, 2017), so providing learners with a clear guideline about how to explore language data would be helpful (Gavioli, 2005; Boulton, 2009). For this reason, some previous studies attempted to help learners handle corpus materials or activities by manipulating instructional or methodological features. For example, Boulton (2010) examined the positive effect of using paper-based materials in DDL on L2 vocabulary learning. By using prepared concordance printouts, Boulton aimed to alleviate learners' burden of dealing with concordance lines in rather unfamiliar computer-assisted environments. However, Boulton (2009) found that the form of KWICs could be effective, even for untrained learners. These apparently contradictory findings about corpus use in language learning call for a meta-analysis that can statistically combine, compare, and contrast the effects of corpus use on language learning in terms of their different methodologies.

3. Present study

Recent meta-analyses have pointed to the value of corpus use in L2 learning (Cobb & Boulton, 2015; Mizumoto & Chujo, 2015; Boulton & Cobb, 2017). For example, Cobb and Boulton's (2015) preliminary meta-analysis included a total of 21 corpus-use studies. They found the overall average ES to be large (Cohen's d = 1.04 for between-group ES; d = 1.68

for within-group ES).[2] In contrast, Mizumoto and Chujo's (2015) meta-analysis on the DDL approach to vocabulary learning, based on 14 studies in Japanese EFL contexts, found a small ES (*d* = .90 for within-group ES).

These two studies, however, have their own limitations. First, as Cobb and Boulton (2015) acknowledged, their meta-analysis was preliminary, in that they calculated only the overall average ES, without taking into account the different methodologies used in the included studies. Mizumoto and Chujo (2015) based their analyses on just one population (Japanese EFL learners), so their conclusion on the effect of DDL on L2 vocabulary learning cannot be generalized to other populations. Furthermore, their meta-analysis did not respond to the call from Cobb and Boulton (2015) about the need to perform moderation estimation.

As a more complete version of Cobb and Boulton (2015), Boulton and Cobb's (2017) meta-analysis included 64 studies representing 88 unique samples and found large ESs for both between-group (*d* = .95) and within-group comparisons (*d* = 1.50). Furthermore, they analyzed moderator variables and found that corpus use had positive effects in almost any language learning situation. However, as in the two previous meta-analyses, Boulton and Cobb selected only a single ES per study for the data analysis, a standard meta-analysis approach in the field, and thus could explore methodological differences only across the included studies.

[2] I follow Plonsky and Oswald's (2014) ES interpretation for L2 studies, where for between-group comparisons (i.e., control/experimental group comparisons) .4, .6, and .9 are considered small, medium, and large ESs, respectively, and .6, 1.0, and 1.4 for within-group comparisons (i.e., pre/post-test designs). Given that ESs should be interpreted within a specific field, Plonsky and Oswald analyzed 346 primary studies and 91 meta-analyses in L2 research and took the 25th, 50th, and 75th percentiles as indicators for the ES interpretation.

Chapter 1 extends the previous meta-analyses to investigate the positive effect of corpus use on L2 learning, but differs from them in my narrower focus on L2 vocabulary learning. Specifically, I explored how corpus use could improve the different dimensions of L2 vocabulary knowledge as part of the analysis of moderator variables. To this end, I chose Henriksen's framework (1999) to distinguish L2 vocabulary knowledge across the following three major dimensions: (1) precise knowledge, i.e., being able to understand the definition of a target lexical item, (2) in-depth knowledge, i.e., being able to state the referential meaning (e.g., synonyms, antonyms) as well as syntactic features (e.g., collocational patterns) of a target lexical item, and (3) productive use ability, i.e., being able to construct a sentence using a target lexical item. Among several other perspectives on L2 vocabulary knowledge, I chose Henriksen's framework because it is widely used; most L2 vocabulary assessments are based on this framework (Gyllstad, 2013).

Furthermore, Chapter 1 is unique in the methodology of meta-analysis used. Instead of computing a single ES per study, I carried out a multilevel meta-analysis (i.e., multiple ESs per study; two-level model) to fully capture large methodological differences *within* (i.e., between effect sizes within studies) and across the included studies. This approach, of course, may raise an issue of statistical independence between the multiple ESs of a study; however, using a multilevel regression analysis should resolve the issue (see Hox, Moerbeek, & van de Schoot, 2010; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013; Pratt, Turanovic, & Cullen, 2016). Also, regression analyses allowed me to statistically compare the different impacts of moderators and to compute adjusted means for each moderator, holding other moderators constant (see the data analysis plan section for details). In this way, I expect a meta-analysis to provide informative pedagogical

implications for the effectiveness of corpus use for L2 vocabulary learning. The following two research questions guided this study:

1. How effective is corpus use in improving L2 vocabulary learning?

2. What are the moderators that influence the magnitude of the effectiveness of corpus use?

4. Methods

In an effort to answer the aforementioned research questions, I focused exclusively on studies based on experimental designs with control conditions for corpus use, unlike the previous meta-analysis studies (i.e., Cobb & Boulton 2015; Mizumoto & Chujo 2015; Boulton & Cobb 2017) which included both within-group (i.e., differences among participants who are in the same group; pre/post-test designs with no control groups) and between-group comparisons (i.e., differences between two or more groups of participants; pre/post-test designs with control groups) in their ES calculations. In some instances, only including one data set (i.e., between-group comparisons) could be less comprehensive than a study based on two (i.e., between-group & within-group comparisons); however, I believe that my decision is more appropriate in the current situation where the primary goal is to examine the overall *treatment effect* of corpus use in L2 vocabulary learning (see Shadish, Cook, & Campbell, 2002). To this end, I took multiple steps to collect and evaluate existing studies for inclusion in the meta-analysis, as described below.

4.1 Literature search

As the first step, I conducted keyword searches (either corpus, corpora, concordance, or data-driven learning) of the literature written in English available up to the year 2016 in major databases, including Education Resources Information Center,

Linguistics and Language Behavior Abstracts, Web of Science, and Dissertation Abstracts Databases. In addition, I manually searched not only in relevant academic journals, including *CALICO Journal*, *Computer-Assisted Language Learning*, *Computer and Education, Language Learning & Technology*, *ReCALL*, and *SYSTEM*, but also through the reference lists in the collected studies and previous meta-analyses in order to find unidentified studies. I identified a total of 52 studies after manually excluding studies unrelated to L2 vocabulary learning.

4.2 Inclusion criteria

First, I decided that a study should implement random control trials along with a control group. If this condition is not met, the study must at least investigate and confirm homogeneity between treatment and control groups and could thus be considered to be a quasi-experimental study. I found that over 30% (19 studies) of the 52 studies did not meet the first criterion. In particular, two of them were small case studies, and one study did not control for baseline differences in a quasi-experimental design. Moreover, 16 of the studies did not have control groups. Only 33 studies remained. Second, a study should report descriptive and/or inferential statistics of post-test scores, which are necessary for calculating post-test ESs. Only two studies did not meet the second criterion, so a total of 31 studies remained. Third, more specifically regarding control conditions, a study should have a conventional (or standard) instruction control group with which the effects of corpus use on L2 vocabulary learning in the experimental group could be compared. Another two studies were excluded because they had control groups that received no instruction at all. In the end, only 29 studies fully matched the suggested three criteria (see studies with an asterisk in the reference section).

4.3 Multilevel effect size calculation

Each post-test and follow-up (e.g., delayed post-test) ES was calculated according to unbiased *d* (Boulton & Cobb 2017). Also known as Hedges' *g* (Hedges and Olkin 1985; see Appendix 1.1 for the equations of the ES calculation), it provides more conservative calculations than Cohen's *d* does, particularly for small samples ($n < 50$; Hedges & Olkin 1985; Huber, 2013). It is calculated by multiplying the so-called bias correction factor ($J = 1 - [3 / \{4 \times df - 1\}]$; Hedges & Okin 1985) by Cohen's *d*. As mentioned earlier, I computed more than one ES per study when possible, so the constructed data set is a two-level model, as shown in Figure 1.2.[3] For example, I computed three post-test ESs for Kaur and Hegelhimer (2005; Study #10), and two post-test ESs for Supatranont (2005; Study #11). I found that Kaur and Hegelhimer not only tested students' in-depth vocabulary knowledge by using two different measurements, which were a cloze test and a sentence-building task, but also examined learners' productive use ability by counting the number of correct words used in students' writing. As a result, Kaur and Hegelhimer had three different measurements in their study, for which I generated three post-test ESs. In line with this approach, in Supatranont's (2005) study, there were two different dependent variables, including definition tests for precise knowledge and cloze tests for in-depth knowledge;

[3] It should be noted here that having multiple ESs for a study does not contribute to the overall results of meta-analysis multiple times more than a study with only one ES. When using multilevel modeling, it explicitly recognizes the membership of ESs (i.e., the nested structure; ESs are nested in studies) and that ESs from the same study are more similar to each other than ESs from other studies. In terms of different contributions of ES to the overall meta-analysis, the sample size (or the standard error) of each ES plays a significant role. Details are discussed in the data analysis section.

therefore, I computed two post-test ESs.

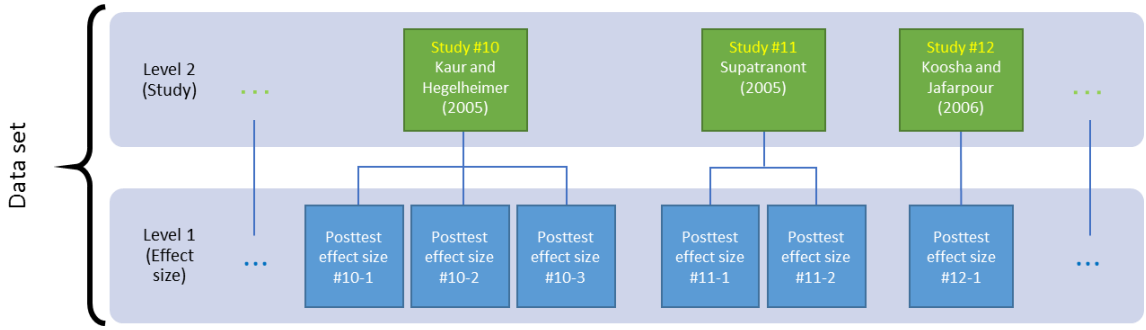

Figure 1.2. Structure of Data Set for Multilevel Meta-Analysis in Case of Post-test Effect Sizes.

In order to check the robustness of the results, I generated gain-score ESs based on pre-test and post-test scores whenever a study used a pre-test-post-test design instead of a post-test only (see Appendix 1.2). I did so because the pre-test-post-test design requires different equations for the ES calculation in order to include pre-test scores in the ES estimation. The post-test differences are often larger than the pre-test differences because of treatment effects (i.e., the treatment does not affect everyone the same way; hence there are likely to be larger post-test differences). Using the aforementioned between-groups equations that use post-test differences instead of pre-test differences would therefore underestimate the ES of pre-test-post-test design studies (see Morris, 2008; Plonsky & Oswald, 2012). The results of this approach showed that the newly generated gain-score ESs were slightly larger than the post-test ESs, though the difference between them was not statistically significant. Taken together, for the sake of robustness I used gain-score ESs over post-test ESs for the relevant studies.

4.4 Publication bias

Given the retrospective nature of a meta-analysis, the inclusion of previous empirical studies is limited to those *identifiable* by a literature search. First, for scholarly

journal articles, it has been said that studies with statistically significant findings are more likely to be submitted and accepted for publication and cited than studies with non-significant findings; such a bias might influence the results of meta-analysis. Furthermore, it has been said that small studies are more likely to show larger ESs; therefore, the ESs computed from a pool of small studies are sometimes contradicted by those in later large studies, which are not published as frequently and quickly as small-scale studies are (see Egger, Smith, Schneider, & Minder, 1997). For this reason, I assessed any possible publication bias (using STATA 14, see Appendix 1.3 for details) among the calculated ESs before carrying out full-scale data analyses. Taken together, the results of these checks (i.e., Funnel plot + Egger's test) revealed no significant publication bias among the calculated post-test ESs.

4.5 Coding procedure

To build a data set for the meta-analysis, I analyzed the previous meta-analyses to identify and develop a coding rubric for ES calculations and analyses of moderator variables. To this end, I decided to adapt the categories commonly used by other meta-analyses (e.g., publication, population, and treatment data) and to especially endorse Boulton and Cobb's (2017) coding strategy for analyzing treatment data (i.e., corpus use). In so doing, I found that five studies could be divided into multiple unique samples because they had separable sub-populations (Stevens, 1991; Cobb, 1999; Tongpoon, 2009; Rezaee, Marefat, & Saeedakhtar, 2015; Vyatkina, 2016). For example, Stevens (1991) had two separable treatment and control comparisons, so I could harvest two unique samples – Stevens 1991a and Stevens 1991b – from the study. Furthermore, in most of the studies, the unique samples included multiple ESs depending on the type of measurements.

In identifying and finalizing the moderator variables and their values, I had to be selective to ensure enough statistical power for regression analyses of moderator variables. For features that may have a bearing on the effectiveness of corpus use, I identified ten variables in the following two categories: (1) publication and population data and (2) treatment data. Table 1.1 shows the descriptive statistics of the identified variables and their values based on the coding scheme.

Table 1.1

Descriptive Statistics of Independent Variables

Variables	Post-test Effect Sizes			Follow-up Effect Sizes		
	Proportion of each value (no. of ES / total no. of ES)	Number of ESs (n = 77)	Number of Unique Samples (k = 38)	Proportion of each value (no. of ES / total no. of ES)	Number of ESs (n = 34)	Number of Unique Samples (k = 13)
1. Publication & Population Data						
(1) Publication date	Mean = 2009.55	Min = 1991	Max = 2016.4	Mean = 2011.94	Min = 2005.9	Max = 2015.8
(2) Publication type						
A. Journal article	73%	56	30	47%	16	8
B. PhD dissertation	23%	18	5	53%	18	5
C. Conference paper / Book chapter	4%	3	3	0%	0	0
(3) Region						
A. Asia	43%	33	14	65%	22	7
B. Middle East	29%	22	15	24%	8	5
C. Other	29%	22	9	12%	4	1
(4) L2 proficiency						
A. Low	26%	20	9	29%	10	3
B. Intermediate	61%	47	22	59%	20	9
C. High	6%	5	2	12%	4	1
D. Mixed	6%	5	5	0%	0	0
(5) Specialty						
A. Languages	17%	13	8	21%	7	3
B. Other	25%	19	11	3%	1	1
C. Mixed	58%	45	19	76%	26	9
2. Treatment Data						
(1) Interaction type						
A. Paper-based	30%	23	13	15%	5	3
B. CALL program	16%	12	7	0%	0	0
C. Concordancer	49%	38	15	76%	26	8
D. Mixed	5%	4	3	9%	3	2
(2) Corpus type						
A. Public corpus (e.g., COCA, BNC, Brown, OANC)	62%	48	22	91%	31	11
B. Local corpus (e.g., own, specialized, graded)	19%	15	9	9%	3	2
C. Pre-selected concordance lines (e.g., corpus-informed materials)	18%	14	7	0%	0	0
(3) L2 vocabulary dimension*						
A. Precise knowledge	29%	22	21	24%	8	8
B. In-depth knowledge	53%	41	32	50%	17	12
C. Productive use ability	18%	14	10	26%	9	5
(4) Training						
A. Not received	16%	12	7	0%	0	0
B. Received	84%	65	31	100%	34	13
(5) Duration						
A. Short (< 2 hours in total or only 1 session)	25%	19	8	12%	4	1
B. Medium (about 3 to 8 sessions)	43%	33	14	59%	20	7
C. Long (≥ 10 sessions in total)	32%	25	16	29%	10	5

Note. ES = effect size.

* The total number of unique samples for this variable is higher than 38 and 13 for post-test and follow-up effect sizes, respectively, because a unique sample may have multiple effect sizes to measure different dimensions of L2 vocabulary knowledge.

4.5.1 Publication and population data

Above all, for the publication data, two variables were coded. First, the date of publication (month and year) was coded. Second, the types of publication were coded (journal article, PhD dissertation, and conference paper or book chapter). This variable shows whether a study was peer-reviewed for the publication, in accordance with the types of publication. I found that most of the post-test ESs came from journal articles (56 post-test ESs, 73%); however, PhD dissertations had a higher proportion of follow-up ESs (18 follow-up ESs, 53%) than did other publication types.

For the population data, three variables were chosen. First, participants' countries of origin were checked. On the initial assumption that the participants have different L1s across studies, I believed that including this difference in the equations would contribute to more reliable results. As the second variable, and for the same reason, learners' overall L2 proficiency was coded. For example, any reported levels of L2 proficiency can be noted, as found in the following official tests, which were frequently available:

- Test of English as a Foreign Language (TOEFL) and Test of English for International Communication (TOEIC) developed by Educational Testing Service (ETS)
- International English Language Testing System (IELTS) and Cambridge English: Preliminary (PET) developed by Cambridge English Language Assessment.

Perhaps because of the aforementioned barriers of DDL, I found rather few ESs from studies of learners with low L2 proficiency (20 post-test ESs, 26%; 10 follow-up ESs, 29%).

For the third variable, on the initial assumption that students specializing in L2 learning would excel in corpus use, I checked the students' specialties, which included those related to language (e.g., English majors), other specialties unrelated to language

(e.g., Engineering, Medicine, Social Science), or mixed specialties (studies that included both language and non-language majors).

4.5.2 Treatment data

Five variables were coded for treatment data (i.e., corpus use) in order to include different ways of implementing corpus use in various contexts. First, assuming that using new technology (e.g., concordancing software) affects learners' experience with corpus materials (Boulton, 2010), I checked how students interacted with corpus data (i.e., interaction type), including paper-based activity, CALL program (e.g., Sketch Engine; Kilgarriff et al., 2014), concordancer-based activity, or a combination thereof (e.g., paper-based + concordancer-based activities). I found that a majority of ESs came from studies that provided computer-based, hands-on concordancing activity (38 post-test ESs, 49%; 26 follow-up ESs, 76%).

Second, different types of corpus data were noted as well. Although most ESs came from studies that used pre-established public corpora (48 post-test ESs, 62%; 31 follow-up ESs, 91%), such as COCA, BNC, Brown, and OANC, I could identify situations where researchers created materials using other collections of language data, such as local corpora of authentic texts, students' textbooks, and graded readers, or they opted to provide concordance lines that were selected to be more suitable for learners (e.g., Frankenberg-Garcia 2012, 2014).

Third, different dimensions of L2 vocabulary knowledge were checked, and I coded the aforementioned three values: precise knowledge, in-depth knowledge, and productive use ability. In essence, how a study measures participants' learning outcomes largely determines the target dimension of L2 vocabulary knowledge. Precise knowledge is

concerned with recall or recognition of the meaning of a target lexical item, which is arguably the most frequently investigated dimension in L2 vocabulary learning research. In-depth knowledge, on the other hand, is measured by investigating participants' knowledge about syntagmatic (e.g., collocation) and paradigmatic (e.g., synonym) relationships between a target lexical item and others. Productive use ability is measured by asking participants to produce sentences containing a target lexical item. I found that about half of ESs came from studies that investigated the in-depth dimension of L2 vocabulary knowledge (41 post-test ESs, 53%; 17 follow-up ESs, 50%).

Fourth, whether participants were trained in corpus use prior to the vocabulary treatment was coded. Given that most of the collected studies focused on the extensive use of concordancer and concordance lines, most ESs came from studies that included a training opportunity (65 post-test ESs, 84%; 34 follow-up ESs, 100%).

Last, the duration of an intervention was taken into consideration. In order to generate comparable criteria, I decided to endorse Boulton and Cobb's (2017) criterion to measure the duration of the treatment: short (less than two hours in total or only one session), medium (about three to nine sessions), and long (ten sessions or more).

To ensure that the data set was generated reliably and accurately, a second rater was trained in the protocol based on the created rubric for the coding procedure. He passed a reliability test that consisted of all codes from eight selected studies (30% of the total number of studies) by achieving 100% agreement with the first rater. After the completion of the entire coding procedure, I found that the inter-coder agreement was perfect.

4.6 Data analysis plan

To answer the research questions about the overall mean ES as well as the impacts of moderator variables, I used STATA 14 (with *meglm* command) to conduct a multilevel regression analysis with the post-test / follow-up ESs as the dependent variables. All the results of multilevel regression analyses were replicated in HLM 7 to confirm their robustness. I used a multilevel model because the unit of data analysis is each ES (level 1), but they are nested in each unique sample (level 2); this hierarchy of the data structure does not satisfy the independence assumption of conventional regression analysis (e.g., ordinary linear square regression; OLS). In other words, a problem comes up when all the observations are simply pooled in the conventional analysis without taking into consideration the nested design because the ESs nested in the same study would not be independent of each other (Raudenbush & Bryk, 2002).

In multilevel models, on the other hand, I can avoid this issue because the analysis retains the nested design of each observation by distinguishing the level 1 variance (i.e., the variance between the ESs within a unique sample) and the level 2 variance (i.e., the variance between the unique samples; see Appendix 1.4 for an example equation of multilevel model regression analysis). In this way, multilevel model analysis has one regression line for each unique sample and they are parallel to the overall regression line (see Woltman et al., 2012 for the details of multilevel modeling). Furthermore, using multilevel regression analysis with the level 1 units allowed me not only to compare and contrast these mean estimates to further analyze whether they are statistically different from each other, but also to compute the adjusted mean ESs (i.e., predictive margins) for

each value of the moderators after controlling for other moderators (or holding others constant).

Based on the discussion above, I first ran unconditional (i.e., no covariate) multilevel models in order to compute the mean ESs for post-tests as well as follow-up tests. The multilevel models used in the meta-analysis differ from common multilevel modeling, because my model is a *variance-known* model. In other words, I already know the level 1 variance because the dependent variable consists of the computed ESs, and I know their sampling errors: the variance (the square of the standard errors) of the ESs (see Appendix 1.4 for the equations of multilevel regression analyses). Following Hox et al.'s (2010) theoretical suggestion and Pratt et al.'s (2016) practical guideline, I built a variance-known multilevel model by including the variance of the ESs as the level 1 variance component. In this way, I could assign different weights to each unique sample according to the precision of the ESs within unique samples, given that the level 1 variance component is used to estimate the level 2 variance (Borenstein, Hedges, Higgins, & Rothstein, 2009).

In addition, for the moderator analysis, I ran two more regression analyses with the variables of interest included on the right-hand side of the equations, one for the publication and population data and the other for the treatment data (see Appendix 1.4 for the equations). In doing so, I abided by the following two rules of thumb for regression analyses to ensure the robustness. First, as mentioned above, I ran two separate regression models for the moderator estimation (i.e., one for publication and population data, the other for treatment data) in order to have each model include a maximum of seven (number of post-test ESs / 10 = 7.7) and three (number of follow-up ESs / 10 = 3.4) predictors in the equations for the post-test ESs (n = 77) and the follow-up ESs (n = 34),

respectively. This is to follow the one-in-ten rule—one of the most conservative perspectives to avoid risk of lack of degree of freedom and thus overfitting the model (see Vittinghoff & McCulloch, 2007), which would mislead and bias inferential statistics of the model. For the follow-up ESs, however, I did not choose certain variables, but found that the variables excluded in the analyses could not be included in the model because the variables had either a multicollinearity issue or small samples.[4]

Second, in order to confirm the findings of multilevel regression analyses, I conducted a clustered regression analysis with the Huber-White corrected estimator of variance and further compared its results with those of multilevel regression analyses (see Appendix 1.5 for the full models). Though this approach does not fully incorporate the hierarchical (i.e., multilevel) data structure, it produces the best linear unbiased estimates that are robust to heteroscedasticity; thus, this approach yields results similar to those of multilevel models (see Angeles & Mroz, 2001; Wooldridge, 2010). Overall, the results of this approach showed that multilevel regression analyses provide similar estimates (i.e., similar coefficients, standard errors, and *p* values), confirming the robustness of the findings.

5. Results

An overview of the post-test ESs by publication year is given in Figure 1.3, which includes bubbles for each post-test ES. The bubbles' size is proportionate to the precision of each post-test ES; therefore, an ES with a high precision (i.e., low standard error) or a large

[4] For example, the value of Other in the Region variable and the value of High in the L2 proficiency variable in the follow-up ES section have identical observations (see Table 1.3); this may cause a multicollinearity issue, so I omitted the variable in the equation to avoid this. In regard to the small sample issue, I can see that the Training variable in the follow-up ES section has only one value (see Table 1.4), and the variable does not have any reference category.

sample size has a bigger bubble. This figure reveals that the studies included began in the 1990s and that a relatively large concentration emerged in the 2010s. In particular, about 60% of the included studies (i.e., 17 studies out of 29) were conducted in the 2010s, perhaps because the broader technical access, mentioned in the Introduction, facilitated dissemination and use of hands-on corpus-related programs in a wide range of pedagogical contexts (Godwin-Jones 2001b; Reppen 2010; Lee and Lee 2015). Nevertheless, I found no association between the magnitude of post-test ESs and publication date (b = .01, SE = .01, p = .24).

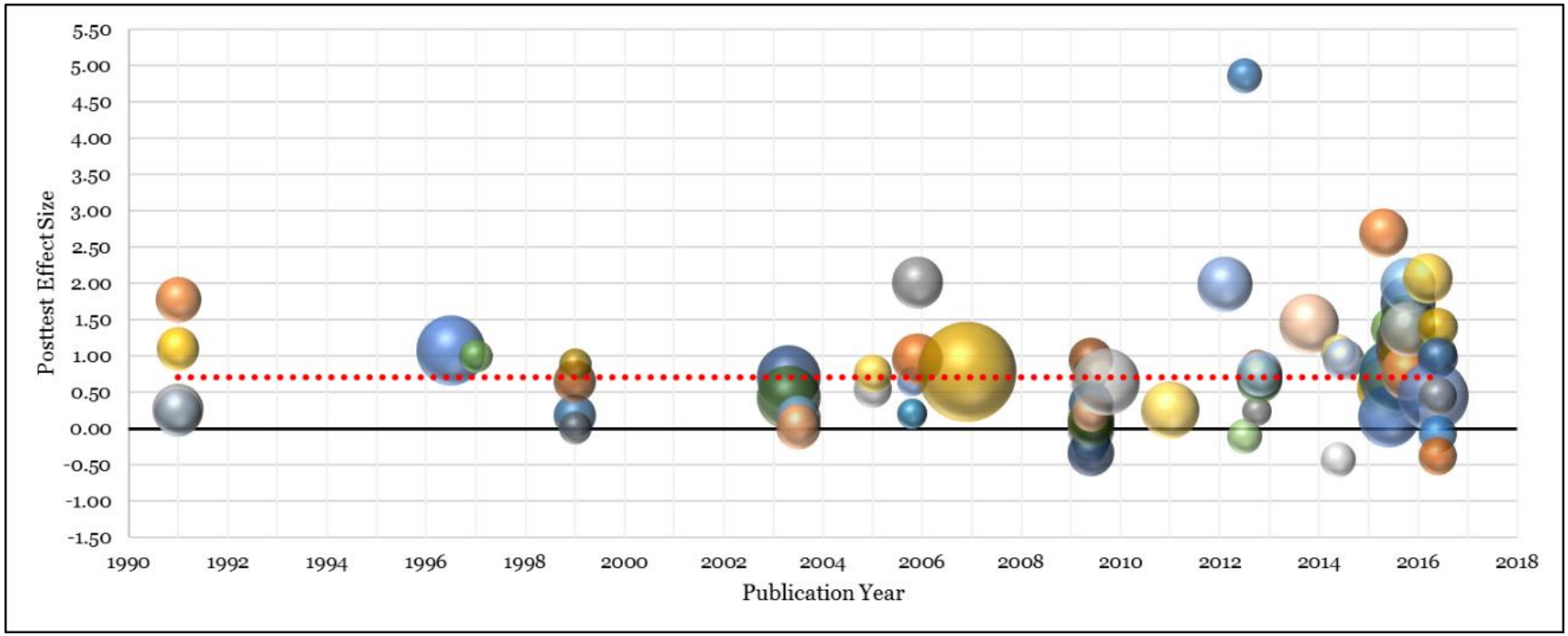

Figure 1.3. Post-test Effect Sizes by Year of Publication.

Note. Bubbles represent each post-test ES. Bubble size is proportionate to the precision of each post-test ES; an ES with higher precision is indicated by a larger bubble.

5.1 Research question 1: How effective is corpus use in improving L2 vocabulary learning?

Table 1.2 presents the weighted mean ESs as well as the variance components as the mean ES estimates. For the post-test ESs, I found a medium-sized effect of corpus use on L2 vocabulary learning (d = .74, SE = .09, p < .001; 77 post-test ESs coming from 38 unique samples), indicating that its impact is in the top 50% of L2 learning tools and instruction in the field of applied linguistics based on the benchmark suggested by Plonsky and Oswald (2014). The variance components of this estimate indicated that, among the model's total variation, about 60% came from within unique samples (i.e., $\sigma_{level\ 1}^2$ = .18, SE = .07, p < .001) and the remaining 40% came from between unique samples (i.e., $\sigma_{level\ 2}^2$ = .12, SE = .07, p < .001). For the follow-up ESs, I found that the positive medium-sized effect of corpus use was long-term (d = .64, SE = .17, p < .001; 34 follow-up ESs coming from 13 unique samples). Furthermore, having both the post-test and follow-up ESs the same size (i.e., medium size; g > .6) is noteworthy, indicating that corpus use as a learning tool is effective in enhancing L2 vocabulary long-term retention (see Min, 2008 for a review on L2 vocabulary learning and retention). For the variance components of the follow-up ES estimates, about 37% of the model's total variation came from level 1 and 63% from level 2. As a result, the findings confirmed that corpus use works as an active student-centered L2 vocabulary learning approach. Furthermore, the estimates of the variance components of the models revealed that there was significant variation in the effect size estimates both within unique samples (level 1) and between unique samples (level 2), thus suggesting that the use of multilevel modeling was necessary.

Table 1.2

Mean Effect Size Estimates

Effect Size Estimates		Post-test Effect Sizes	Follow-up Effect Sizes
1. Weighted mean effect size		.74*** (.09), 95% CI [.57 ~ .91]	.64*** (.17), 95% CI [.31 ~ .97]
2. Variance components	Level 1: Effect size level	.18*** (.07)	.15*** (.08)
	Level 2: Study level	.12*** (.07)	.26*** (.14)
Number of ESs (*n*)		77	34
Number of Unique Samples (*k*)		38	13

Note. CI = confidence interval; ES = effect size.

*** $p < .001$

In addition to the multilevel ESs, I computed a single average ES per study and confirmed both the short-term and the long-term medium-sized effect of corpus use (38 post-test ESs for unique samples, $d = .78$, $SE = .08$, $p < .001$; 13 follow-up ESs for unique samples, $d = .70$, $SE = .15$, $p < .001$). Also, I present a forest plot of these average ESs across the studies in accordance with the standard meta-analysis practice in the field (see Appendix 1.1 for the equations and procedures of the ES calculation at study level, and see Appendix 1.6 for the ES estimates and the forest plots).

5.2 Research question 2: What are the moderators that influence the magnitude of the effectiveness of corpus use?

Tables 1.3 and 1.4 describe the results of both simple and multiple regression analyses for the identified nine moderators from the publication, population, and treatment data. In particular, the columns for the simple regression analyses show the descriptive statistics for each moderator, such as numbers of ESs and unique samples, and weighted means for each value of the moderator. The columns of multiple regression analyses present the inferential statistics (i.e., the adjusted means and their contrasts) to describe how each value of the moderator influences the magnitude of the effectiveness of corpus use along with the other moderators.

Table 1.3

Descriptive and Inferential Statistics of Moderator Variables I (Publication and Population Data)

Moderator Variables	Post-test Effect Sizes					Follow-up Effect Sizes				
	Simple Regression			Multiple Regression		Simple Regression			Multiple Regression	
	n	*k*	Weighted means	Adjusted means	Contrasts	*n*	*k*	Weighted means	Adjusted means	Contrasts
1. Publication & Population Data										
(1) Publication type										
A. Journal article	56	30	.79*** (.09)	.74*** (.08)	vs. B: .31 (.21)	16	8	1.01*** (.14)	.99*** (.14)	-
B. PhD dissertation	18	5	.29 (.17)	.42* (.18)	vs. C: -.66 (.48)	18	5	.06 (.16)	.08 (.16)	vs. B: -.91*** (.23)
C. Other[a]	3	3	1.29*** (.34)	1.08* (.45)	vs. A: .34 (.46)	-	-	-	-	-
(2) Region										
A. Asia	33	14	.49*** (.12)	.53*** (.11)	vs. B: -.52** (.19)	22	7	.26 (.16)		
B. Middle East	22	15	1.06*** (.13)	1.05*** (.13)	vs. C: .52** (.20)	8	5	1.09*** (.22)		Omitted[d]
C. Other[b]	22	9	.68*** (.15)	.53*** (.15)	vs. A: -.00 (.20)	4	1	1.13** (.38)		
(3) L2 proficiency										
A. Low	20	9	.50** (.16)	.47*** (.13)	vs. B: -.23 (.15)	10	3	.18 (.29)	.29 (.18)	vs. B: -.28 (.23)
B. Intermediate	47	22	.71*** (.10)	.69*** (.09)	vs. D: -.05 (.38)	20	9	.73*** (.18)	.57***(.13)	vs. C: -.23 (.40)
C. High	5	2	1.11*** (.32)	1.27*** (.31)	vs. A: .80* (.36)	4	1	1.13* (.48)	.79*(.37)	vs. A: .51 (.45)
D. Mixed	5	5	1.21*** (.26)	.74* (.35)	vs. C: -.52 (.46)	-	-	-	-	-
(4) Specialty										
A. Languages	13	8	.90*** (.19)	.62** (.20)	vs. B: .08 (.24)	7	3	1.03** (.33)	.36 (.29)	vs. B: .13 (.56)
B. Other[c]	19	11	.81*** (.17)	.54*** (.14)	vs. C: -.22 (.18)	1	1	.77 (.65)	.23 (.49)	vs. C: -.33 (.51)
C. Mixed	45	19	.64*** (.12)	.75*** (.09)	vs. A: .14 (.24)	26	9	.50** (.18)	.56*** (.12)	vs. A: .20 (.34)
Number of ESs (*n*)	77					34				
Number of Unique Samples (*k*)	38					13				

Note. Standard errors are in parentheses. A dash in a cell indicates that the variable does not have relevance to the cell. ES = effect size.

[a] e.g., conference paper, book chapter

[b] e.g., Europe, US

[c] e.g., Engineering, Medical, Social Science

[d] The value of Other in the variable and the value of High in the L2 proficiency variable in the follow-up ES section have identical observations (i.e., $n = 4$, $k = 1$), and this may cause a multicollinearity issue, so I omitted the variable in the equation to avoid this.

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

5.2.1 Publication and population data

In Table 1.3, I have the descriptive and inferential statistics of the four variables in the publication and population data, including (1) publication type, (2) region, (3) L2 proficiency, and (4) specialty. First, for the publication type, the simple regression for the post ESs showed a medium effect for journal articles and a large effect for other publication types, such as conference papers and book chapters. However, there was a negligible marginal effect for PhD dissertations. Also, the follow-up ESs had a similar pattern, where I found a large effect for journal articles and a negligible effect for PhD dissertations. After controlling for other variables in the publication and population data (or keeping other variables at their averages), however, multiple regression analyses showed that there were no statistically significant differences for the post-test ESs, indicating that the publication types do not have different effect sizes of corpus use on the assumption that other publication and population variables are the same across the publication types. Nevertheless, for the follow-up ESs, I found a large, statistically significant difference between PhD dissertations and journal articles (*ES difference*; $d = .91$, $SE = .23$, $p < .001$). Similar findings were reported in one of the previous meta-analysis studies (Boulton & Cobb, 2017), and I agree with their interpretation that among many possible reasons this finding may indicate both submission bias (i.e., researchers are more likely to submit studies with higher impact to a journal) and acceptance bias (i.e., journals are more likely to accept studies with higher impact). A similar issue was discussed earlier in the publication bias section.

Second, the post-test ESs of the region variable indicated that the L2 learners from the Middle East had a mean ES of 1.06 ($SE = .13$, $p < .001$), which is higher than the ESs of

the learners in either Asia (d = .49, SE = .12, p < .001) or Europe and the US (d = .68, SE = .15, p < .001). The adjusted means and their contrasts from multiple regression analyses showed that the higher performance of the Middle Eastern learners is statistically significant when compared to their peers (*ES difference*; d = .52, SE = .20, p < .01). This finding corresponds to that of Boulton and Cobb (2017), who found this result to be counter-intuitive given the similar deductive-oriented L2 learning cultures both in Asia and the Middle East. Moreover, they suggested that the lower ESs for Europe and the US may be due to the similarity between DDL and traditional teaching in their regions.

Third, the L2 proficiency variable showed that, for the post-test ESs, corpus use had a small effect on learners with low L2 proficiency (d = .50, SE = .16, p < .01), but had either medium or large effects for those with intermediate (d = .71, SE = .10, p < .001) or high L2 proficiency (d = 1.11, SE = .32, p < .001). When holding other variables constant, I found a medium, statistically significant difference between the learners with low proficiency and those with high proficiency (*ES difference;* d = .80, SE = .36, p < .05). For the groups of learners with mixed L2 proficiency levels, I found that multiple regression analyses substantially lowered their adjusted means to .74 (SE = .35, p < .05). Although the simple regression indicated a large effect for this group of learners (d = 1.21, SE = .26, p < .001), I cautiously suggest (because it is based on very limited data; there were only 5 ESs coming from 5 unique samples) that by including covariates in the equation, multiple regression analyses produced more reliable results than the simple regression. For the follow-up ESs, I found that the small effect of corpus use became negligible in the long term for low proficiency levels (d = .18, SE = .29, p > .05), but positive effects on L2 vocabulary learning for the learners with intermediate and high L2 proficiency remained (although from very

limited data, as there were only 4 ESs coming from 1 unique samples for high proficiency levels). Taken together, I found that corpus use could be more effective for intermediate and high-proficiency learners than for low-proficiency learners.

Fourth, for the specialty variable, I found that the learners performed well in their corpus-based instruction / activities no matter what their specialties were. The simple regressions for both the post-test and follow-up ESs show that students specializing in the language-related disciplines tended to have slightly higher post-test (d = .90, SE = .19, p < .001) as well as follow-up ESs (d = 1.03, SE = .33, p < .01); however, after keeping other variables at their averages, I could not find any statistically significant difference between the specialties. This finding did not agree with my initial assumption, but it could be explained by the dominant role of the L2 proficiency level variable included in multiple regression analyses. In other words, I can assume that learners' specialty does not have a large effect on L2 vocabulary learning via corpus use once I control for L2 proficiency.

Table 1.4

Descriptive and Inferential Statistics of Moderator Variables II (Treatment Data)

Moderator Variables	Post-test Effect Sizes					Follow-up Effect Sizes				
	Simple Regression			Multiple Regression		Simple Regression			Multiple Regression	
	n	*k*	Weighted means	Adjusted means	Contrasts	*n*	*k*	Weighted means	Adjusted means	Contrasts
2. Treatment Data										
(1) Interaction type										
A. Paper-based	23	13	.67*** (.15)	.55*** (.14)	vs. D: -.75* (.34)	5	3	.89** (.34)	.70*** (.19)	vs. C: .31 (.22)
B. CALL program	12	7	.73*** (.21)	.70** (.23)	vs. A: .15 (.26)	-	-	-	-	-
C. Concordancer	38	15	.71*** (.13)	.72*** (.14)	vs. B: 0.02 (.31)	26	8	.46* (.19)	.39*** (.09)	vs. D: -.72* (.29)
D. Mixed	4	3	1.24*** (.32)	1.30*** (.32)	vs. C: .58 (.38)	3	2	1.06* (.43)	1.11*** (.27)	vs. A: -.31 (.22)
(2) Corpus type										
A. Public corpus[a]	48	22	.72*** (.11)	.65*** (.11)	vs. B: 0.05 (.25)	31	11	.57** (.17)		
B. Local corpus[b]	15	9	.77*** (.19)	.59** (.20)	vs. C: -.39 (.36)	3	2	1.06* (.44)	Omitted[f]	
C. Pre-selected[c]	14	7	.78*** (.21)	.98*** (.28)	vs. A: .34 (.34)	-	-	-		
(3) L2 vocabulary dimension[d]										
A. Precise knowledge	22	21	.46*** (.13)	.40** (.13)	vs. B: -.51** (.16)	8	8	.38 (.20)	.29 (.17)	vs. B: -.48* (.23)
B. In-depth knowledge	41	32	.92*** (.10)	.91*** (.10)	vs. C: .36 (.20)	17	12	.88*** (.16)	.77*** (.13)	vs. C: .59* (.23)
C. Productive use ability	14	10	.53** (.19)	.55** (.18)	vs. A: .15 (.21)	9	5	.20 (.24)	.18 (.17)	vs. A: -.11 (.24)
(4) Training										
A. Not received	12	7	.76*** (.21)	.58** (.22)		-	-		Omitted[g]	
B. Received	65	31	.73*** (.10)	.72*** (.08)	vs. A: .14 (.24)	34	13	.64*** (.17)		
(5) Duration[e]										
A. Short	19	8	.71*** (.17)	.68** (.21)	vs. B: .13 (.28)	4	1	1.13*** (.27)	1.24*** (.22)	vs. B: 1.01*** (.25)
B. Medium	33	14	.51*** (.13)	.55*** (.13)	vs. C: -.35 (.18)	20	7	.10 (.12)	.23* (.11)	vs. C: -.52* (.21)
C. Long	25	16	.97*** (.13)	.90*** (.14)	vs. A: .22 (.29)	10	5	1.08*** (.16)	.75*** (.16)	vs. A: -.49 (.28)
Number of ESs (*n*)				77					34	
Number of Unique Samples (*k*)				38					13	

Note. Standard errors are in parentheses. A dash in a cell indicates that the variable does not have relevance to the cell. ES = effect size.

[a] e.g., COCA, BNC, Brown, OANC

[b] e.g., own, specialized, graded

[c] e.g., providing carefully selected or syntactically modified concordance lines; corpus-informed materials

[d] The total number of unique samples for this variable is higher than 38 and 13 for post-test and follow-up effect sizes, respectively, because a unique sample may have multiple effect sizes to measure different dimensions of L2 vocabulary knowledge.

[e] Short: less than 2 hours in total or only 1 session; Medium: about 3 to 9 sessions; and Long: 10 sessions or more in total

[f] The value of Local corpus in the variable and the value of Mixed in the Interaction type variable in the follow-up ES section have identical observations (i.e., $n = 3$, $k = 2$), and this may cause a multicollinearity issue, so I omitted the variable in the equation to avoid this.

[g] The variable has only one value and does not have any reference category; therefore, I omitted the variable in the equation.

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

5.2.2 Treatment data

In Table 1.4, I present the descriptive and inferential statistics of the five variables for the treatment data, including (1) interaction type, (2) corpus type, (3) L2 vocabulary, (4) training, and (5) duration. First, for the interaction type, I found that paper-based, CALL program-based, and concordancer-based activities or instruction all had a medium effect on improving L2 vocabulary knowledge (d = .67, SE = .15, p < .001; d = .73, SE = .21, p < .001; d = .71, SE = .13, p < .001, respectively), and that using a combination of these types (e.g., paper-based activity + concordancer) had the largest ES (d = 1.24, SE = .32, p < .001). Multiple regression analyses for the post-test ESs revealed that there was a medium-sized difference between using only paper-based and using paper-based + additional interaction types (*ES difference;* d = .75, SE = .34, p < .05). Similarly, multiple regression analyses for the follow-up ESs revealed that using multiple interaction types for corpus use produced superior results in the long term as well. Although the analyses of the follow-up ESs appeared to favor the paper-based type of corpus use over the concordancer-based type, the ES difference was not significant (*ES difference;* d = .31, SE = .22, p > .05).

Second, for the post-test ESs for the second moderator in the treatment data, I found medium-sized effects of corpus use on L2 vocabulary learning for all the different corpus types. According to the adjusted means from multiple regression analyses, I found that careful selection of concordance lines by teachers or researchers had a large impact on improving L2 vocabulary knowledge (d = .98, SE = .28, p < .001), whereas using either public or local corpora had a medium-sized ES (d = .65, SE = .11, p < .001; d = .59, SE = .20, p < .01; respectively). For the follow-up ESs, the positive effects of using either public or local

corpora remained long-term. In regard to the remaining coefficients, their small sample sizes did not allow me to explain further.

Third, the analyses of the post-test ESs for the L2 vocabulary variable indicated that corpus use improved all three dimensions of L2 vocabulary knowledge. In particular, although corpus use had a large effect on improving in-depth knowledge (d = .92, SE = .10, p < .001), it had only a small effect for precise knowledge and productive use ability (d = .46, SE = .13, p < .001; d = .53, SE = .19, p < .01; respectively). For the ES difference between the dimensions, the results of multiple regression analyses revealed that corpus use is more effective for expanding in-depth knowledge of L2 vocabulary than for increasing learners' precise knowledge of L2 vocabulary (*ES difference;* d = .51, SE = .16, p < .01). For the follow-up ESs, corpus use was effective only for in-depth knowledge (d = .88, SE = .16, p < .001), as its effects on improving the other dimensions became negligible in the long term. In brief, this pattern corresponds to the findings of a recent meta-analysis on computer-mediated textual glosses (Abraham, 2008), where learners benefited from glosses less in terms of productive lexical knowledge than in terms of receptive knowledge (i.e., precise + in-depth knowledge). It also accords with the general structure of vocabulary knowledge among L2 learners, with receptive knowledge being greater than productive knowledge (Laufer & Paribakht, 1998). It is interesting to note that corpora use led to greater gains in in-depth knowledge than in precise (i.e., definitional) knowledge, and a possible explanation for this finding may be related to the nature of corpus use that offers multiple samples of target vocabulary. This finding is one of the unique contributions of Chapter 1 and I include an in-depth discussion in the following section.

Fourth, I found that the learners performed almost equally well for their L2 vocabulary learning based on corpus use whether they had had a training opportunity or not. Although multiple regression analyses revealed that receiving training could have a slightly larger effect than not receiving any training when keeping other treatment-related variables at their averages, the ES difference was negligible and was not statistically significant (*ES difference; d* = .14, *SE* = .24, $p > .05$). All the studies that had follow-up ESs provided training opportunities, so no further investigation was conducted for this variable.

Fifth, for the duration of instruction, I found that the interventions with ten sessions or more had larger ESs than interventions with less than ten sessions. However, there was no statistically significant difference across the different lengths of the intervention, according to the results of multiple regression analyses. Further exploration of the follow-up ESs did not give us a clear picture of the impact of the duration of instruction, again because the samples were small.

6. Discussion and conclusion

A meta-analysis was conducted to investigate the overall effect of corpus use on L2 vocabulary learning and to identify moderators that may influence the effectiveness of corpus use. Based on the calculated 77 post-test ESs from 38 unique samples and 34 follow-up ESs from 13 unique samples, the meta-analysis shows a medium-sized effect on L2 vocabulary learning, with the greatest benefits for promoting in-depth knowledge to learners who have at least intermediate L2 proficiency. Corpus use was also more effective when the concordance lines were purposely selected and provided and when learning materials were given along with hands-on corpus-use opportunities. Moreover, I found that

corpus use is still effective even without prior training and remains effective regardless of the corpus type or the length of the intervention. In the following, I discuss the findings in terms of corpus use for constructive L2 vocabulary learning and different dimensions of L2 vocabulary knowledge and also discuss evidence-based considerations in implementing corpus use.

6.1 Overview: Corpus use for constructive L2 vocabulary learning

Overall, an overall medium effect in multilevel meta-analyses of corpus use for both the short (d = .74) and the long term (d = .64) is an important indication of the self-driven construction of L2 vocabulary learning. Given that the studies included in the meta-analysis were mostly concerned with 'direct' (as opposed to 'indirect') use of corpora in language teaching (Leech, 1997), in which learners gain direct access to corpus materials, I assume that the learners in the sampled studies apparently explored the data independently for their linguistic inquiry and language learning. This process, also known as DDL (Johns, 1991), has been at the heart of corpus use in language teaching (Leech, 1997) and is associated with important interrelated educational constructs, such as learner autonomy (Gavioli, 2009), motivation (Curado Fuentes, 2015), and discovery learning/constructivism (Flowerdew, 2015).

For this reason, the fact that learners construct or discover L2 vocabulary knowledge by using a corpus-based approach has important implications for language teaching. As Bernardini (2004) highlights, discovery learning "encourages learners to follow their own interests" (p. 23), and learners may pave different learning paths even with the same learning materials. In other words, they independently and individually construct their own knowledge. Bernardini further notes that this change in the learning

approach influences the previous roles of teacher and learners as well as those of material developers and curriculum designers. In view of constructivism, although the findings of this meta-analysis point to the value of a corpus-based approach, *how* learners construct their knowledge deserves more attention. As Flowerdew (2015) suggests, learners are expected to tackle corpus materials with "higher order cognitive skills" (p. 18), which would interact with the learners' current L2 proficiency level and other moderators (see below for the discussion on this issue). This very complicated process would be better explored through a more qualitative lens; such an investigation, integrated with a quantitative approach, could help teachers and researchers contemplate more effective ways to develop individualized instruction.

6.2 Corpus use and dimensions of L2 vocabulary knowledge

Furthermore, my focus on specific dimensions of L2 vocabulary knowledge led us to understand that several moderators may come into play in the effectiveness of corpus use in L2 vocabulary learning. The most salient and intriguing finding from this approach is related to the dimensions of L2 vocabulary knowledge, with in-depth properties of target lexical items being learnt best. A possible explanation for this finding may lie in the KWIC format, which offers multiple contextual instances of vocabulary (Johns, 1994). This feature helps learners refer to multiple samples of how a target lexical item is used differently in various contexts. Learners would then be able to understand in-depth properties of a target lexical item, such as paradigmatic and syntagmatic relationships between a target lexical item and others (Wolter, 2006), which accords with predictions made by proponents of corpus use in vocabulary teaching and the DDL approach (e.g., Johns, 1991; Kita & Ogata, 1997; Cobb, 1999; Hill, 2000).

Moreover, this finding fits in with views of researchers, such as Bernardini (2004) or Flowerdew (2015), who introduced diverse ways of using DDL in L2 instruction. Although they do not specifically refer to the vocabulary knowledge framework like the one used in this study (i.e., Henriksen, 1999), it can be inferred from their use of expressions such as 'collocates' and 'synonyms and/or antonyms', that in-depth knowledge is expected to be developed by learners' exploration of concordance lines. This is a promising finding for corpus use as an L2 vocabulary teaching approach, considering that this dimension of knowledge is not frequently dealt with in instructed L2 vocabulary research presumably because the primary focus in this literature is on establishing the link between form and meaning of target vocabulary (Schmitt, 2008). On this point, Schmitt (2008, p. 334) suggests that an explicit approach be used first to establish this "form-meaning link", and "exposure approach' could later be used to develop more 'contextual knowledge" (i.e., in-depth knowledge). Although he did not elaborate on the details of this approach, I suggest, based on the findings, that corpus use could be an appropriate mechanism for this.

6.3 Evidence-based considerations for corpus use in L2 vocabulary learning

Given the considerations about implementing corpus use discussed above, such as the suitability of language data and mastery of corpus consultation skills, the findings of the meta-analysis offer some evidence-based suggestions on how to effectively implement DDL in L2 learning. Above all, one noticeable factor that influences the effectiveness of corpus use is whether the language data provided in corpus use were suitable for the learners. In the initial stages of adapting corpus materials to classroom contexts, the belief among teachers that low-proficiency learners cannot handle corpus data well for learning appears to have impeded its wide application (Leńko-Szymańska & Boulton, 2015). The finding that

learners with low L2 proficiency do not benefit as much from corpus use confirms this belief. As Chujo, Oghigian, and Akasegawa (2015, p. 111) note, "if DDL is to be considered for low-proficiency learners, there is a need to rethink available corpora".

The results of the moderator analysis also pointed to the importance of some methodological features that could contribute to learners' engagement with the given corpus materials for their independent discovery learning. For example, I found that providing L2 learners with pre-selected, comprehensible concordance lines appears more effective in supporting their corpus-based activities, which corresponds to previous empirical findings (e.g., Frankenberg-Garcia, 2012, 2014). Also, this is in line with recommendations about the use of corpus materials that are more finely tuned to learners' L2 proficiency level (Allan, 2009; Poole, 2012; Chujo et al., 2015) and about learner-friendly concordancer software specifically designed for L2 learning (Chujo et al., 2015; Lee et al., 2015). I also verified that both paper-based and computer-based corpus activities have medium-size effects on improving L2 vocabulary learning, thus suggesting that the way corpus materials are presented to learners may not be a critical variable. In addition, confirming Boulton's (2009) finding that DDL can be effective even for untrained learners, the results showed that corpus-based activity without any prior training is just as effective as when training was received. Also, the finding that the duration of corpus-based activity was not associated with the effectiveness of corpus use corresponds to previous findings (Boulton & Cobb, 2017).

Taken together, I believe that learners' L2 proficiency plays the most significant role in DDL activities, though the evidence I found from the meta-analysis is based on limited data (there were only 5 ESs from 2 unique samples for high proficiency levels). Given that

independent discovery learning requires substantial levels of linguistic knowledge and inference on the part of learners (Johns, 1991), the importance of learners' L2 proficiency is expected. It is true, however, that some recent studies (e.g., Vyatkina, 2016) revealed that DDL could also be effective for beginning learners, and the findings of the meta-analysis support this argument. Nevertheless, considering the different ESs for each L2 proficiency level—the small effect for beginners and medium or large effects for more advanced learners—I found that more proficient learners might benefit more extensively. For example, the ES difference between the low and high proficiency levels for the post-test ESs is around .80; that is, the mean of the group of students at the high L2 level is higher than about the 79th percentile of the group of students with low L2 proficiency (Cohen, 1988). In other words, less than 21% of low-proficiency students who participated in the included studies could perform as well as or better than a student with high L2 proficiency—all other variables being equal. For this reason, I believe that the difference between the adjusted means of L2 proficiency levels has to be approached differently from the other estimated moderators. As Flowerdew (2015, p. 18) suggests, "constructivism [which theoretically underlies the DDL approach] may not be ideal for all students" because of individual learner differences. I think that L2 proficiency may be an important factor among the differences that influence the effectiveness of corpus use for L2 vocabulary learning and deserves more attention in future research.

6.4 Limitations and suggestions

Despite the contributory findings, this study is not without its limitations. The sample size was not large enough to draw more reliable and comprehensive results, particularly for L2 proficiency. For example, although I could draw statistically significant

estimates for this issue (i.e., weighted and adjusted means for each value of the variable), there were only five post-test ESs (6%) and four follow-up ESs (12%) for high L2 proficiency. Although the intermediate L2 proficiency level I used could, in theory, be divided into low (B1) and high intermediate (B2) levels to match the Common European Framework of Reference for Languages (CEFR) guideline, making subjective judgments would not be methodologically justified because many of the included studies did not provide detailed information from which I could draw an objective distinction. Mainly because of the small sample size for this moderator, I did not have enough statistical power to compute interaction effects between moderator variables accurately. Similarly, in addition to the moderators identified in the included studies, there may be other moderators that could influence L2 vocabulary learning, such as corpus size, purpose of using the corpus (e.g., to answer specific linguistic inquiries or to acquire L2 vocabulary), or contexts (e.g., foreign or second language learning). Unfortunately, the small sample size for Chapter 1 did not allow me to conduct a reliable moderator estimation for these variables. Having more studies on pedagogical use of a corpus would make it possible to perform such an analysis and, thus, possibly suggest further pedagogical implications.

In light of these limitations, which call for the accumulation of more empirical evidence, I recommend that future research in L2 studies include clear descriptions of students' L2 proficiency levels as well as information about their L2 vocabulary teaching contexts, in order to minimize meta-analysts' "informed guesswork" (Boulton & Cobb, 2017, p. 28). In addition, an investigation into interactions between L2 proficiency level and the dimensions of L2 vocabulary knowledge should yield important pedagogical implications for effective corpus use for L2 vocabulary learning. Finally, further

development of computer technology may continue to spawn more effective and efficient means of incorporating corpus use into L2 vocabulary learning, which will also, in turn, merit empirical investigation.

CHAPTER 2: Effects of Concordance-based Electronic Glosses on Second Language Vocabulary Learning[5]

1. Introduction

Unprecedented technological change is transforming classroom environments, often leading students to read electronic texts on computer screens (i.e., digital reading) instead of paper-based textbooks. Digital reading may offer some potential advantages. For example, vocabulary learning through reading could benefit from multimedia environments that provide textual (e.g., synonyms, definitions), audio (e.g., pronunciation, sound effects), or visual supports (e.g., pictures, videos; Anderson-Inman & Horney, 2007; Nation, 2009; Yanguas, 2009). Among these digital scaffolding tools, the focus of this study is on electronic glosses (hereinafter, e-glosses) for textual supports. Given that digital reading environments are more versatile and dynamic than their paper-based counterparts, the potential of e-glosses has been a subject of scholarly interest for L2 vocabulary and reading research (Abraham, 2008; Chun, 2011).

Traditional glosses, which provide supplementary information for vocabulary in reading texts, have been highlighted as an effective tool for vocabulary learning, particularly in learning meanings of unfamiliar words when reading a lexically challenging text (Nation, 2009; Schmitt, 2008). On the other hand, e-glosses may have different *formats* on the computer screen (e.g., AbuSeileek, 2011; Lee & Lee, 2015; Chen & Yen, 2013), or may be filled with different *types* of glossary information (e.g., Lee & Lee, 2015; Poole,

[5] The text of this chapter is a reprint of the material as it appears in Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology, 21*(2), 32–51. https://doi.org/10125/44610. I was the primary investigator and author of this paper, and the co-authors directed and supervised research which forms the basis for the paper.

2012). This means that digital reading can include various types of glossary information for its target vocabulary, regardless of length. In this study, I endeavored to adopt a new type of glossary information: concordance lines.

However, it is surprising that there have been only a limited number of empirical studies on this issue to date (e.g., Lee & Lee, 2015; Poole, 2012), considering the strong theoretical justification for the idea of using this type of information for L2 vocabulary learning, such as input enhancement (i.e., key words are salient in each sentence; Chapelle, 2003), noticing hypothesis (i.e., learners will notice a target word while exposed to its occurrences in multiple contexts; Schmidt, 2001), and involvement load hypothesis (i.e., readers will be involved in meaning inferences; Laufer & Hulstijn, 2001).

The issue I attempt to address by examining the value of concordance lines as effective glossary information for L2 learners' acquisition of word meaning is the gap between the theoretical supports and empirical evidence that should be bridged for finding a more effective pedagogical approach to L2 vocabulary learning. To this end, I tested the effects of two different types of e-glosses, with the first type providing the concordance lines of target vocabulary only, and the second type providing the concordance lines plus the definition of target vocabulary, under a repeated-measure design (i.e., within-subject; see *Experimental Conditions and Design* for details). I also analyzed log data related to the participants' interactions with e-glosses in order to gauge the extent to which they consulted the glossed items and comprehended concordance glossary information. Along with results from the experimental phase, interview data with a subset of the participants and the record of their implementation of e-glosses aided in understanding of the learners' complex interactions with e-glosses.

2. Background

The conceptual foundation of Chapter 2 flows from three interrelated topics: (1) the role of electronic glosses in digital reading environments; (2) the benefits and limitations of using concordance lines for vocabulary learning; and (3) prior research on learners' interactions with e-glosses, and its implication for using concordance lines as a type of glossary information in e-glosses.

2.1 Digital reading and e-glosses

Anderson-Inman and Horney (2007) suggest that digital reading would offer promising opportunities for readers in terms of accessibility and supportiveness by providing various types of digital scaffolding tools alongside the text. In their view, e-glosses can serve as effective supports for transforming a plain electronic text into a "supported eText" (p. 153). In a similar sense, e-glosses have been supported within several theoretical and pedagogical backgrounds (Lee & Lee, 2015; Chun, 2011). Above all, digital reading increases the likelihood of target vocabulary being noticed by readers because these items, which are hyperlinked to e-glosses, can be made visually salient in a variety of styles (Chapelle, 2003; Chun, 2001). Therefore, e-glosses have the potential to contribute to a reader's learning of unfamiliar vocabulary when reading electronic texts (i.e., noticing hypothesis, Schmidt, 2001). Furthermore, by giving readers more control over their reading processes (Leu, Kinzer, Coiro, Castek, & Henry, 2013), digital reading constructs an "interaction" among readers, texts, and scaffolding materials (i.e., e-glosses; Chapelle, 2003, p. 25). Lastly, unlike the print form, digital reading is not limited by spatial restrictions (Lee & Lee, 2015); thus, digital platforms may have e-glosses filled with an

abundance of lexical information such as multiple concordance lines for target vocabulary (Nation, 2009).

In light of these virtues of digital environments, a number of empirical studies have demonstrated the positive effect of e-glosses on L2 learners' vocabulary learning (see Abraham, 2008 for a meta-review). Furthermore, the interest of the research community has recently shifted toward the *format* of glossing (e.g., AbuSeileek, 2011; Chen & Yen, 2013) as well as the *type* of glossary information (e.g., Lee & Lee, 2015; Poole, 2012). Regarding the former, a small number of empirical studies (e.g., AbuSeileek, 2011; Chen & Yen, 2013) have been conducted on comparing the effects of different formats of e-glosses (e.g., pop-up type, marginal type), but the findings of these studies have not been consistent. Research on different types of glossary information for e-glosses is even scarcer, in particular, regarding the use of concordance lines as glossary information. In the following section, I will first review the literature dedicated to the use of the corpus for vocabulary learning, and then introduce two studies that have used concordance data in e-gloss format.

2.2 Concordance lines as vocabulary learning resources

The use of corpora in L2 vocabulary learning has attracted the interest of the research community for the following reasons. First, inferring the meaning of an unfamiliar word is considered an effective strategy for learning vocabulary (e.g., Fraser, 1999; Schmitt, 1997); and learners, in theory, are supposed to make a more informed and accurate guess of the meaning of an unfamiliar word when exposed to multiple contextual instances surrounding a target word (Johns, 1986). Second, allowing learners to infer meaning from examples is thought to generate a high level of learner involvement, which may lead to

greater retention (Laufer & Hulstijn, 2001). Third, providing multiple instances of target vocabulary in a wide range of sentential contexts is believed to enhance learners' awareness of target vocabulary, thus accelerating their vocabulary acquisition (Chapelle, 2003; Schmidt, 2001).

Although they did not utilize a corpus analysis, early empirical studies attempted to confirm the effect of using example sentences excerpted from corpus-based dictionaries on L2 vocabulary learning (e.g., Laufer, 1993; Summers, 1988). For example, Summers (1988) examined the effects of example sentences for the participants' vocabulary comprehension and production. She designed three different conditions, with types of information selected from dictionaries: definitions, example sentences, and both definitions and example sentences. Although the experimental conditions led participants to have better results than the control condition, there was no significant difference across different conditions with respect to the participants' vocabulary comprehension and production. With a similar research objective, Laufer (1993) tested the use and comprehension of 18 target vocabulary with 43 EFL undergraduate students, providing four different conditions (i.e., definitions, examples, definitions followed by examples, and examples followed by definitions). The results indicated that the combinations of definitions and examples were more effective than definitions only or examples only for the participants' vocabulary use and comprehension. Moreover, Laufer found that definitions might contribute more to improving comprehension than examples, whereas the contributions of these two components were similar for the production counterpart.

Recent studies showed more positive results, probably because of more diverse experimental conditions thanks to the state-of-the-art corpus technology (e.g., Chan & Liou,

2005; Cobb, 1999; Frankenberg-Garcia, 2012, 2014). For example, Cobb (1999) conducted an empirical study with two different vocabulary learning conditions: concordance-based vocabulary learning (e.g., the use of a concordance program) and traditional vocabulary learning (e.g., the use of dictionaries and word lists). The results of Cobb's study with 20 adult Chinese EFL students showed that the former treatment yielded more gains in terms of the learners' knowledge of vocabulary. In Chan and Liou (2005), 32 Taiwan college students completed web-based practice units incorporated with a bilingual concordancer, and the results showed a significant collocation improvement with the use of a corpus example, as well as an on-line concordance program during their vocabulary practice. Recently, Frankenberg-Garcia's studies (2012, 2014) confirmed the positive effects of concordance lines on L2 vocabulary learning. Taking into consideration that examples should include enough contextual clues for comprehension, she carefully selected concordance lines from multiple corpora. The results indicated that these examples were effective for EFL students in encoding (e.g., correcting typical L2 mistakes, 2012; writing sentences using target vocabulary, 2014) and decoding (e.g., understanding target vocabulary, 2012; 2014).

However, when it comes to L2 vocabulary learning in the e-gloss format, empirical findings have been inconclusive (e.g., Lee & Lee, 2015; Poole, 2012). For example, Poole (2012) compared the effects of syntactically modified concordance lines and dictionary definitions of glossed words as two different types of glossary information, and could not find any statistical difference between these two types in improving the participants' vocabulary acquisition. Similarly, Lee and Lee's (2015) study re-examined the effects of these two different types of glossary information. Unlike Poole's (2012) study, Lee and Lee

did not modify concordance lines to the level of the participants. The results showed that participants who received dictionary definitions made higher vocabulary gains than those who had concordance lines as their glossary information.

These inconclusive results concerning the value of concordance lines may be explained by previous suggestions that learners inferencing meaning from context may do so ineffectively (e.g., Schmitt, 2008), and thus retain wrongly inferred meanings in their lexicons (Mondria, 2003). Similarly, it should also be noted that the learners might not be able to understand all the given concordance lines to successfully elicit the meaning of target vocabulary item. In short, the use of concordance lines, which inevitably involves the inference of meaning, has not received the attention it deserves.

One way of overcoming the lack of evidence is to enable learners to confirm the meaning they inferred from contexts (e.g., Cobb, Greaves, & Horst, 2001; Fraser, 1999). As Godwin-Jones (2001b) suggested, if the learners' meaning inferences can later be confirmed, their inaccurate inferences will be minimized. However, to date, there has been no empirical study to support this suggestion, in particular, none in digital reading environments.

2.3 Learners' interactions with e-Glosses

Along with the learners' inaccurate meaning inferences, another practical issue to consider in using concordance lines as glossary information is the learners' interactions with e-glosses, including implementation rate of clicking e-glosses. By tracking user behavior, previous studies have focused on conditions that made learners click e-glosses (e.g., Chun, 2001; Laufer, 2000). One of the major findings from these studies is that learners largely prefer a type of lexical information that requires a relatively low level of

cognitive load. For example, Laufer (2000) found that participants did not make use of example sentences of a target word as a type of glossary information in digital reading environments; rather they opted for word definition. It is likely that concordance lines as glossary information require a relatively high level of cognitive load for learners to process; it can be expected that this type of glossary information may not be overwhelmingly favored by learners. Hence, understanding the learners' interactions with e-glosses seems to be an important issue in examining the effects of e-glosses for vocabulary learning.

However, there have been few empirical efforts to assess the implementation rates and observe the specific behaviors of learners' consultations of concordance lines in e-glosses. For example, Poole (2012) did not include any research methods to figure out how and to what extent they interacted with glossary information for their understanding and learning. This is a problem in other studies on e-glosses, which are limited to focusing only on the results of vocabulary tests, based on the assumption that treatments have been ideally employed without properly understanding learners' behaviors. Along the same line, it is noteworthy that the aforementioned theoretical supports for using concordance lines as glossary information are based on the premise that learners would be likely to devote their full attention to that kind of lexical information.

In light of this gap in the literature, I will not only examine the effects of providing a confirmation process along with concordance lines for the meaning inferences, but also observe the learners' implementation rate of consulting glossary information as well as their clicking behaviors with e-glosses. I hypothesize that this will be a significant step toward overcoming the limitations of concordance lines as glossary information, and propose future pedagogical directions for L2 vocabulary learning.

2.4 Research questions

1. What effects do the two different vocabulary learning conditions in digital reading environments have on the meaning-recall of target vocabulary for English as a Foreign Language (EFL) adult learners?
    A. How significant are the effects of receiving different treatments on the meaning-recall of target vocabulary?
    B. How significant are the effects of providing the treatments in different orders?
2. How do the participants interact with e-glosses when reading, and what are the pedagogical implications of these interactions?
    A. What are the average implementation rates of and amount of time spent on consulting e-glosses during reading?
    B. What are their relations to the results of a meaning-recall post-test?
    C. How are the participants' clicking behaviors different across target vocabulary?

3. Methods

To answer the research questions, I carried out an experiment that was based on a repeated measures design. This enabled me to deliver different reading conditions and measure their effects on the participants' meaning-recall of target vocabulary in a controlled way with a reliable level of confidence. This section discusses the different aspects of the research method of Chapter 2: the description of the participants, experimental design, target reading materials, and an outline of the procedure and data analysis.

3.1 Participants

A total of 138 undergraduate South Korean EFL learners participated in the study. All of them were 21 years old and had ten years of English learning experience in formal school contexts. The average score of the participants on the Test of English for International Communication (TOEIC) was 732, thereby indicating that they were independent English users (B1–2) based on the Common European Framework of Reference for Languages (CEFR), according to the testing publishers (Educational Testing Service, 2016). The data collection was conducted during their enrollment in a mandatory English course. The participants were from nine intact classes, taught by three different instructors with the same textbook and curriculum at the time of the study.

3.2 Experimental conditions and design

I first developed three different versions of the digital reading materials, one with e-glosses for concordance lines (hereinafter, CONC) and another with e-glosses for concordance data supplemented with dictionary definitions (hereinafter, CODI), and the other without any e-glosses, which served as a control (hereinafter, CTRL). A pilot study was conducted to determine the appropriate number of concordance lines for glossary information; in general, students in the study (n = 45) pointed to three examples as the most manageable number for inferring the meaning of target items, without being distracted (Mean = 2.91, SD = .73). In this way, the aforementioned three instructors and myself carefully selected three concordance lines for each target vocabulary item from multiple reference corpora (e.g., Open American National Corpus, British National Corpus, Brown Corpus) for the participants' effective meaning inferences in light of their proficiency levels (see Frankenberg-Garcia, 2012, 2014 for similar efforts). Appendix 2.1

further explicates how I chose the most appropriate three examples for the word "inflection" in order to highlight its specific meaning used within its context (i.e., a change in the pitch or tone of a person's voice).

For the dictionary definitions of target vocabulary, this study opted to use the Merriam-Webster on-line dictionary, which has been widely used in the participants' college. The rationale behind choosing L2 dictionary definitions for glossary information instead of first language (L1) was twofold: the participants were intermediate-level learners with a vocabulary of more than 2,000 words (i.e., learners with a vocabulary of less than 2,000 often have comprehension problems with L2 glosses; Nation, 2009); and L2 dictionary definitions were considered to be more appropriate language input in this experimental condition, where L2 concordance lines were provided as glossary information.

To further obtain data on the participants' behaviors with e-glosses and glossary information, this study used a free on-line survey tool (i.e., SurveyMonkey) to provide glossary information in a popup window format. Specifically, every glossed lexical item was hyperlinked to a pop-up style window, which presented concordance lines of the item upon the participants' activation. For the CODI, an additional window was designed to provide the definition of a target item, which was activated when a user clicked the "next" button after reading concordance lines in the previous window (see Figure 2.1). In this manner, data regarding the participants' clicks on glossed items were collected to analyze the implementation rate of and length of time they spent with each e-gloss until closing the window. Furthermore, a checkbox was included in front of each concordance sentence so that the participants could report which of the concordance lines they understood. This e-

gloss format was designed in such a way that this study could measure how and to what extent the participants interacted with the glossed words, and whether they understood or consulted glossary information during their reading.

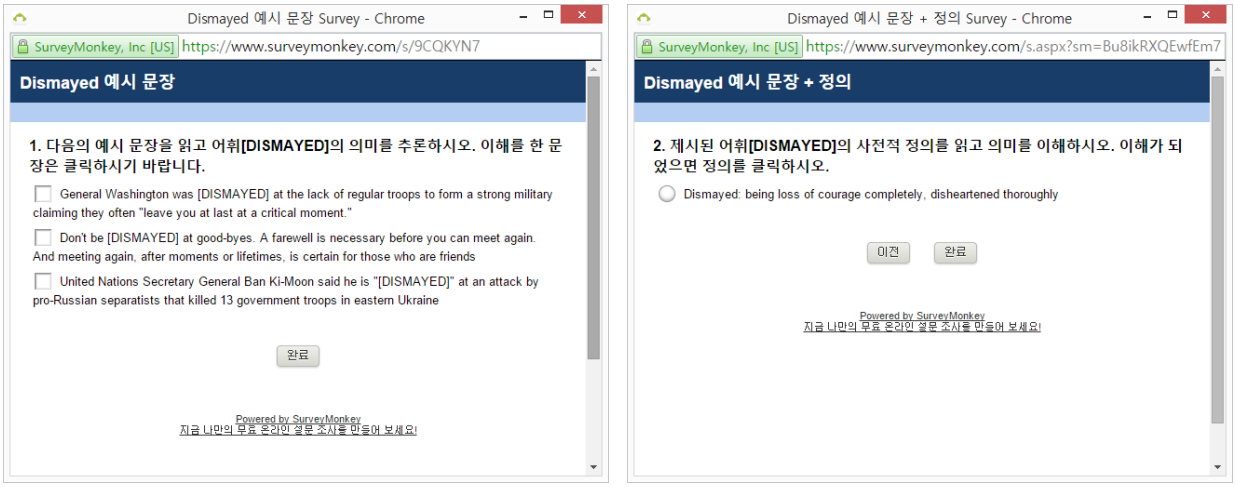

Figure 2.1. Glossary Information.
Note. Concordance lines (left) and dictionary definitions (right).

In order to have the participants exposed to all the conditions, a repeated measures design was adopted. This design allows each participant to experience all the conditions, including the control condition. In repeated measures experiments, it is important to confirm that any order effects (i.e., the effects that the order of presenting three different conditions might have on the results) do not exist. As part of this effort, counter-balancing was taken into consideration when designing the group formation, as shown in Table 2.3. For example, there were six possible orders to consider all the possible combinations of the three conditions: (1) CTRL-CODI-CONC, (2) CODI-CONC-CTRL, (3) CONC-CTRL-CODI, (4) CTRL-CONC-CODI, (5) CONC-CODI-CTRL, and (6) CODI-CTRL-CONC. Since there were nine intact classrooms in total, there was a random assignment of six different orders to the six classrooms, and the remaining three classrooms were randomly assigned to one of those six orders. The results of the data analyses confirmed that there were no significant order

effects from the experimental design (see non-significant coefficients of the order effect variable in Table 2.3).

Table 2.1
Study Design

Order	Classroom	Trial 1	Trial 2	Trial 3
Order 1 (*n* = 29)	1, 2	CTRL	CODI	CONC
Order 2 (*n* = 29)	3, 4	CODI	CONC	CTRL
Order 3 (*n* = 28)	5, 6	CONC	CTRL	CODI
Order 4 (*n* = 15)	7	CTRL	CONC	CODI
Order 5 (*n* = 15)	8	CONC	CODI	CTRL
Order 6 (*n* = 16)	9	CODI	CTRL	CONC

3.3 Material

Three reading texts were extracted from *Cutting Edge Advanced* (Cunningham, Moor, & Carr, 2003). The length of each reading was 459, 479, and 519 words, respectively. I chose 30 potential target vocabulary from these texts and selected ten target vocabulary per text based on the results of a pilot test with 45 students similar in profile to the participants in the study (see Appendix 2.2 for a list of target vocabulary and Appendix 2.3 for the texts and their hyperlinks).

3.4 Testing and scoring

A total of four meaning-recall tests of vocabulary were conducted, at the beginning of the study as well as after each reading activity. In these tests, participants were asked to write down the meaning of a target word either in English or their L1 (i.e., Korean). When scoring, a total of two points were allotted for each item. One point was given when students gave a partially correct meaning of each target vocabulary item, while two points were given for a completely correct meaning. One of the three instructors and myself graded the vocabulary test. Both raters scored fifteen percent of the testing sheets for the

purpose of checking inter-rater reliability; and the reliability was found to be .93 (Cohen's Kappa, p = .01). Any discrepancies were resolved through discussion.

3.5 Procedure

At the beginning of the study, the participants were given a pre-test with all 30 target vocabulary items from the three target texts; as previously mentioned, the target vocabulary were selected from a pilot study. Then, a computer workshop was given after the pre-test to give the participants some basic knowledge concerning the definition of concordance lines and how they could infer the meaning of a glossed word by consulting the given lines.

The main reading tasks were conducted two weeks after the pre-test. As part of the effort to minimize the potential impact of instructors, all the reading materials, including testing ones, were designed in a way in which each individual could complete all the activities without any further guidance or instruction. Each task was performed weekly to prevent any possible carry-on effects (i.e., effects that carry over from one condition to another). In each reading session, the participants were asked to read the text with their own laptops for 15 minutes, and this reading was followed by an immediate post-test for 5 minutes on a different web page.

3.6 Interview

After the experiment, in order to understand how the participants interact with e-glosses across the target vocabulary, interviews were carried out with a purposely stratified sample of three participants: one student of advanced proficiency (C1), one at the upper intermediate level (B2), and one intermediate user (B1). In the interview, these participants were presented with the three different texts they read, and were asked to

give their opinions about each target vocabulary item and how much additional glossary information was needed in comprehending its meaning. The post-interviews were conducted in their first language, audio-recorded, and partially transcribed and translated.

3.7 Data Analysis

Statistical analyses were performed using STATA (Version 14.0). Prior to the regression analyses, correlations between predictor variables (i.e., independent variables in the regression equations) were examined in order to control for multicollinearity. The results showed that predictor variables were not strongly related ($r < .8$). Then, residualized change regression analyses with the Huber-White standard errors (i.e., controlling for heteroskedasticity) including the cluster adjustment (i.e., ensembling multiple test results at the student level) were conducted for the first research question, followed by additional analyses for the robustness checks (i.e., fixed-effect adjustments, simple change regression analyses; see Appendix 2.4 for the variables and equations for these regression models and details).

For the second research question, the number of clicks of all the glossed words and the length of time spent consulting glossary information were analyzed, along with the participants' reports on the number of concordance lines they had comprehended. In particular, the amount of time the participants spent on consulting glossary information was analyzed by excluding potential outliers (e.g., those who did not spend enough time on making meaning inferences or those who spent too much time on each target vocabulary item, for example, if they left the pop-up window open). The interquartile range (IQR) rule was applied in this case. For example, the first quartile (Q1) and the third quartile (Q3) were calculated, based on the time the participants spent on each of the target vocabulary.

Then the IQR was calculated (IQR = Q3 – Q1), and the lower boundary (Q1 – 1.5 × IQR) and the upper boundary (Q3 + 1.5 × IQR) were computed. If the time one spent on consulting glossary information was outside this range, then this click was considered an outlier. Combining all of the above, the implementation rates of the participants' clicking the target vocabulary and consulting glossary information were analyzed. A paired *t*-test was further conducted to compare the mean difference between implementation rates for the two experimental conditions (i.e., CODI and CONC). Moreover, a multiple regression analysis with Huber-White standard errors was conducted in order to confirm possible associations between the participants' clicking behaviors and the meaning-recall rate of the target vocabulary.

3.8 Limitations

There are two limitations of Chapter 2. First, delayed tests were not conducted because of the participants' limited availability. Within a brief time period, I decided to provide them with all the conditions without employing delayed tests, rather than to randomly assign them into one of the three conditions (i.e., CTRL, CONC, and CODI) with delayed tests. So I was not able to assess retention of vocabulary. Second, while the inclusion of an experimental condition with definitions alone would have allowed me to measure the effects of dictionary definitions in CODI more accurately, scheduling considerations (i.e., participant availability) made the use of a control group more feasible than a definition-only group. This allowed me to examine the effects of concordance lines as glossary information, as well as the effects of the confirmation of meaning inferences through dictionary definitions. These effects have been examined only to a limited extent in the previous literature (unlike definition-only, which has received considerable attention).

While the experimental design of Chapter 2 was suitable for my goals, future research with a definition-only condition will be valuable for illuminating the pedagogical implications of exposure to different types of glossary information.

4. Results

This section presents the findings in two parts, with the first part reporting the results of vocabulary tests based on a set of multiple regression models, and the second part presenting the results concerning the participants' clicking behaviors in different experimental conditions along with the interactions of (1) individual lexical items, (2) their recall rates, and (3) the participants' clicking behaviors.

4.1 Results of vocabulary recall tests

Table 2.2 provides descriptive statistics for the scores of the three conditions on the vocabulary recall tests. Out of 138 participants, a total of six could not complete all three reading tasks, and thus were excluded in the analyses. Overall, the participants demonstrated significant gains in learning vocabulary for all the conditions, according to the results of paired *t*-tests ($p < .001$).

Table 2.2
Descriptive Statistics for the Vocabulary Tests

Conditions	Pre-test Mean (*SD*)	Post-test Mean (*SD*)	*t*-test	*t* value
CTRL ($n = 132$)	.27 (.64)	3.97 (3.17)	Pre < Post	– 13.45***
CONC ($n = 132$)	.20 (.44)	6.24 (4.17)	Pre < Post	– 17.00***
CODI ($n = 132$)	.14 (.43)	8.89 (4.60)	Pre < Post	– 22.19***

*** $p < .001$

Regarding the first research question, the results from the residualized change model, as shown in Model 1 in Table 2.3, revealed that there was a significant treatment

effect depending on the different conditions (b = 2.51, p < .001) when controlling for three learner variables (i.e., pre-test scores, English proficiency, and gender).

For the next step, dummy variables for the three different conditions were plugged into the regression model to compare the participants' post-test scores under these conditions (see Model 2 in Table 2.3). The estimated coefficients for the dummy variables implied that CTRL would, on average, lead a participant to get a 2.32 lower vocabulary score than CONC (b = – 2.32, p < .001), and that one would, on average, get a 2.69 higher vocabulary score if the participant were given CODI rather than CONC (b = 2.69, p < .001). As a result, it can be interpreted that one would, on average, learn about one more target vocabulary items out of 10 total, or partially learn about two more target vocabulary if the participants were given CODI rather than CONC.

In addition, when the order effect product term (i.e., the interaction effect of providing different experimental conditions in different orders; conditions × trial) was added to the model, no significant effect was found (b = .25, p > .05), with the treatment effect depending on the different conditions remaining statistically significant, as shown in Model 3 in Table 2.3.

The additional analyses for the robustness checks also confirmed the aforementioned findings. The full results regarding these analyses, including fixed-effect adjustments (i.e., Model 4 in Table 2.3) and simple change models (i.e., Models 1, 2, and 3 in Table 2.4), are described in Appendix 2.5.

Table 2.3

Regression Models of the Vocabulary Tests (Residualized Models)

	Dependent variable: Vocabulary recall post-test (*n* = 132 participants × 3 conditions = 396 observations)			
Independent variables	Model 1 (Residualized model)	Model 2 (Dummy variables)	Model 3 (Order effect added)	Model 4 (Classroom fixed-effects)
Conditions	2.51***			
	(.21)			
CTRL		-2.32***	-1.84**	-1.97*
		(.41)	(.66)	(.88)
CODI		2.69***	2.21***	2.34**
		(.44)	(.62)	(.88)
Trial			.86	1.00
			(.50)	(.79)
Order effect			.25	.18
Trial × Condition			(.28)	(.38)
Pre-test	.40*	.40*	.41*	.41*
	(.20)	(.20)	(.18)	(.19)
English proficiency	.01***	.01***	.01***	.01***
	(.00)	(.00)	(.00)	(.00)
Gender	-.32	-.32	-.39	-.70
	(1.54)	(1.54)	(1.55)	(.96)
Constant	-9.52***	-4.63*	-7.19***	-6.46***
	(1.86)	(1.81)	(1.86)	(1.81)
R^2	.307	.308	.370	.352

Note. Standard errors are in parentheses. CONC is the omitted condition category in Models 2, 3, and 4.

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

Table 2.4

Regression Models of the Vocabulary Tests (Simple Change Models)

	Dependent variable: Vocabulary gains (Post – Pre; *n* = 396)		
Independent variables	Model 1 (Simple change model)	Model 2 (Dummy variables)	Model 3 (Classroom fixed-effects)
Conditions	2.02***		
	(.54)		
CTRL		-1.83**	-1.97*
		(.66)	(.88)
CODI		2.20***	2.34**
		(.62)	(.87)
Trial	.82	.82	.96
	(.50)	(.50)	(.79)
Order effect	.25	.25	.18
Trial × Condition	(.28)	(.28)	(.38)
English proficiency	.01***	.01***	.01***
	(.00)	(.00)	(.00)
Gender	-.39	-.39	-.71
	(1.53)	(1.53)	(.95)
Constant	-11.04***	-7.13***	-6.45***
	(2.15)	(1.81)	(1.79)
R^2	.360	.360	.348

Note. Standard errors are in parentheses. CONC is the omitted condition category in Models 2 and 3.

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

4.2 Results concerning the participants' clicking behavior

As for the second research question, I first examined whether CODI and CONC resulted in different implementation rates. As shown in Table 2.5, the participants, on average, showed an implementation rate of about 83%. Despite the different amounts of vocabulary gains between CODI and CONC ($t = -3.41$, $p < .001$), the implementation rates were not significantly different between the two conditions ($t = -.92$, $p > .05$). Moreover, the participants spent similar amounts of time looking up glossary information in CODI and CONC ($t = -.66$, $p > .05$). Lastly, the number of concordance lines of which the participants reported comprehension was not substantially different between the conditions ($t = -.52$, $p > .05$).

Further regression analysis was performed to explore a relationship between the participants' clicking behaviors and their recall rates of target vocabulary. The dependent variable was the meaning-recall scores of the target vocabulary, whereas the independent variables included (1) the average amount of time (in seconds) spent on each piece of glossary information, (2) the average number of concordance lines each participant reported to have comprehended, (3) the rates of clicking for each e-gloss, and (4) the condition variable (dummies for CODI and CONC; see Table 2.6). The results indicated that the condition variable had a significant effect on predicting the recall score (b = 11.70, p < .001), whereas the number of concordance lines each participant clicked did not (b = – 15.86, p > .05). On the other hand, the amount of time spent on glossaries had a negative effect on predicting the dependent variable, albeit a very weak one (b = – .28, p < .05).

Table 2.5
Recall Scores and Clicking Behaviors between CONC and CODI

Conditions	CONC (n = 30)	CODI (n = 30)	t-test	t value
Vocabulary recall test score	27.57 (12.33)	39.23 (14.10)	CONC < CODI	– 3.41***
Rates of clicking each e-gloss	.82 (.14)	.85 (.11)	CONC ≒ CODI	– .92^{ns}
Average time spent	37.22 (21.82)	4.88 (2.99)	CONC ≒ CODI	– .66^{ns}
Number of concordance lines they comprehended	1.52 (.37)	1.57 (.32)	CONC ≒ CODI	– .52^{ns}

Note. Standard deviations are in parentheses.
^{ns} p > .05, *** p < .001

Table 2.6
Influences of Participants' Clicking Behaviors on Vocabulary Tests

Independent variables	Coefficients (Standard errors)
Average time spent	– .28* (.11)
Rates of clicking each e-gloss	6.69 (34.74)
Number of concordance lines they comprehended	– 15.86 (8.73)
Condition (*CODI*)	11.70** (3.43)
Constant	3.85 (15.19)
R^2	.26

Note. CONC is the omitted condition category, so the condition coefficient is relative to this group.
* $p < .05$, ** $p < .01$

The results of the post-interview revealed a complex picture of the participants' vocabulary learning, as described in Figure 2.2. In other words, the close analyses of each target vocabulary item, along with the participants' clicking behaviors and recall scores, pointed to complex interactions of (1) participants' clicking behaviors, (2) the nature of selected concordance lines, (3) the surrounding context of target vocabulary item, and (4) the participants' prior knowledge of target vocabulary.

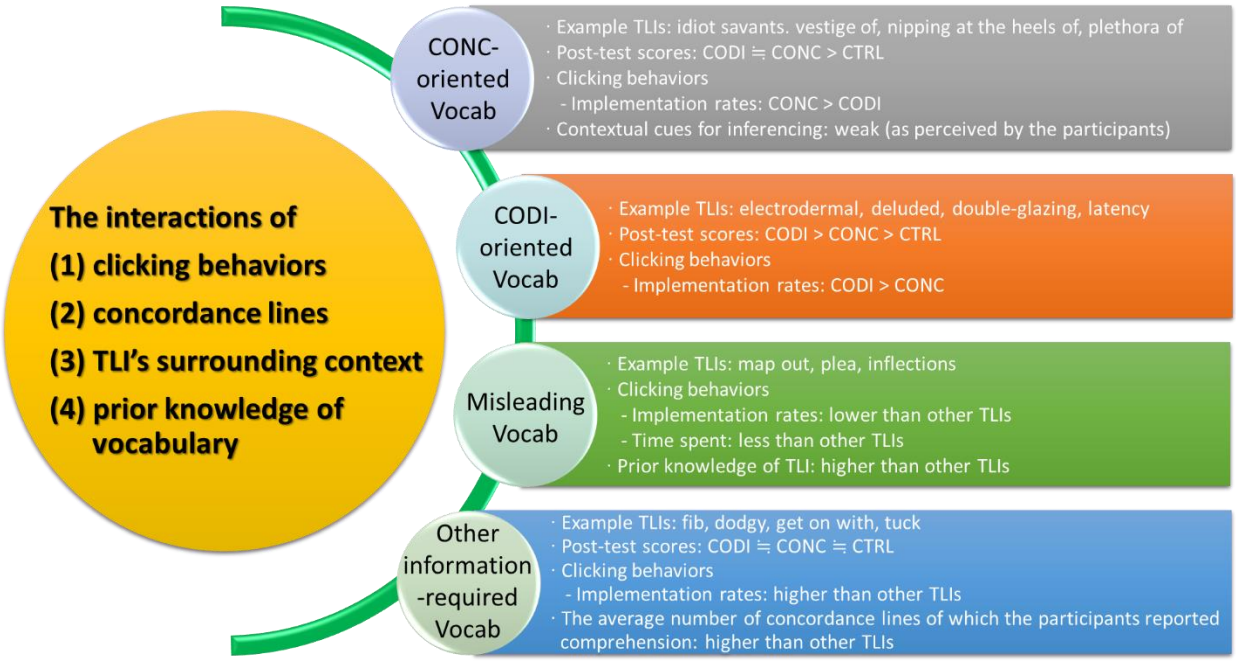

Figure 2.2. Four Emerging Patterns of Target Vocabulary from the Participants' Interactions with E-glosses.

The first group of the target vocabulary with similar patterns included "idiot savants," "vestige of," "nipping at the heels of," and "plethora of." These patterns were as follows: (1) the participants' meaning-recall scores in CTRL were on average close to zero (indicating that the participants in this condition failed to infer the meaning of these target vocabulary), and there were only small differences in the meaning-recall score between CONC and CODI; (2) the participants' implementation rates were higher on average in CONC than CODI; and (3) the participants lacked contextual cues in inferring the meaning of target vocabulary, as can be seen from the response (below) to the target vocabulary item "plethora of."

> **C1** (the interviewees' names are replaced by their proficiency levels: C1, B2, and B1): To be honest, I don't know about this item … I think I can guess its meaning from the previous paragraph … but not from sentences surrounding this item.

B2: I think I can guess what it means … but I am highly uncertain.

With these target vocabulary items, the participants were thus not able to infer the meanings properly, but concordance lines significantly contributed to their meaning inferences, whereas the confirmation of their meanings through dictionary information was not obligatory for most of the participants. Having seen the patterns of the first group, the words that fall into this category may be called "CONC-oriented vocabulary," meaning that concordance lines as glossary information are not only beneficial, but also sufficient for accurate meaning inferences.

The second set of target vocabulary that showed consistent patterns included "electrodermal," "under duress," "in the vicinity of," "deluded," "double-glazing," and "latency." For these target vocabulary, three observations were made. First, the participants' average meaning-recall score was highest for CODI, lower for CONC, and lowest for CTRL. Second, the participants' implementation rates were higher in CODI than in CONC. Third, the participants were able to make some inferences about the meaning of target vocabulary item based on the surrounding context and the part of the target vocabulary item (i.e., morpheme), as can be seen from the interviewees' comments on "electrodermal."

C1: When I see this word, "electrodermal" … it reminds me of the word "electronic." Considering previous words, such as "blood pressure" and "breathing rate" … this word could be related to the physical signs of the human body.

B2: I see this word consists of "electro" and "dermal" … and I know both of these words. After reading the previous and next sentences … I was able to figure out that this word may indicate a sort of electronic sign from human skin.

B1: I think this word is highly related to the term electronic. I don't see much of a problem for guessing the meaning of this word.

The interviewees also made a similar response to "latency," indicating that they were able to make inferences based on the surrounding context. However, the comparison of the total meaning-recall scores between CODI (total score = 36) and CONC (total score = 24) suggests that the confirmation of the meaning of this word enhanced the participants' comprehension. Thus, these words may be called "CODI-oriented vocabulary," thereby indicating that meaning inferences followed by the confirmation of correct meaning inferences would result in the most positive learning outcome for these target vocabulary items.

The third group of the target vocabulary, which included "map out," "plea," and "inflections," had the following patterns: (1) the participants' implementation rates were 10% lower for these items when compared with those rates of all other target vocabulary (which was about 80%); (2) the participants spent much less time on reading glossary information of these target vocabulary (on average, 23 seconds, compared with the average of 45 seconds for the other set of the target vocabulary); and (3) the participants were fairly confident in their inferred meanings of these target vocabulary, as can be seen from their responses to the target item, "map out," as below.

C1: I already knew this expression … "to map" is to draw a map … to me, there is no need to have extra help for this easy phrase.

B2: I do not see any necessity for accessing additional information for "map out."

B1: I am not sure about this phrase … but it seems straightforward ... I think … it is to draw something. I don't think I need more information for this word.

It was found that the interviewees did not attempt to find contextual meanings of the expression "map out," which refers to "to plan" in the target context. It appeared that the three interviewees knew about the word "map," but did not go further to explore the meaning of "map out." The results of the post-test also support the comments of the interviewees. In particular, the majority of the participants' wrong answers were related to "a map" or to "drawing a map" (47 out of 97 wrong responses). In light of the rather low implementation rate for these target vocabulary, and the interviewees' confidence and misjudgment, I call those kind of terms, "misleading vocabulary." These items require particular attention from instructors, who will need to make sure their learners would not make wrong meaning inferences.

The final set of the target vocabulary items included "fib," "dodgy," "get on with," and "tuck," and had the following pattern: (1) the participants' average implementation rate was high (approximately 88%); (2) the average number of concordance lines for which each participant reported to have comprehended their meanings (M = 1.92) was higher than that for the rest of the target vocabulary items combined (M = 1.72); and (3) CONC and CODI did not result in higher recall scores than CTRL. In other words, these target vocabulary items were highly consulted, and their concordance lines were

comprehensible to the participants. However, CONC and CODI were not necessarily more beneficial to the participants' meaning-recall than CTRL. So, it seems that these target vocabulary items require other glossary information that was not provided in this study (e.g., L1 equivalents) for higher meaning-recall rates. Based on this insight, these words are called "other information-required vocabulary."

5. Discussion

The first research question investigated whether two different vocabulary learning conditions (i.e., CONC and CODI) would make any differences in undergraduate EFL learners' meaning-recall knowledge of target vocabulary. With regard to this research aim, the results showed that the participants fared better in CONC than CTRL. The finding here supports several theoretical hypotheses that would confirm the use of concordance lines for vocabulary learning, such as the noticing hypothesis (Schmidt, 2001) and the involvement load hypothesis (Laufer & Hulstijn, 2001). In light of previous concerns about using concordance lines as glossary information (e.g., Cobb et al., 2001; Fraser, 1999; Godwin-Jones, 2001b), I cautiously suggest that a few steps undertaken in Chapter 2 may account for the aforementioned positive results. That is, through a carefully planned pilot study, I examined the most appropriate number of concordance lines for their meaning inferences (i.e., three), and had multiple discussions with the instructors of the target classes in selecting example sentences from concordance lines, which were deemed fine-tuned to the participants' level of English proficiency (see Frankenberg-Garcia, 2012, 2014 for similar efforts).

Moreover, CODI was more beneficial to the participants' meaning-recall than CONC, thus supporting prior findings that the additional confirmation of an inferred meaning

supports students' making more accurate meaning inferences from concordance lines (e.g., Cobb et al., 2001; Fraser, 1999). It also accords with Laufer's (1993) study, in which the combination of definition and example sentences resulted in the highest comprehension gains.

The results related to the second research question showed that a holistic account of the participants' meaning-recall is complex, after a close analysis of the interactions concerning (1) the participants' clicking behaviors, (2) the difficulty of selected concordance lines, (3) the surrounding context around target vocabulary, and (4) the participants' prior knowledge of target vocabulary. In particular, I have shown that the participants interacted rather differently with each set of the target vocabulary. That is, a majority of the target vocabulary (e.g., "electrodermal," "latency") were best recalled in CODI, in accordance with the results of the first research question. The superiority of CODI over the other two conditions discussed above may be attributable to the fact that most of the target vocabulary fall into this group. On the other hand, some target vocabulary (e.g., "idiot savants," "vestige of") were recalled fairly well even without the aforementioned confirmation process. These items were concordance-oriented; if concordance lines were judiciously selected for them, then their recall could be guaranteed. Another group of the target vocabulary (e.g., "map out," "inflections") misled learners into thinking that their meanings were easy to infer or were already known to them. In such a case, learners may easily make a wrong inference. These are the lexical items that should be dealt with very carefully by an instructor, as a wrongly inferred meaning could be retained in the learners' vocabulary system (Mondria, 2003). Finally, there were a small number of words (e.g., "fib," "dodgy") that were not recalled well even with concordance lines and dictionary

definitions. It can be assumed that these words may be better retained by learners if other lexical information is provided.

The aforementioned categorization of the target vocabulary may not be equally applicable in other contexts. It is highly likely that learners from different pedagogical contexts, even with the same level of English proficiency as those in Chapter 2, may interact differently with the aforementioned target vocabulary items. My intention was to raise researchers' awareness of the possibility of the dynamic interactions of (1) the learners' prior knowledge of target vocabulary, (2) the comprehensibility of glossary information, and (3) their actual utilization of such glossary information. As an example, the participants' implementation rates in this study ranged from about 50% to 100%, depending on target vocabulary. Through the interviews, it was found that some participants may opt not to use the given glossary information by mistakenly thinking that they already know the meaning of the target vocabulary items. On the other hand, the implementation rate and the participants' self-reported understanding level of concordance lines did not always correlate with the recall rates, thereby implying that L2 vocabulary learning may be subject to the aforementioned dynamic interactions. I believe that the innovation of Chapter 2 lies in demonstrating that some glossary information, such as concordance lines, may involve more unexpected interactions with L2 learners when compared with traditional dictionary information.

6. Conclusion

Chapter 2 investigated the effects of, and clicking behaviors related to, two different vocabulary learning conditions in digital reading environments, with one providing concordance lines only and the other providing concordance lines along with definitions as

glossary information. Based on the findings, I conclude that providing concordance lines along with the subsequent confirmation of the inferred meanings is more effective than providing concordance lines only, which in turn results in better meaning-recall than no glossary information. Furthermore, I have shown that a particular lexical item may need different treatments for it to be recalled most efficiently and effectively through the close analyses of the interactions of (1) the participants' clicking behaviors, (2) the difficulty of selected concordance lines, (3) the surrounding context around target vocabulary, and (4) the participants' prior knowledge of target vocabulary. While the findings should not be interpreted as leading to a prescriptive method for teaching these target vocabulary items, they nevertheless can provide important guidelines for future L2 vocabulary research and teaching. One promising direction for future research would be to compare the effects of CODI with an experimental condition with dictionary definitions alone on the meaning acquisition of L2 vocabulary with delayed tests, in particular on different sets of lexical items, which would compensate for the primary limitation of this study. Future research may benefit from the use of vocabulary measures other than meaning-focused tests, which may further reveal the effectiveness of concordance lines as glossary information for improvements in more productive aspects of lexical competence.

CHAPTER 3: Unearthing Hidden Groups of Learners in a Corpus-based Second Language Vocabulary Learning Experiment[6]

1. Introduction

Most quantitative analyses can be categorized into two main types: variable-centred and person-centred (see Bergman & Magnusson, 1997). The former is used to examine associations between variables, whereas the latter is used to identify groups of individuals with similar values across variables. To illustrate this using the main topic of the dissertation as an example, a variable-centred analysis would investigate whether providing different types of glossary information is associated with L2 vocabulary learning, whereas a person-centred analysis would explore whether there exist different learner types when various types of glossary information are provided.

The use of person-centred analysis in the field of L2 research dates back to the early 1980s. It has been adopted to identify hidden patterns or groups among L2 learners in terms of L2 aptitude (e.g., Hummel & French, 2016; Skehan, 1986), L2 motivation (e.g., Csizér & Dörnyei, 2005; Papi & Teimouri, 2014), and L2 learning approaches or strategies (e.g., Yamamori, Isoda, Hiromori & Oxford, 2003). By using a battery of language tests and questionnaires, these studies have revealed that it is possible to identify different affective, cognitive, and achievement profiles of L2 learners.

[6] The text of this chapter is a reprint of the material as it appears in Lee, H., Warschauer, M., & Lee, J. H. (2018a). Advancing CALL research via data mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL*. Advance online publication. https://doi.org/1.1017/S0958344018000162. I was the primary investigator and author of this paper, and the co-authors directed and supervised research which forms the basis for the paper.

The primary goal of Chapter 3 is to unearth hidden groups of learners in a corpus-based L2 vocabulary learning experiment using a data mining technique that is frequently used in the field of computer science. To this end, I employed a model-based clustering technique, which uses statistical criteria to determine an optimum number of groups. This feature distinguishes itself from traditional clustering methods used thus far in the field of L2 research (e.g., hierarchical clustering or partitioning clustering; Csizér & Dörnyei, 2005; Skehan, 1986). Although these traditional clustering methods have potential to exert strong power in identifying similar groups of learners "based on strength of and relationships among several [outcome] variables" (Papi & Teimouri, 2014, p. 495), they are considered heuristic because they are not based on formal models, thus requiring researchers to make a subjective decision on the optimal number of clusters (see Meila & Heckerman, 2001; Witten, Frank, Hall & Pal, 2016). Adopting more advanced techniques from a cutting-edge data mining approach might help us unearth hidden groups or patterns from a dataset in a more reliable and precise way.

Moreover, scant attempts have been made in L2 and computer-assisted language learning (CALL) research to understand possible hidden groups or patterns from data obtained from *experimental* designs aimed to compare effects of different types of treatment. To address this paucity, I used data obtained from a previous experiment in an instructed L2 context, in which concordance lines excerpted from corpora were provided as glossary information in CALL reading environments (Chapter 2). On the assumption that learners would make meaning inferences of target vocabulary by exploring the concordance lines provided (i.e., DDL), their lexical inferences via the multiple contextual examples excerpted from corpora can be more accurate when additional confirmation

opportunities via dictionary definitions of target vocabulary are provided. A previous variable-centred analysis revealed that, on average, DDL was effective but that providing an additional definition glossary led to even more L2 vocabulary gains (Chapter 2). Given that clustering groups of similar individuals in terms of the aforementioned variables (i.e., L2 aptitude, motivation, and strategy use) has provided a new angle on the problem, adopting data mining techniques for experimental data could reveal information that is equally important, if not more so. For example, when there are significant differences between treatment and control groups in terms of outcome variables, it is possible that "there is considerable variation within these groups" (Staples & Biber, 2015, p. 243). A close examination of this variation could thus reveal some interesting findings about the effect of a target treatment on certain groups of learners.

Overall, I believe that adopting a data mining approach makes a significant contribution in that we can glean information about how individuals behave in each learning condition, subsequently allowing us to provide personalized instruction. As Chapter 3 is about CALL reading environments equipped with glossary information with one type of this information being corpus-based input (i.e., concordance lines), in the following sections I review the literature on different learner types in L2 vocabulary learning with different glossary types and corpus-based L2 vocabulary learning, I then present a description of methods, findings, and implications.

1.1 Different learner types in L2 vocabulary learning with different glossary types

There have been continuous empirical efforts in the L2 vocabulary learning and CALL literatures to examine the impact of different glossary types, such as L1 and L2 glosses (e.g., Yoshii, 2006), multimedia glosses (e.g., Lomicka, 1998; Yanguas, 2009),

glosses in different positions on screen (e.g., AbuSeileek, 2011; Chen & Yen, 2013; Lee & Lee, 2013, 2015), and concordance-based glosses (e.g., Chapter 2; Poole, 2012). However, only limited attempts have been made to understand different learner types in L2 vocabulary learning with different glossary types (e.g., Chun, 2001; Plass, Chun, Mayer & Leutner, 1998).

In one example, Plass et al. (1998) developed a computer program called *CyberBush* to improve L2 reading comprehension by offering different types of glossary. Participants in their study were German learners who read a story that consisted of 762 German words. Participants could click on verbal information (i.e., text translated into their first language) and/or multimedia cues (i.e., picture or video clips in their first language) on 24 target lexical items. Based on log-file data of learners' choices of glossary types, the results revealed that on average students achieved better vocabulary gains when they selected both glossary types rather than selecting one or none. Further, Plass et al. found different learner types who showed diverse preferences toward glossary types to maximize their reading comprehension.

In another example, Chun (2001) developed a web program called *netLearn* to investigate L2 learners' preferences between instructor-created glosses (internal glossary) and electronic dictionaries (external glossary). In addition to log-file data of learners' clicks on websites, a small number of participants was asked to conduct think-aloud protocols as they read. Post-intervention interviews were then conducted with a few selected students. The results from both the quantitative and qualitative data revealed that on average participants showed better reading comprehension when they were provided access to both internal and external glossaries than when provided access only to an external

glossary. Further, Chun identified different learner types who had varied beliefs in which glossary type was more helpful and showed contrasting strategies for L2 vocabulary learning during their activities.

Overall, only a few studies in the field of L2 vocabulary learning have examined different learner types in CALL experiments, and researchers have emphasized that more research is needed to tackle this issue. With similar purpose, a data mining technique that can provide a statistical basis for similarity-based aggregating (Witten et al., 2016) could be used to identify different learner types based on data obtained from an experiment.

1.2 Positive impact of corpus-based glossary information and different learner types

Inferring meanings of unfamiliar L2 vocabulary, also known as L2 lexical inferencing, is considered a successful vocabulary learning strategy (e.g., Fraser, 1999; Nassaji, 2003; Schmitt, 2000). For this reason, it is believed that corpus use can promote self-driven L2 vocabulary learning by allowing learners to explore authentic language data on their own or with some level of guided induction (e.g., Cobb, 1999; Johns, 1991; Chapter 2). Learners can discover linguistic features of target vocabulary such as its contextual meanings and collocation patterns as they are induced by multiple contextual examples excerpted from corpora (DDL; Johns, 1991). Recent meta-analysis studies (e.g., Boulton & Cobb, 2017; Chapter 1) have confirmed the overall positive effect of corpus use on L2 vocabulary learning, though they did not distinguish the studies focusing on the effects of independent lexical inferencing and those on the effects of guided induction in DDL.

However, to the best of my knowledge, little has been studied about how learners may differentially benefit from corpus-based glossary information (Boulton, 2009; Flowerdew, 2008; Chapter 2). For example, in Chapter 2, I found four emerging patterns of

vocabulary learning from the participants' interactions with e-glosses. However, the results were not about learner types but about vocabulary types, such that vocabulary items may need different glossary types for them to be learned most effectively. Furthermore, although recent meta-analyses (Boulton & Cobb, 2017; Chapter 1) included a moderator analysis to identify factors related to learner types in corpus-based L2 learning, these studies failed to identify such factors because of limited data and incomplete reporting of the included studies. Nevertheless, I believe that corpus-based glossary information may not be appropriate for all types of learners. For example, there could be different learner types, considering that language data excerpted from corpora could be beyond some learners' comprehension levels (Chapter 1) or could impose different amounts of cognitive load (Lee & Lee, 2015). Thus, Chapter 3 aimed to unearth different learner types in a corpus-based L2 vocabulary learning experiment via a data mining approach.

2. Present study

Data mining is generally categorized into two types: (1) supervised and (2) unsupervised (Witten et al., 2016). A distinguishing factor of these two types is whether data mining is being implemented with or without a response variable (i.e., predefined labels or classifications of observations). For example, supervised data mining is mainly used to predict or classify observations from a new dataset based on the algorithm computed from pre-classified observations in an original dataset. Conversely, unsupervised data mining is mainly used to analyse a given dataset without predefined labels or classifications and identify hidden structures of the dataset. In Chapter 3, I focused on *unsupervised* data mining because the goal was to unearth possible different learner types from experimental data. As part of this approach, I also hypothesized that using a model-

based clustering technique based on statistical distributions and probabilities in identifying clusters (see Fraley & Raftery, 2002) would expand our perspective and advance CALL research.

Based on the suggestions of researchers (e.g., Chun, 2001; Chapter 2; Plass et al., 1998) and limited empirical evidence on learner types (e.g., Boulton & Cobb, 2017; Chapters 1 & 2), I conjecture that there may be different learner types who show different learning patterns to maximize their L2 vocabulary learning. Such differential learning patterns would deviate from previous variable-centred findings on the correlation between the amount of glossary information provided and L2 vocabulary knowledge gains at the group level (Chapter 2). I believe that this study provides a clearer picture on how to explore different learner types when a data mining approach (as part of person-centred analysis) is accompanied by variable-centred analysis. I address the following three guiding research questions:

1. Are there hidden groups of learners in a corpus-based L2 vocabulary learning experiment?
2. If so, how similar or different are they from each other in terms of maximizing their L2 vocabulary learning?
3. What is the role of L2 proficiency in relation to different learner types?

3. Methods

In this study, I used experimental data collected in Chapter 2 where I addressed effects of different glossary types on L2 vocabulary learning, including: (1) concordance lines and definitions; (2) concordance lines only; (3) no glossary – all of which were equipped with English reading tasks in CALL reading environments. In this section, I

describe the methodological aspects involved in comparing these three conditions as well as the details of the data analysis plan.

3.1 Participants

A total of 132 (6 were excluded from original 138 participants due to missing values) L2 undergraduate students with a wide range of academic majors in South Korea participated in this study. They were convenience samples from six intact EFL classes. I adopted a repeated-measures design to have each participant experience all three reading tasks in a random order. To determine students' L2 (English) proficiency, I collected their scores on the Test of English for International Communication (TOEIC) developed by the Educational Testing Service (2016). The average TOEIC score of these participants was 732, indicating that they were at an intermediate level of English proficiency (CEFR B1; TOEIC scores between 550 and 785). According to a *post hoc* power analysis, the sample size was large enough to yield sufficient statistical power for the study design.[7]

3.2 Materials

I extracted three reading texts, each consisting of approximately 500 words, from newspaper articles on social issues from an English textbook (Cunningham, Moor & Carr, 2003). Through a pilot study using students with similar characteristics to the participants of Chapter 2, I determined key features of glossary information (e.g., a preferred and manageable number of sample sentences for students' L2 vocabulary learning) and 30 unfamiliar lexical items from the reading texts, including nouns (e.g., *endowments*, *fib*), verbs (e.g., *traipse*, *tuck*), adjectives (e.g., *mucky*, *dodgy*), and multi-lexical items (e.g., *be*

[7] Faul, Erdfelder, Lang, and Buchner's (2007) *G*Power* software is one of the most popular tools for power analysis. Results indicated that the required sample sizes were 12, 28, or 163 for large, medium, and small effect sizes, respectively, for three conditions under a repeated-measures design with a power of 0.80.

beset with, *in the vicinity of*). Each reading text included 10 of the target lexical items evenly distributed throughout the text.[8]

The reading texts were made compatible with three conditions of glossary information.[9] For the concordance lines and dictionary definitions condition (CODI), each target vocabulary item was hyperlinked to a pop-up window showing three pre-selected concordance lines via a customized tool. After consulting these concordance lines, learners could confirm the inferred meaning through dictionary definition of the target vocabulary item.[10] For the concordance lines only condition (CONC), target vocabulary items were hyperlinked to a pop-up window for concordance lines. For the no glossary or control condition (CTRL), target vocabulary items were underlined without providing lexical information.

3.3 Study design and previous findings

To prevent possible order effects, I computed six order types to distribute the three tasks: i.e., (1) task 1 - task 2 - task 3, (2) task 1 - task 3 - task 2, (3) task 2 - task 1 - task 3, (4) task 2 - task 3 - task 1, (5) task 3 - task 1 - task 2, and (6) task 3 - task 2 - task 1. Participants within condition were then randomly assigned to these order types. After a pre-test of target lexical items, the participants completed the three reading tasks in an ordinary classroom with their own laptops, followed immediately by post-tests of target lexical items (e.g., task 1 - post-test 1 - task 2 - post-test 2 - task 3 - post-test 3). The pre-test

[8] The list of 30 target vocabulary items and their definitions can be found in Appendix 2.2.
[9] To ensure learners' comprehension, I carefully selected concordance lines from the BNC, the OANC, and the Brown Corpus. Detailed information about the process of sentence selection can be found in Appendix 2.1. The reading texts can be found in Appendix 2.3
[10] I used the Merriam-Webster Online Dictionary for dictionary definitions of target vocabulary items.

and each post-test took 5 minutes, and each reading task took 15 minutes. Scheduling considerations did not allow me to include a definition-only condition.

In the post-tests, the participants were required to write the meaning of each target lexical item in either the L1 (Korean) or the L2 (English). Their answers were given a zero point for inaccurate meaning, one point for partially correct meaning, or two points for completely correct meaning.

In Chapter 2, I found that on average the participants achieved higher post-test scores than their pre-test scores and that their gains were statistically significant ($p < .001$) within each of the different conditions. When compared between the conditions, I found that the participants under the CODI condition achieved the highest post-test scores on average, followed by those under the CONC condition and then those under the CTRL condition. The achievement differences between the conditions were statistically significant ($p < .05$). It was also found that there was no order effect and that having different sets of target lexical items in each reading task did not affect participants' overall vocabulary gains (see Chapter 2, for more details).

3.4 Data analysis plan

To answer the first research question, as part of data mining approach, I used a model-based clustering technique with *R* 3.4.2 for Windows along with the *mclust* package version 5.3 (entitled "Gaussian mixture modelling for model-based clustering"; Fraley, Raftery, Scrucca, Murphy & Fop, 2017; Scrucca, Fop, Murphy & Raftery, 2016). The *mclust* package identifies clusters in given data based on pre-defined Gaussian mixture models (GMMs; see Fraley & Raftery, 2002; Jung, Kang & Heo, 2014). Provided that any given data can be multi-dimensional, models can be distinguished from each other in terms of their

distributions, volumes, shapes, and orientations (see Appendix 3.1 for detailed information about pre-defined GMMs included in the *mclust* package).[11]

In general, any statistical technique works best with normally distributed data. Although model-based clustering has been reported to be robust for non-normal datasets (Mun, von Eye, Bates & Vaschillo, 2008; Yeung, Fraley, Murua, Raftery & Ruzzo, 2001), the method assumes that identified clusters are "concentrated locally about linear subspaces" by using pre-defined normally distributed models to statistically unearth hidden groups of cases (Fraley & Raftery, 1998: 586). This indicates that it would also work best with data that are normally distributed, or Gaussian.

For this reason, I checked the normality of students' performance data, such as pre- and post-test scores, by using univariate and multivariate normality tests after applying each of four commonly used data-transformations: logarithm, square root, reciprocal, and reverse score (see Field, 2009; Pires & Branco, 2010, for data-transformations). First, I used a skewness and kurtosis test for each variable (i.e., the Improved D'Agostino Test suggested by Royston, 1991). This test compares the symmetry and flatness of the distribution of a variable to the normal distribution. Second, I used two widely-used tests for multivariate skewness and kurtosis (i.e., the Henze-Zirkler, 1990, and the Doornik-Hansen, 2008, tests). These tests examine "whether the marginal distributions and linear combinations of *x*-variables are normal and whether observations of pairs of *x*-variables show the elliptical appearance of the equal-density contours" (Tacq, 2010: 338; see this reference for more information about the multivariate normal distribution).

[11] Scrucca *et al.* (2016) provide more details on the model-based clustering technique, including examples with short scripts of *R* code.

The results indicated that the square root transformation satisfied the Gaussian mixture assumption for the post-test scores according to the normality tests at the 1% significance level. Although the pre-test scores were important indicators of students' achievement, I found that the data values, which measured participants' unfamiliarity with target vocabulary items, could not be normally distributed, as both the mean and standard deviation were very close to zero. It was further found that gain scores, which can be calculated by subtracting pre-test scores from post-test scores, failed the normality tests under all the data-transformations.[12] As a result, I used the post-test scores for the model-based clustering, and included the pre-test scores as covariates in the following data analysis to take account of the baseline differences.

To answer the remaining research questions, I conducted logistic regression and multiple regression analyses. With pre-test scores and gender as covariates, L2 proficiency scores were analysed to: (1) check how participants of different cluster memberships were similar to or different from each other; (2) understand the role of L2 proficiency in relation to different learner types in different learning conditions. To ensure consistency, all data analyses followed by the model-based clustering used the square roots of the vocabulary test scores.

4. Results

4.1 RQ #1. Hidden groups in a corpus-based L2 vocabulary learning experiment

The results of the data mining technique indicated that there were two clusters across different glossary types for maximizing L2 vocabulary learning potentials.

[12] Furthermore, according to Fitzmaurice, Laird, and Ware (2012), using pre-test scores as covariates is more appropriate than using gain scores for the current study design (i.e., a randomized controlled trial). See Maris (1998) for an in-depth discussion about covariance adjustment versus gain scores.

Specifically, I found the best fit at the two-cluster solution according to the lowest BIC value (BIC = −511.92; see Appendix 3.1 for detailed information about how the optimal model can be determined by BIC via the *mclust* package). This result indicated that there were different groups of learners in terms of the square roots of their post-test scores in three conditions.

4.2 RQ #2. Similarities and differences in L2 vocabulary learning between hidden groups

Table 3.1 presents the descriptive statistics of vocabulary test scores for the two different learner types, as well as the total sample. Corresponding to the overall finding that all participants, on average, gained a statistically significant amount of L2 vocabulary knowledge within each of the three conditions, both the two learner types achieved higher vocabulary post-test scores than pre-test scores in all conditions and the differences were statistically significant ($p < .001$).

Table 3.1
Descriptive Statistics of Vocabulary Test Scores

	Learner Type	Condition	SQRT of Pre-test	SQRT of Post-test	*t*-test[a]	ANOVA[b]
Before clustering	Total Sample (N = 132)	CTRL	.21 (.47)	1.76 (.93)	18.11***	$F(2, 393) = 49.46$***
		CONC	.19 (.40)	2.30 (.97)	25.34***	CTRL < CONC < CODI
		CODI	.13 (.36)	2.85 (.89)	34.79***	
After clustering	Cluster 1 (n = 82)	CTRL	.27 (.52)	2.17 (.65)	21.83***	$F(2, 243) = 44.60$***
		CONC	.22 (.43)	2.77 (.66)	33.30***	CTRL < CONC < CODI
		CODI	.17 (.41)	3.17 (.63)	38.83***	
	Cluster 2 (n = 50)	CTRL	.13 (.36)	1.10 (.94)	6.86***	$F(2, 147) = 22.35$***
		CONC	.13 (.36)	1.54 (.90)	10.83***	CTRL < CONC < CODI
		CODI	.06 (.24)	2.32 (1.00)	15.96***	

Note. Standard deviations are in parentheses. SQRT = square root.
[a] *t*-tests examined whether the square roots of post-test scores were different from the square roots of pre-test scores.
[b] ANOVA results compared the square roots of post-test scores between the conditions.
*** $p < .001$

To further analyse how the two groups of learners were similar to or different from each other across the different glossary types, I conducted an ANOVA by using the square roots of the post-test scores, which were used for the model-based clustering technique. Roughly speaking, the two learner types appeared to exhibit L2 vocabulary learning patterns largely corresponding to the previous finding at the group level (i.e., CTRL < CONC < CODI; see Total Sample row in Table 3.2) but with different magnitudes of gains (i.e., higher or lower gains according to the size of *t*-values from *t*-tests).

First, the finding that the participants performed well under the CONC and CODI conditions, when compared to CTRL, indicates that both the two learner types might have harnessed corpus-based language input as a source of their vocabulary learning, supporting researchers' advocacy for providing concordance lines as glossary information for L2 vocabulary learning (e.g., Frankenberg-Garcia, 2012, 2014; Chapter 2; Poole, 2012).

Second, participants' higher scores under the CODI condition compared to the CONC condition indicated that providing definitions of target lexical items in addition to concordance lines had a positive impact on L2 vocabulary learning. This condition likely enabled learners to confirm their inferred meaning of target lexical items (e.g., Cobb, Greaves & Horst, 2001; Fraser, 1999). This finding also corresponds to Godwin-Jones's (2001b) idea that learners' inaccurate inferences from inductive learning would be minimized if their meaning inferences can later be confirmed through definition of target lexical items.

However, results from the multiple regression analysis with the square roots of the vocabulary post-test scores as the dependent variable and the square roots of the

vocabulary pre-test scores, gender, and L2 proficiency scores as independent variables or covariates reported a slightly different picture.

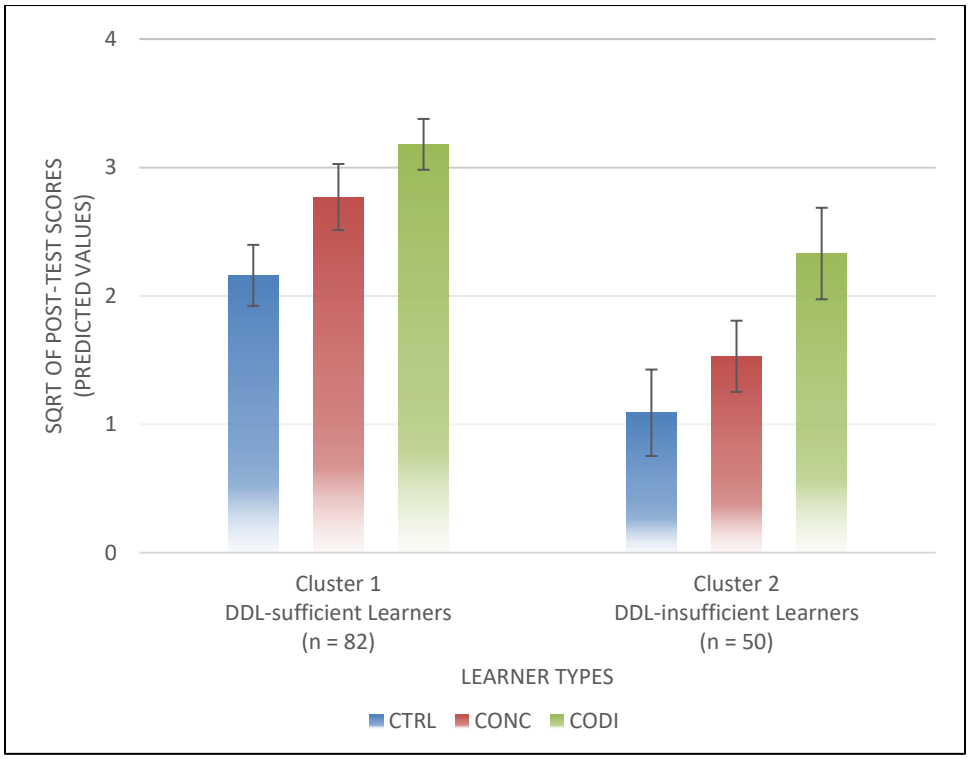

Figure 3.1. Two Clusters (learner types) and Their L2 Vocabulary Learning Patterns
Note. Values are from Table 3.2. Error bars represent 95% confidence intervals.

I obtained bar graphs (Figure 3.1) by using the predicted values of the dependent variable (i.e., the square roots of the post-test scores across three different conditions) from the multiple regression analysis (see Table 3.2). Although the bars largely correspond to the overall finding (i.e., CTRL < CONC < CODI), examination of error bars suggested that some of the differences between the conditions could be statistically insignificant (i.e., overlapping error bars). In brief, the vocabulary gains of Cluster 1 in CONC and CODI conditions and the gains of Cluster 2 in CTRL and CONC were not statistically different, respectively (see Table 3.2). I named these two clusters as follows: (1) "DDL-sufficient learners"; (2) "DDL-insufficient learners."

Table 3.2
Predicted Values of Vocabulary Post-test Scores

	Learner Type	Predicted values			Contrasts at 5% significance level
		CTRL	CONC	CODI	
Before clustering	Total Sample (N = 132)	1.76 (.14)	2.30 (.11)	2.86 (.10)	CTRL < CONC < CODI
After clustering	DDL-sufficient learners (n = 82)	2.16 (.12)	2.77 (.13)	3.18 (.10)	CTRL < CONC ≤ CODI
	DDL-insufficient learners (n = 50)	1.09 (.17)	1.53 (.14)	2.33 (.18)	CTRL ≤ CONC < CODI

Note. Values came from multiple regression analysis with the square roots of the vocabulary post-test scores as the dependent variable and vocabulary pre-test scores, gender, and L2 proficiency scores as independent variables or covariates. See Appendix 3.2 for the complete results of the analysis. Standard errors are in parentheses.

First, I named Cluster 1 "DDL-sufficient learners" (n = 82) to refer to those who achieved the highest post-test scores when they performed DDL under CONC and CODI conditions (no statistically significant difference between these two conditions; t = 2.04, p = .08). This learning pattern indicated that receiving concordance lines as glossary information was sufficient for these learners to be successfully induced to discover the meaning of target lexical items by independently exploring concordance lines. Further, given that this learner type has higher English proficiency (TOEIC scores: M = 759.04, SD = 81.07) than the average L2 proficiency level of the "DDL-insufficient learners" (M = 688.50, SD = 86.39), I can speculate that DDL can be more effective for learners with higher L2 proficiency (Boulton, 2009; Flowerdew, 2015; Leńko-Szymańska & Boulton, 2015).

Second, I named Cluster 2 "DDL-insufficient learners" (n = 50) because they were unlikely to achieve significantly higher scores when they received concordance lines only as glossary information compared to under the CTRL condition where no glossary was provided (no difference between these two conditions; t = 1.52, p = .17). In other words, DDL was not an effective way of learning L2 vocabulary for this learner type compared to "DDL-sufficient learners." As suggested by previous researchers (e.g., Huang, 2011; Lee &

Lee, 2015), students of this learner type may find that the concordance lines provided are incomprehensible or unsuitable in accordance with their relatively low L2 proficiency level (*M* = 688.50, *SD* = 86.39). This assumption, as well as findings of a study by Chun (2001) that participants generally prefer glossary information that requires a lower cognitive load, enabled us me suggest that "DDL-insufficient learners" probably learned target lexical items by leveraging dictionary definitions under the CODI condition (Cobb et al., 2001; Fraser, 1999) without benefiting much from DDL activities.

4.3 RQ #3. Role of L2 proficiency in relation to different learner types

Finally, I used the English proficiency (TOEIC) variable to check if it was related to cluster membership. In addition to the simple comparison between "DDL-sufficient learners" and "DDL-insufficient learners" in terms of L2 proficiency level, I conducted a logistic regression analysis to predict probabilities of a participant falling into specific learner types (see Table 3.3).

Table 3.3
Results of Logistic Regression Analysis

DV: Whether DDL-sufficient learners or DDL-insufficient learners	Total Sample (*N* = 132)	
	Odds Ratio	SE
Condition		
CTRL	0.988	(0.023)
CONC	(reference)	
CODI	1.017	(0.018)
TOEIC	1.010***	(0.002)
SQRT of Pre-test	1.401	(0.559)
Female	0.778	(0.784)
Constant	0.000***	(0.001)

Note. DV = dependent variable. SQRT = square root. SE = standard error.
*** $p < .001$

The results revealed that there was a statistically significant odds ratio of 1.01 ($\chi^2(5) = 77.69$, $SE = .00$, $p < .001$) for "DDL-sufficient learners", indicating that the odds for being this learner type increase about 1% for every one point increase in TOEIC score, and that students with higher L2 proficiency levels were more likely to fall in this learner type. On the other hand, the odds for being the other learner type ("DDL-insufficient learners") increase about 1% for every one-point decrease in TOEIC score; thus, those with lower L2 proficiency levels were more likely to fall into the "DDL-insufficient learner" type.

To further investigate the role of L2 proficiency, I conducted a multiple regression analysis with L2 proficiency as an independent variable for each treatment condition. As shown in Table 3.4, the results indicated that L2 proficiency had positive associations with vocabulary post-test scores for "DDL-sufficient learners" in the CONC ($b = .002$, $SE = .001$, $p < .05$) and CODI ($b = .002$, $SE = .001$, $p < .05$) conditions, and for "DDL-insufficient learners" in the CONC ($b = .003$, $SE = .001$, $p < .05$) condition. Given that L2 proficiency was significantly associated with L2 vocabulary learning across all three conditions including CTRL in the full sample, the null impact of L2 proficiency under CTRL for the two learner types indicated an exclusive role of L2 proficiency in determining the success of DDL tasks. Along these lines, the null impact of L2 proficiency under CODI for "DDL-insufficient learners" partly corroborated one of my speculations about possible absence of DDL efforts of this learner type in CODI. Overall, although the practical magnitude of the impact was not large enough to draw a meaningful interpretation (i.e., small coefficients of the L2 proficiency variable), I found that L2 proficiency played a significant role in influencing successful DDL.

Table 3.4
Role of L2 Proficiency Identified from Regression Analysis

	Learner Type	Coefficient of TOEIC		
		CTRL	CONC	CODI
Before clustering	Total Sample (N = 132)	.002* (.001)	.004*** (.001)	.003*** (.001)
After clustering	DDL-sufficient learners (n = 82)	.001 (.001)	.002* (.001)	.002* (.001)
	DDL-insufficient learners (n = 50)	-.002 (.002)	.003* (.001)	.002 (.002)

Note. Values came from multiple regression analysis with the square roots of the vocabulary post-test scores as the dependent variable and vocabulary pre-test scores, gender, and L2 proficiency scores as independent variables or covariates. See Appendix 3.2 for the complete results of the analysis.
* $p < .05$, *** $p < .001$

5. Discussion

This chapter adopted a data mining approach to unearth hidden groups of learners in an instructed L2 vocabulary learning context using a model-based clustering technique. In doing so, I was able to extend the previous finding that more glossary information led to better overall learning outcomes (Chapter 2), suggesting that identified learner types from the clustering technique may not exactly follow the overall learning pattern. I found that participants exposed to all three learning conditions through a repeated-measures design could be divided into two learner types: (1) "DDL-sufficient learners", whose ability to make use of concordance lines made them suitable for DDL (cf. Johns, 1991), but who did not benefit as much from additional dictionary information as had been expected (e.g., Godwin-Jones, 2001b; Chapter 2); and (2) "DDL-insufficient learners", who did not benefit much from only receiving concordance lines as glossary information. That is, by using data mining, this study shed light on unidentified learner types overshadowed by the *average* obtained through data analysis at the group level. The two learner types had distinctively different learning patterns, so combining them produced a poorly defined "one size fits all"

conclusion. I thus offered support to what Staples and Biber (2015, p. 243) asserted, that clustering techniques can "provide a bottom-up way to identify *new* groups that are better defined with respect to target variables."

In this way, I was able to identify the complex role of learners' L2 proficiency level in this CALL intervention. The model-based clustering technique I employed here is designed to identify statistically more similar and homogeneous groups of learners, and this led me to recognize a strong predictive power of L2 proficiency not only *between* but also *within* the identified learner types. Between the two learner types, I found that students with higher L2 proficiency were more likely to be "DDL-sufficient learners", and those with lower L2 proficiency were more likely to be "DDL-insufficient learners." Within each learner type, data mining helped us find that learners' high levels of L2 proficiency were especially crucial for successful DDL activities, unlike the previous finding at the group level, where students' overall L2 proficiency level is expected to influence the magnitude of L2 vocabulary gains equally across all conditions. That is, I found more direct statistical evidence that DDL can be beneficial to vocabulary learning for L2 learners in general, but the impact can be greater for those with higher L2 proficiency (e.g., Boulton, 2009; Flowerdew, 2015; Leńko-Szymańska & Boulton, 2015). This may bring us one step closer to understanding "the types of learners who take most readily to DDL or extract most benefit from it" (Boulton, 2009, p. 87), corroborating the significant role of L2 proficiency in DDL highlighted by recent meta-analyses (Boulton & Cobb, 2017; Chapter 1).

Given that corpus use for L2 vocabulary learning is "an active, creative, and socially interactive process" (Rüschoff & Ritter, 2001, p. 223), more research is required to understand which learner factors could affect successful DDL activities, such as motivation

(e.g., Gass, Behney & Plonsky, 2013), strategy use (e.g., Tseng & Schmitt, 2008), use of knowledge sources (e.g., Nassaji, 2003), learning styles (e.g., Flowerdew, 2008), and working memory capacities (e.g., Martin & Ellis, 2012). Again, findings of Chapter 3 highlight that results from aggregating individuals (as in the case of data analysis at the group level; Chapter 2) should be interpreted cautiously, as providing only a partial perspective on L2 vocabulary learning.

6. Implications and limitations

This chapter has several implications for future research on L2 learning. It is expected that the use of data mining techniques in analysing experimental datasets will expand research paradigms in several possible ways. First, I recommend the use of data mining techniques in addition to variable-centred statistical analysis. Such an implementation could either produce similar results across different analyses (and thus enhance the validity of the overall result) or present a conflicting finding that could provide useful information about hidden groups of learners with different profiles. Second, researchers may administer a series of pre-tests and questionnaires on individual differences and use data mining techniques to examine if their participants can be clustered in meaningful ways prior to an intervention (see Staples & Biber, 2015, for similar suggestions). Researchers could then further examine possible interaction effects between each cluster and target intervention(s), which could provide valuable findings regarding personalized instruction.

The limitation of Chapter 3 is that I could not fully harness the dataset by excluding the vocabulary pre-test scores in the model-based clustering technique. I used these scores to interpret the results of clustering in a statistically robust way, and yet I wonder how

results might have differed if the pre-test score variable had been normally distributed, in which case it would have been possible to include it in the data mining. According to my clustering simulation with the six variables (i.e., square roots of pre-test and post-test scores for CTRL, CONC, and CODI), I found six clusters with uneven sizes (large differences between the size of the clusters; i.e., 9, 12, 12, 13, 35, and 51, respectively), which could be a convincing sign of bias (see Firooz, 2015, for a discussion). Moreover, the small sample size of the dataset could be another limitation in this case, because the number of variables (i.e., degree of data dimensionality) generally requires a corresponding sample size. Although there is no rule of thumb about the sample size necessary for clustering techniques, one suggests using 5×2^k (k = number of variables) as the minimum sample size (see Dolnicar, 2002, for a review), which would be a minimum of 320 samples for this case. Overall, it is complex and difficult to evaluate findings of unsupervised data mining due to an absence of *true* labels or classifications. For this reason, future studies on this topic with larger samples are warranted to understand how individual learners gain their L2 vocabulary knowledge in diverse ways.

7. Conclusion

Using a data mining technique at the individual level, the results indicated that there were different learner types who exhibited learning patterns that differed slightly from the previous finding at the group level – a positive association between the amount of glossary information provided and post-test scores of L2 vocabulary knowledge. As a result, I identified that individual learners might require different accommodations to maximize their L2 vocabulary learning potentials. For example, for the "DDL-sufficient learner" type, receiving concordance lines was enough for successful L2 vocabulary learning, while the

process of confirming inferred meaning did not substantially contribute to L2 vocabulary learning. The vocabulary post-test scores of "DDL-insufficient learners" also corresponded to a concern that DDL may not be as effective for some learners (Schmitt, 2008). Therefore, closer attention to individual types of learners is required (e.g., Boulton, 2009; Flowerdew, 2015; Lee & Lee, 2015; Leńko-Szymańska & Boulton, 2015). If L2 researchers can implement similar approaches in their future studies, this could contribute to a better understanding of CALL environments equipped with different learning accommodations and the development of more personalized instruction.

CHAPTER 4: Role of Learner Factors in Corpus-based L2 Vocabulary Learning

1. Introduction

This chapter investigated learner factors in corpus-based L2 vocabulary learning, where learners are encouraged to analyze and explore corpora (e.g., Sinclair, 2004) to resolve linguistic inquiries. Also known as DDL, corpus-based L2 vocabulary learning has been becoming increasingly popular, as it offers a vast amount of comprehensible language input to L2 learners (Krashen, 1985; Laufer, 1991). Moreover, corpora, as well as their analysis tools, have become more available and accessible in recent years (e.g., Biber, Conrad, & Reppen, 1998; Lee, Lee, & Sert, 2015; Sinclair, 2004). This growing popularity has been supported by both theory and empirical evidence. First, corpus use has been shown to allow learners to construct their L2 vocabulary knowledge independently by exploring compiled linguistic data such as concordance lines that provide multiple sentential examples of how a target word is used (Johns, 1994). Second, corpus tools (i.e., concordances) display the typed item in the center of multiple concordance lines, a format called "Key Word in Context" (KWIC; see Figure 1.1), and this heavily exposes learners to target vocabulary items (Ellis, 2002). Such exposure makes the target items more salient (Chapelle, 2003) and thus increases the possibility of learner attention to and acquisition of the target items (Schmidt, 2001). Third, cumulative empirical evidence has supported the effectiveness of DDL as an L2 vocabulary learning tool (Bouton & Cobb, 2017; Chapter 1). For example, based on 64 studies, Boulton and Cobb found that DDL was largely effective in L2 learning in general (d = .95). Based on a meta-analysis of 29 studies (Chapter 1), I found that DDL has a moderate impact on L2 vocabulary acquisition (d = .74) and retention (d =

.64). Despite the evidence for DDL's effectiveness in L2 vocabulary learning, one cannot overlook the wide variation among learners regarding their success in general L2 learning (see Dörnyei, 2005). For example, Chapter 3, which used a data mining technique to uncover different L2 vocabulary learning patterns from experimental data, found that learners responded differently to DDL, with significant variations in their L2 vocabulary gains. Similarly, two recent meta-analysis studies (i.e., Boulton & Cobb, 2017; Chapter 1) found that L2 proficiency had a statistically significant effect on corpora-based L2 vocabulary learning.

Along the line, it has been long assumed that learner factors, especially cognition-related ones, play significant roles in corpus-based vocabulary learning because of the heavy cognitive load involved in DDL, as learners may need to autonomously search materials for target linguistic items while being immersed in language data, some of which may be beyond their comprehension (Boulton, 2009a; Chapters 1 & 2). Thus, in addition to L2 proficiency, other learner factors, such as strategy use and working memory, may play significant roles in easing learners' cognitive loads. However, in-depth investigations on how learners differentially construct their L2 vocabulary knowledge during DDL activities is largely absent from the literature (Boulton, 2009a; Chapter 2). To address this gap, Chapter 4 investigated the role of learner factors such as L2 proficiency, strategy use, and working memory in corpus-based L2 vocabulary acquisition and retention.

2. Background

Laufer and Hulstijn (2001) hypothesized that incidental L2 vocabulary learning (usually through reading) occurs when learners notice an unknown word. This involves a search for the meaning of the target word and the connection of its form and meaning,

which results in successful lexical inferencing. Similarly, de Bot, Paribakht, and Wesche (1997) proposed an incidental vocabulary acquisition model (see Figure 4.1) that points to the important role of *cognition* in L2 vocabulary learning. According to this model, a learner may proceed in the following steps when confronted with target vocabulary: (1) the mental lexicon components determine if a given word is unknown; (2) when successfully decoded, the target word's string of letters (i.e., form) needs to be matched with a lexeme in the mental lexicon, which then has to be matched with the syntactic and semantic features of the target word; and (3) finally, comprehension of the target word will be successful if the lemma is connected with one or more concepts. As de Bot et al. (1997) highlighted, the interactions between these steps may not constitute a simple linear relation but a complex process that requires various types of strategies and knowledge sources to bridge the gap between the form and meaning of a target word. Considering that multiple examples are given to learners for their lexical inferencing in corpus-based vocabulary learning, it is evident that cognition-related learner factors do, in fact, matter in successful DDL.

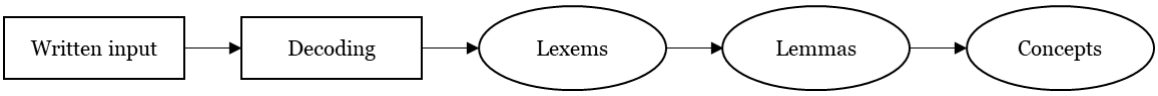

Figure 4.1 Incidental Vocabulary Acquisition Model, adapted from de Bot et al. (1997).

Moreover, as Johns (1991, 1994, 1997) and Lewis (1993) suggested, a DDL activity is not merely reading example sentences containing target vocabulary, but a learning process that can be considered "an active, creative, and socially interactive process" (Rüschoff & Ritter, 2001, p. 223), which comprises the following three stages: (1) learners observe and research the L2 learning materials (the "observe" stage); (2) learners build a hypothesis about language features, such as contextual meaning and syntactic usage (the

"hypothesize" stage); and (3) learners test their hypothesis through practice, improvisation, or classification (the "experiment" stage). That is, the learning process of corpus-based L2 vocabulary learning involves continuous *cognitive* efforts to observe, hypothesize, and experiment with multiple inferences for successful vocabulary learning. For this reason, this process may be influenced by cognition-related learner factors, such as L2 proficiency, strategy use, and working memory, each of which will be discussed below in terms of their relevance to corpus-based L2 vocabulary learning.

2.1 L2 proficiency and corpus-based L2 vocabulary learning

In his review, Boulton (2009a) documented that some researchers had claimed that DDL placed a heavy cognitive load on learners because of its content and KWIC format (i.e., multiple concordance lines of target vocabulary). For example, multiple concordance lines shown in the corpus analysis programs (see Figure 1.1) are randomly selected regardless of levels of difficulty and relevance (Boulton, 2009a; Chapter 1). To allay these concerns, researchers have suggested using customized (e.g., Cobb, 1997; Lee et al., 2015) or graded reader corpora (e.g., Allan, 2009), providing simplified (e.g., Poole, 2012) or pre-selected (e.g., Frankenberg-Garcia, 2014; Lee & Lee, 2015) concordance lines, or offering an additional confirmation opportunity to double-check learners' lexical inferencing (e.g., Chapter 2; Godwin-Jones, 2001). Generally, these suggestions have been effective in improving L2 vocabulary learning through DDL, and this could indicate that L2 proficiency is crucial to the success of DDL.

In Chapter 1, I concluded that DDL was generally effective for learners with different L2 proficiency levels but that learners with higher L2 proficiency benefited the most from DDL. Chapters 2 and 3 offered similar conclusions and well explained the complex role of

L2 proficiency in corpus-based L2 vocabulary learning. First, in Chapter 2, I conducted a corpus-based experiment, where L2 learners received concordance lines as glossary information, and found that participants demonstrated higher L2 vocabulary gains on average in the treatment condition (i.e., they received concordance lines as glossary information) than the control condition (i.e., no glossary information was received), and their achievement was significantly associated with their L2 proficiency. Second, in Chapter 3, I re-analyzed their previous data (Chapter 2) to investigate this at the individual level and employed data mining techniques to reveal hidden patterns. The results revealed two different groups of learners based on vocabulary gains, and the group of learners with higher L2 vocabulary gains was found to have significantly higher L2 proficiency than the other group. Taken together, as several other researchers have argued (Boulton, 2009a; Flowerdew, 2015; Leńko-Szymańska & Boulton, 2015), in Chapters 2 and 3, I found that DDL is generally effective across different proficiency levels, even for lower-level (e.g., Boulton, 2009b) or beginner-level (Vyatkina, 2016) learners; however, its effectiveness increases among high-proficiency learners.

2.2 Strategy use and corpus-based L2 vocabulary learning

Johns (1991, 1994) highlighted that the DDL approach is based on inductive learning *strategies*, in that learners observe linguistic input, perceive similarities among and differences across concordance lines, and hypothesize and test their lexical inferences (i.e., the observe, hypothesize, and experiment stages). Furthermore, Sun (2003) found that learners who were familiar with inductive learning and thinking strategies tended to explore concordance lines better. Still, there is little empirical evidence explaining how and when learners use strategies in successful DDL.

Partly because strategy use is a malleable and teachable factor in successful language learning (Schmitt, 2000, 2008), the literature suggests that strategy use is one of the most important learner factors in L2 vocabulary learning (see Tseng, Dörnyei, & Schmitt, 2006; Tseng & Schmitt, 2008). Moreover, there have been continuous efforts to investigate learners' use of inferencing strategies in L2 vocabulary learning (e.g., Anvari & Farvardin, 2016; Fraser, 1999; Nassaji, 2003; Shen, 2017).

Nassaji and colleagues have conducted several empirical studies to explore the role of lexical inferencing in DDL. Based on previous studies, Hu and Nassaji (2014) defined 12 lexical inferencing strategies that could be divided into four categories (e.g., Haastrup, 1991; Hu & Nassaji, 2012; Huckin & Bloch, 1993; Nassaji, 2003, 2004, 2006; Paribakht & Wesche, 1999; Pressley & Afflerbach, 1995 as cited in Hu & Nassaji, 2014). This included (1) form-focused strategies: analyzing (i.e., "analyzing a word using knowledge of prefixes, suffixes, punctuation, or grammar," Hu & Nassaji, 2014, p. 30), associating (i.e., "attempting to infer the meaning of the target word by associating the word with other similar words," p. 30), and repeating (i.e., "repeating the target word or part of the text containing the target word out aloud," p. 30); (2) meaning-focused strategies: using textual clues (i.e., "guessing the meaning of the target word by using the surrounding context clues," p. 30), using prior knowledge (i.e., "using prior knowledge or experience to infer the word meaning," p. 30), and paraphrasing (i.e., "paraphrasing or translating part of the text that contains the target word," p. 30); (3) evaluating strategies: making inquiries (i.e., "questioning their own inferences," p. 31), confirming/disconfirming (i.e., "confirming or disconfirming the inferences made by using the information in the text," p. 31), and commenting (i.e., "making evaluative comments about the target word ," p. 31); and (4)

monitoring strategies: stating the failure/difficulty (i.e., "making statements about the failure of inferencing or the difficulty of the target word," p. 31), suspending judgment (i.e., "postponing the inference making and leaving it for a later time," p. 31), and reattempting (i.e., "discarding the old inference and attempting to make a new one," p.31). Dividing learners into successful and unsuccessful groups based on their lexical inferencing skills, they found a statistically significant association between frequent use of monitoring strategies and successful lexical inferencing.

This set of lexical inferencing strategies has been widely adopted in studies on strategy use in L2 vocabulary learning (e.g., Anvari & Farvardin, 2016; Hermagustiana, 2017). Hermagustiana (2017) replicated Hu and Nassaji (2014) using a reading task with 10 target words. Findings from a think-aloud protocol confirmed the use of 12 strategy types in four major strategy categories. In addition, Anvari and Farvardin (2016) found that the quality of learners' strategy use played an important role in successful lexical inferencing.

Provided that DDL does not only require learners to explore multiple sentence contexts, but also involves multiple lexical inferencing processes, I assume that there will be unique DDL-related strategies that will influence learner success in addition to the previously defined lexical inferencing strategies. In addition, DDL may place a heavier cognitive load on learners' lexical inferencing than incidental vocabulary learning in a single context because of additional learning processes; therefore, management of the resulting cognitive load may be related to working memory.

2.3 Working memory and corpus-based L2 vocabulary learning

Defined as "the temporary storage and manipulation of information that is assumed to be necessary for a wide range of complex cognitive activities" (Baddeley, 2003, p. 189), working memory is one of the major factors that contribute to individual differences in L2 learning (e.g., Martin & Ellis, 2012; Williams, 2012; see Linck, Osthus, Koeth, & Bunting, 2014 for a review). Likewise, Hu and Nassaji (2012) highlighted that more research is needed on the role of other learner factors, including working memory, to further understand successful L2 lexical inferencing. Provided that working memory is directly and indirectly related to vocabulary learning as part of foundational cognition (Kim, 2017), I believe that learners' cognitive capacities are crucial for corpus-based L2 vocabulary learning, not only in terms of "memory storage, attentional control, and manipulation of information in the service of complex cognition" (Tasi, Au, & Jaeggi, 2016, p. 69) but also "encoding, maintenance, and manipulation of speech-based information" (Gathercole, Willis, Emslie, & Baddeley, 1992, p. 887). The latter capacity is also known as verbal working memory (Gathercole, Alloway, Willis, & Adams, 2006), and Chapter 4 refers to this view of working memory in the establishment of the DDL model.

In L2 research, as Martin and Ellis (2012) summarized, one's capacity to store and process verbal information is referred to as working memory, whereas phonological short-term memory is referred to as the capacity to store memories only. As a sub-component of working memory, phonological short-term memory is known to be another predictor of L2 learning ability, as it holds memory traces for several seconds before they fade from the phonological store (Baddeley, 2003). It has been reported that working memory and phonological short-term memory play independent roles in successful L2 learning

(Gathercole et al., 2006; Kormos & Sáfár, 2008). In a study with adolescent L2 participants of an intensive, 1-year English language program, Kormos and Sáfár (2008) confirmed that working memory generally affected their L2 vocabulary learning, whereas phonological short-term memory played a limited role in higher proficiency-level vocabulary learning. Martin and Ellis (2012) found somewhat different results, as the adult L2 learners in their study demonstrated a positive correlation between working memory and L2 vocabulary production, as well as a positive correlation between phonological short-term memory, L2 vocabulary production, and comprehension.

In sum, the empirical evidence suggests that working memory and phonological short-term memory are related but independent components of successful L2 vocabulary learning, although the research on how these two constructs interact in language learning remains inconclusive (Martin & Ellis, 2012; Weekes, 2018). In Chapter 4, I have focused solely on working memory and its role in corpus-based L2 vocabulary learning.

3. Present study: A hypothesized model of DDL

Chapter 4 investigates the role of learner factors in corpus-based L2 vocabulary learning, with the aim of establishing a model of DDL. To this end, the goals are to explore (1) what types of lexical inferencing strategies are used and how they are used by learners in successful DDL activities; (2) how learner factors, such as L2 proficiency, strategy use, and working memory, work together in successful corpus-based L2 vocabulary acquisition and retention; and (3) whether my hypothesized model of DDL (see Figure 4.2) fits the collected data—where L2 proficiency and strategy use directly and indirectly contribute both to L2 vocabulary acquisition and retention, and working memory directly contributes to vocabulary acquisition while indirectly contributing to retention.

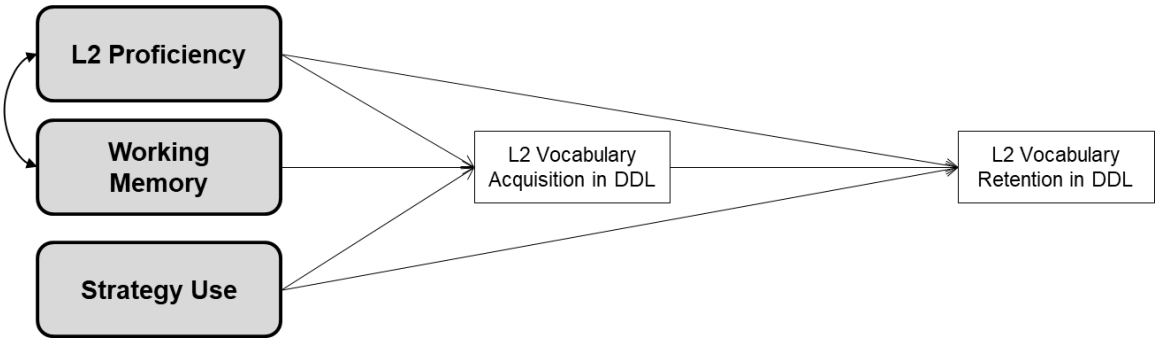

Figure 4.2 A Hypothesized Model of Data-Driven Learning.

In addition, I hypothesize that (1) learners will demonstrate unique DDL-focused strategy use as a task-specific cognitive learner factor (Johns, 1991, 1994; Lewis, 1993), and these strategies will be related to successful DDL (e.g., Sun, 2003); (2) learners with higher L2 proficiency and working memory will benefit more from DDL than those with lower L2 proficiency and working memory (Chapters 1 & 2; Gathercole et al., 2006); (3) L2 proficiency and working memory will be closely related and, as general cognitive learner factors, they will help to manage cognitive load (e.g., Gathercole et al., 2006; Kormos & Sáfár, 2008); and (4) working memory will not have a direct contribution to retention because it is temporal in nature (Baddeley, 2003).

4. Methods

4.1 Participants

A total of 35 college students participated in this study.[13] In an elective English teaching related course where 60 students from a wide range of majors enrolled, a brief

[13] The sample size may not be large enough for an SEM model based on the rule-of-thumb for the minimum sample for multivariate analyses, such as the 10 cases per variable rule or the $5 \times 2k$ rule (k = number of variables; see Dolnicar, 2002 for a review). However, recent simulation studies (i.e., Sideridis, Simos, Papanicolaou, & Fletcher, 2014; Wolf, Harrington, Clark, & Miller, 2013) have emphasized the limitations of commonly cited rules-of-thumb and recommended, rather, that small sample sizes are sufficient (e.g., 30 cases for a one-latent-variable SEM model with four variables; Wolf et al., 2013). Furthermore, the required sample size for an SEM model is largely affected by how much the data set satisfies statistical assumptions

introduction of the research, including its purpose and objective, procedures, and compensation, was announced in the first class, and the participants volunteered after completing the informed consent process. Except for one student majoring in Economics, 34 students majored in English education. Their ages ranged between 19 (sophomores) and 21 (juniors and seniors), and they generally had around 10 years of formal English learning experience (i.e., in elementary, middle, and high school). One participant had to withdraw from the study due to personal reasons, so 34 students completed the necessary materials and tasks (i.e., there were no missing values) and were compensated for their time.

4.2 Reading passage, target vocabulary, and concordance lines

To ensure successful DDL, learners should be able to comprehend the given reading passage to infer and acquire the meanings of target vocabulary. Thus, I first chose a passage entitled "What didn't come to pass" excerpted from Cunningham, Moor, and Carr (2003). Second, I analyzed the text and selected nine target vocabulary items, including three verbs (i.e., crack, traipse, tuck), adjectives (i.e., dodgy, lumbering, mucky), and nouns (i.e., cryogenics, double-glazed windows, grannies). Third, I retrieved lists of concordance lines of each target vocabulary item to select the five most comprehensible and suitable sample sentences for the participants.[14] See Appendix 4.1 for the reading passage and Appendix 4.2 for the selected concordance lines.

(e.g., multivariate normality), if the data has any missing values (Schreiber, Nora, Stage, Barlow, & King, 2006), and, more importantly, whether the data obtains an overall good model fit.

[14] I used the process of selecting example concordance lines used in Chapter 2, which is as follows: (1) sample sentences should be comprehensible for students and should not have unfamiliar words and phrases, (2) sample sentences should have obvious clues to infer the contextual meaning of target vocabulary used in the passage, and (3) sample sentences that may induce faulty or irrelevant meaning inferences should be excluded.

4.3 Measures

4.3.1 L2 proficiency

To measure the participants' L2 proficiency, I used the Vocabulary Size Test developed by Nation and Beglar (2007). As a multiple-choice vocabulary test, it measures L2 vocabulary knowledge by asking learners to select one of four given choices, such as words, expressions, or phrases that best match the target word. For non-native speakers of English, the test consists of 140 items that cover 14,000 word families sampled from British National Corpus (BNC) frequency lists. The maximum possible score for the test is 140. In terms of validity, the reported Rasch-based reliability measure is .96 (Beglar, 2010).

4.3.2 Working memory

A listening span task (adapted from Martin & Ellis, 2012) was used to measure the participants' working memory. The participants listened to sets of sentences (ranging from to two to four sentences each) and had to decide whether each sentence was grammatically correct. After each set of sentences, they were asked to recall the last word (which were all one-syllabled) of each sentence. For a practice trial, they listened to three sets of two sentences along with the feedback. They heard a total of 12 sets (four sets of two, three, and four sentences in random order); therefore, the maximum possible score for this task was 36. Cronbach's alpha (α) across these 36 items was .76.

4.3.3 Vocabulary tests

Pre-, post-, and follow-up vocabulary tests were conducted before, during, and after the DDL task (in 2-week intervals) to measure participants' prior knowledge, vocabulary acquisition, and retention, respectively. To ensure the reliability of the scoring, 10 post-vocabulary tests (29% of 34 post-vocabulary tests) were randomly selected and given to an

L2 vocabulary researcher (i.e., the second rater). A total of 2 points were allotted for each item, with 2 points awarded for a correct answer, 1 point for a partially correct answer, and 0 points for an incorrect answer. Therefore, the maximum possible score for each vocabulary test was 18. Cohen's kappa coefficient (*k*) for the inter-rater reliability was .92 (SE = .05, $p < .001$).

4.3.4 Reading comprehension test

A reading comprehension test was used to ensure that students' DDL activities did not interfere with their understanding of the text. The students were required to summarize the text they had read. To ensure high inter-rater reliability, the second rater and I scored 10 randomly selected reading tests (29% of 34 reading comprehension tests) together using a 4-point scale. Cohen's kappa coefficient (*k*) for inter-rater reliability was .88 (SE = .16, $p < .001$).

4.4 Procedure

During the first session, the participants completed a consent form and then received a short introduction on the study. Next, the participants completed the vocabulary size test and the pre-vocabulary test on the target vocabulary. Two weeks later, in the second session, the researcher and a research assistant met each student individually. The participants completed the listening span task, performed a DDL activity after a brief training session, and took the post-vocabulary and reading comprehension test. Two weeks later, in the third session, the participants took the follow-up vocabulary test.

5. Data analysis and results

To achieve the research goals, I used a mixed-method approach. The qualitative component included observing, coding, and analyzing learners' L2 lexical and DDL strategy

use. The quantitative component included measuring learner factors, examining L2 vocabulary acquisition and retention, and identifying associations among factors to fit the hypothesized model.

5.1 Qualitative component: L2 lexical inferencing and DDL strategy use

To investigate strategy use in corpus-based L2 vocabulary learning, I used a think-aloud protocol, which has been widely used in applied linguistics to investigate learners' thinking during an L2 task (Ericsson & Simon, 1993). Thus, each participant was asked to verbalize their thoughts during their DDL activities. Before performing the task, they watched a video clip recorded by myself on how to perform a think-aloud protocol. During their activities, they received feedback when necessary.

Think-aloud protocols are often combined with other methods (e.g., video-recording) to triangulate findings (Deschambault, 2017), as learners often do not verbalize all of their thought processes. In Chapter 4, the think-aloud process was video-recorded upon the students' consent. When the learners did not speak, the video data helped me to check visual clues of their DDL activities (e.g., eye-gaze patterns).

To ensure inter-coder reliability, the second rater and I qualitatively analyzed 10 randomly selected video clips (29% of 34 video clips) to determine qualitative codes and themes. We referred to the 12 lexical inferencing strategies suggested by Hu and Nassaji (2014), and used an inductive approach to identify any emerging codes. We utilized Microsoft Word's Memo feature to mark and label the codes on the transcripts, and then used Microsoft Excel for the coding framework when an agreement was reached.

For the first cycle, we used both process coding (i.e., coding for a word or phrase that captures action) and simultaneous coding (i.e., providing multiple codes for the same

text; see Saldaña, 2016) to capture how the participants responded to the language input using strategies. We found nine (grouped into three categories) of the 12 lexical strategies (Hu & Nassaji, 2014) in our data: analyzing, associating, and repeating (form-focused strategies); using textual clues, using prior knowledge, and paraphrasing (meaning-focused strategies); and making inquiries, confirming/disconfirming, and commenting (evaluating strategies). In addition, we identified three unique DDL-focused strategies: exploring, cross-checking/double-checking, and synthesizing. Table 4.1 displays the strategies used by the participants during the DDL activities.

Table 4.1
The 12 Lexical Inferencing Strategies in Corpus-Based L2 Vocabulary Learning

Category	Freq.	%	Strategy	Freq.	%
Form-focused strategies	158	14%	Analyzing	45	28%
			Associating	66	42%
			Repeating	47	30%
Meaning-focused strategies	247	21%	Using textual clues	166	67%
			Using prior knowledge	28	11%
			Paraphrasing	53	22%
Evaluating strategies	201	17%	Making inquiry	27	13%
			Confirming/disconfirming	22	11%
			Commenting	152	76%
DDL-focused strategies	562	48%	Exploring	269	48%
			Cross-checking/double-checking	173	31%
			Synthesizing	120	21%
Total	1,168	100%			

For the second cycle, we used pattern coding to understand how the three newly found codes under the DDL-focused category were related to other strategies and categories. Most notably, we found that the DDL-focused strategies were used mostly *between* concordance lines, whereas the remaining nine strategies were used *within* concordance lines. Furthermore, we found that the DDL-focused category was the most

frequently used, and its three strategies were used more often than other strategies, except for strategies involving the use of textual clues and comments.

Table 4.2
Three DDL-focused Strategies and Their Definitions and Examples

Strategy	**Definition**	**Example**
Exploring	Reading multiple example sentences to infer the word meaning while judging the difficulties or relevancies of the sentences.	Target Word: lumbering (ID: 26-13) **[Sentence #1]** I don't get it. **[S #2]** It is an adjective and seems to relate to something old. **[S #3]** Slow? **[S #4]** Something old and slow.
Cross-checking/ Double-checking	Revisiting example sentences to check or confirm previous inferences after another DDL.	Target Word: cryogenics (ID: 30-13) [S #1] TW is a noun. [S #5] Hmm, TW is about freezing people. [S #1] No, this is not about that. **I need to see another sentence then.** [S #4] I think TW is about freezing and defrosting people. **Let's go back. [S #1]** It does not make sense here. **[S #3]** It is a field of research related to freezing people for medical purposes. **Oh, now it makes sense. [S #1] So, this sentence** is about a salesman who works in this field of research. Now it makes sense.
Synthesizing	Making conclusive comments about the TW based on previously made multiple inferences and judgments made by DDL.	Target Word: cracked (ID: 6-16) Okay, in the first sentence he cannot break the level, the third sentence is about getting into the hall of fame, and the fourth sentence is about going beyond a wall or barrier. **Taken all together**, I think the meaning of TW is to go beyond or breakthrough a level or barrier.

Note. Bold text is specific to each DDL-focused strategy.

Furthermore, we found that participants used monitoring strategies, such as stating the failure/difficulty, suspending judgment, and reattempting strategies suggested by Hu and Nassaji (2014), but these strategies were used as DDL-focused strategies, as shown in Table 4.2. For example, the exploring strategy describes when learners read multiple example sentences to infer word meaning while judging the difficulties or relevancies of the sentences, which involves stating the failure/difficulty strategy. The cross-checking/double-checking strategy involves learners revisiting example sentences to check

or confirm previous inferences, which includes the reattempting strategy. The synthesizing strategy describes when learners make conclusive comments about the target vocabulary based on previously made, multiple inferences and judgments through DDL, which involves the suspending judgment strategy. After this initial phase of coding, Cohen's kappa coefficient (*k*) for the inter-coder reliability reached .86 (SE = .03, $p < .001$).

5.2 Quantitative component

5.2.1 Descriptive statistics and correlations

After identifying and quantifying learners' strategy use, structural equation modeling (SEM) was employed to examine if the collected data fit the hypothesized model of DDL. I first explored the relationships among the variables to understand how they work together to contribute to successful corpus-based L2 vocabulary acquisition and retention using descriptive statistics and correlations. It should be noted that I did not use each of the 12 strategies as a variable for the strategy use aspect, considering that the sample size of Chapter 4 was not large enough to accommodate many variables; therefore, I used four variables––the form-focused, meaning-focused, evaluating, and DDL-focused strategies––to assess learners' strategy use.

The descriptive statistics and correlations are displayed in Table 4.3. First, the descriptive statistics indicated that, on average, the participants acquired about five or six new vocabulary items and retained about three or four new vocabulary items. For L2 proficiency, the results indicated that participants had an average vocabulary size of around 7,650 word families.[15] Last, for working memory, the participants had an average

[15] According to Nation and Beglar (2007), successful ESL college students at an English-speaking university had a vocabulary size of around 5,000-6,000 word families.

listening span task score of 25.59, getting about 71% of the 36 tasks correct. correlations did not differ much from my hypothesis. The post-vocabulary test was significantly related to the follow-up vocabulary test (r = .84). In terms of the independent variables, the post-vocabulary test was related to L2 proficiency (r = .36), DDL-focused strategy use (r = .46), and working memory (r = .40). In the case of the follow-up vocabulary test, it was not related to working memory as I expected, but to strategy use, such as the meaning-focused (r = .41) and DDL-focused (r = .44) strategy types. Contrary to my expectation, it was not related to L2 proficiency. Concerning the independent variables, I found that L2 proficiency was significantly related to working memory (r = .48). Strategy use was unrelated to L2 proficiency or working memory in general, except for a significant association between working memory and form-focused strategy use (r = .35). More importantly, I found that the four categories of strategy use were not related to each other, although they were all used by the participants in their DDL activities.

Table 4.3
Descriptive Statistics and Correlations

		Post-vocab. test	Follow-up vocab. test	L2 vocab. size	Strategy use				WM
					F.	M.	E.	D.	
Post-vocabulary test		-							
Follow-up vocab. test		.84***	-						
L2 proficiency		.36*	.31	-					
Strategy use	Form	.16	.18	.26	-				
	Meaning	.29	.41*	–.16	.20	-			
	Evaluating	.30	.21	–.07	–.22	.19	-		
	DDL	.46**	.44**	–.15	.24	.23	.29	-	
Working memory		.40*	.31	.48**	.35*	.11	.04	–.03	-
Mean		11.35	7.44	76.5	4.64	7.26	5.91	16.53	25.59
(SD)		(3.49)	(3.55)	(12.08)	(2.10)	(2.55)	(3.18)	(3.46)	(4.79)

Note. Values are correlation coefficients and those without an asterisk(s) are non-significant.
* p < .05, ** p < .01, *** p < .001

5.2.2 Testing the assumptions of SEM

SEM was employed as the primary data analysis method to examine if the collected data fit the hypothesized model of DDL. Before employing SEM, five statistical assumptions for SEM were checked (Acock, 2013; Kline, 2012), beginning with a check for assumptions regarding sub-regression models for SEM. Shadish, Cook, and Campbell (2002) suggested that (1) the residuals (errors) be identically and independently distributed (i.e., normality of residuals; Shapiro-Wilk test, 1965; $p > .05$), (2) the variance of the residuals should be constant across all values of the independent variables (i.e., homoscedasticity of residuals; Cameron & Trivedi's test, 1990; $p > .05$), and (3) the independent variables should not be linear combinations of one another (i.e., multicollinearity; variance inflation factors—VIF—for each independent variable should not be greater than 5). Next, Acock (2013) and Kline (2012) suggested confirming that (4) the joint distribution of the dependent variables is multivariate normal (Henze-Zirkler test, 1990; $p > .05$). Furthermore, the model fit indices of an SEM model should be checked to ensure that it is legitimate for valid inferential statistics (Acock, 2013; Kline, 2012). That is, it should be confirmed that (5) the model fits the data by satisfying goodness of fit indices, such as a chi-square test (the model fit should not be significantly poorer than a saturated model, $p > .05$), root mean square error of approximation (RMSEA < .05), comparative fit index (CFI > .95), Tucker Lewis index (TLI > .95), and standardized root mean square residual (SRMR < .05).

In Chapter 4, the hypothesized SEM model had two dependent variables (i.e., post- and follow-up vocabulary tests); therefore, assumption checks were conducted for each regression model. For the first model with the post-vocabulary test as the dependent variable, the data passed the Shapiro-Wilk test (normality of residuals; $p = .23$) and

Cameron & Trivedi's test (homoscedasticity; p = .58). In addition, the mean VIF was found to be 1.70, and the VIF for each variable was less than 5 (Min−Max: 1.32−2.53). For the second model with the follow-up vocabulary test as the dependent variable, the data passed the Shapiro-Wilk test (normality of residuals; p = .42) and Cameron & Trivedi's test (homoscedasticity; p = .33). Moreover, the mean VIF was found to be 1.87, and the VIF for each variable was less than 5 (Min−Max: 1.33−2.55). Furthermore, the two dependent variables passed the Henze-Zirkler test for multivariate normality of dependent variables at a 5% significant level, and the estimated SEM model had acceptable model fit indices (χ^2 = 8.43, p = .39; RMSEA = .04; CFI = .99; TLI = .98; SRMR = .02; see Figure 4.3).

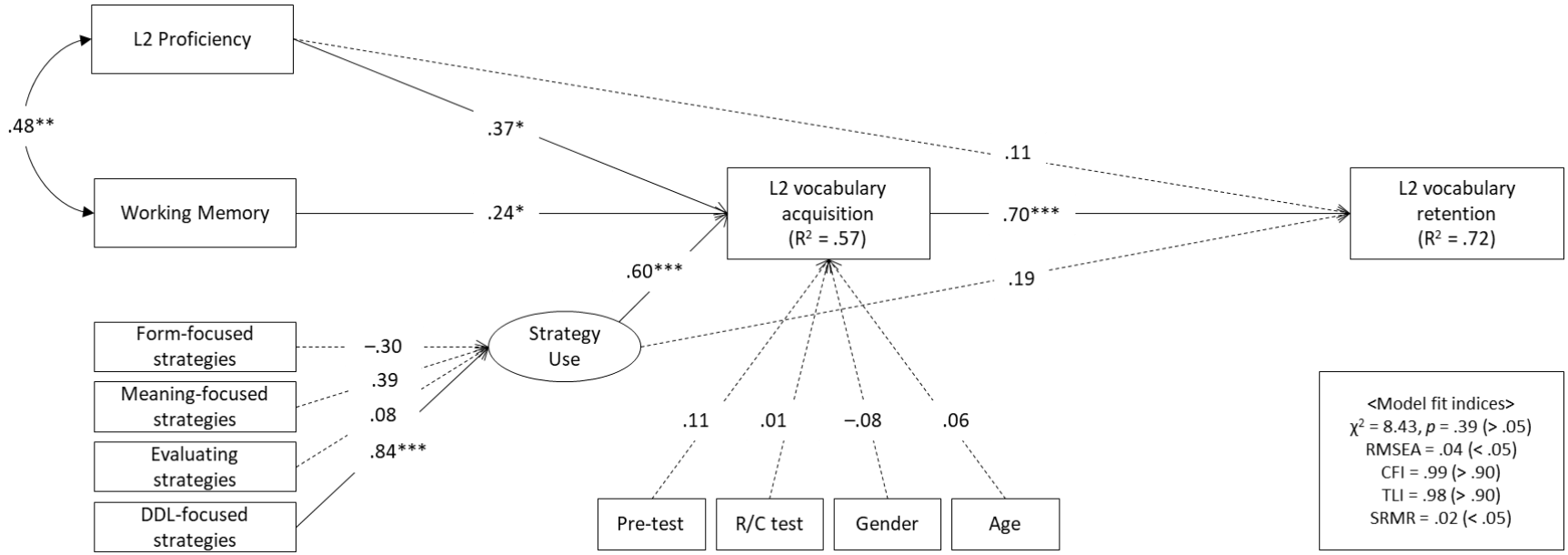

Figure 4.3 A Structural Equation Model for DDL.

Note. Values are standardized path coefficients for the associations of L2 proficiency, working memory, and strategy use to corpus-based L2 vocabulary acquisition and retention after controlling for pre-vocabulary test, reading comprehension test, gender, and age. The solid lines represent statistically significant associations, whereas the dashed lines represent non-significant associations. The model passed the Henze-Zirkler test for multivariate normality of dependent variables at a 5% significance level. Strategy use is a composite latent variable (or a formative construct); therefore, its error variance was fixed at zero.

* p < .05, ** p < .01, *** p < .001

5.2.3 Structural equation model

Figure 4.3 illustrates the estimated SEM model with the associations of L2 proficiency, working memory, and strategy use, such as form-focused, meaning-focused, evaluating, and DDL-focused strategies, to corpus-based L2 vocabulary acquisition and retention after controlling for the pre-vocabulary test, reading comprehension test, gender, and age. According to the *R*-squared results, the model explained about 57% of the variance for L2 vocabulary acquisition and about 72% of the variance for L2 vocabulary retention. Table 4.4 shows the total effects (direct + indirect effects) of the independent variables on corpus-based L2 vocabulary acquisition and retention. First, according to the standardized path coefficients displayed between the variables in Figure 4.3 and the total effects represented in Table 4.4, the results indicated that DDL-focused strategy use (β = .50, p < .001)—mediated by the strategy use latent variable (i.e., .84×.60 = .50)—, working memory (β = .24, p < .05), and L2 proficiency (β = .37, p < .05) were directly related to L2 vocabulary acquisition after controlling for the pre-vocabulary test, reading comprehension test, gender, and age. Form-focused (β = –.30×.60 = –.18, p > .05), meaning-focused (β = .39×.60 = .23, p > .05), and evaluating strategies (β = .08×.60 = .05, p > .05) were not significantly associated with vocabulary acquisition. To identify which learner factors contributed more to vocabulary acquisition, I tested the equalities of the standardized path coefficients (i.e., Wald Chi-Squared Test). Although the effect of DDL-focused strategy use (β = .50) was descriptively larger than that of L2 proficiency (β = .37), followed by that of working memory (β = .24), the results indicated that the differences between these three learner factors were not statistically significant (p > .05).

Table 4.4
Total Effects of Independent Variables on Corpus-Based L2 Vocabulary Acquisition and Retention

Dependent variables	Independent variables		Direct effect	Indirect effect	Total effect
Post-vocabulary test	L2 proficiency		.37* (.15)	(no path)	.37* (.15)
	Strategy use	Form	–.18 (.13)	(no path)	–.18 (.13)
		Meaning	.23 (.14)	(no path)	.23 (.14)
		Evaluating	.05 (.13)	(no path)	.05 (.13)
		DDL	.50*** (.11)	(no path)	.50*** (.11)
	Working memory		.24* (.10)	(no path)	.24* (.10)
Follow-up vocabulary test	L2 proficiency		.11 (.11)	.26* (.11)	.37*** (.11)
	Strategy use	Form	–.06 (.05)	–.13 (.10)	–.18 (.13)
		Meaning	.08 (.10)	.16* (.08)	.24 (.16)
		Evaluating	.01 (.04)	.03 (.09)	.05 (.13)
		DDL	.16 (.13)	.35** (.11)	.51*** (.11)
	Working memory		(no path)	.17* (.08)	.17* (.08)
	Post-vocabulary test		.70*** (.12)	(no path)	.70*** (.12)

Note. Standard errors are in parentheses. Values are standardized path coefficients in the DDL model (Figure 4.3) after controlling for pre-vocabulary test, reading comprehension test, gender, and age. For meaning-focused and evaluating strategy use variables, the difference between the total effects and their parts (direct + indirect effects) is due to rounding. * $p < .05$, ** $p < .01$, *** $p < .001$

Second, concerning L2 vocabulary retention, Table 4.4 displays the total effects of the independent variables. The post-vocabulary test ($\beta = .70$, $p < .001$), DDL-focused strategy use ($\beta = .51$, $p < .001$), L2 proficiency ($\beta = .37$, $p < .001$), and working memory ($\beta = .17$, $p < .05$) contributed significantly to vocabulary retention. Vocabulary acquisition ($\beta = .70$, $p < .001$) was directly related to retention; therefore, as a mediating variable, it allowed the independent variables to indirectly contribute to retention. In the case of meaning-focused strategy use, although it had a significant indirect effect ($\beta = .16$, $p < .05$), it did not have a significant total effect ($\beta = .24$, $p > .05$) after combining with its non-significant direct effect ($\beta = .08$, $p > .05$). Again, I tested the equalities of the standardized path coefficients, but the results revealed that the differences between the post-vocabulary test ($\beta = .70$), DDL-focused strategy use ($\beta = .50$) and L2 proficiency ($\beta = .37$) were not

statistically significant. The total effect of working memory (β = .17) was significantly smaller than that of the post-vocabulary test (χ^2 = 21.74, p < .001) and DDL-focused strategy use (χ^2 = 4.94, p < .05). Finally, the difference between L2 proficiency and working memory was not statistically significant (χ^2 = 1.38, p > .05).

6. Discussion

DDL has received much attention as an effective method to improve L2 vocabulary, in that it does not only provide learners with large amounts of authentic language input for linguistic inquiries (Chapter 1), but also encourages them to develop their L2 vocabulary knowledge independently (i.e., discovery learning; Flowerdew, 2015). Chapter 4 investigated the role of cognition-related learner factors, such as L2 proficiency, strategy use, and working memory, in determining the success of L2 vocabulary learning using DDL. Overall, I found that L2 proficiency, DDL-focused strategy use, and working memory were both directly and indirectly associated with L2 vocabulary acquisition and retention. The findings of Chapter 4 extend our understanding of the learning mechanisms behind DDL, in the following important ways.

6.1 DDL-focused strategy use in DDL

First, I identified learners' use of three unique DDL-focused strategies––exploring, cross-checking/double-checking, and synthesizing––and found that they *largely* contributed both to L2 vocabulary acquisition (β = .50) and retention (β = .51). When compared to other cognition-related factors (i.e., L2 proficiency and working memory), the effect of DDL-focused strategy use was the largest for both vocabulary learning and retention; moreover, its impact on vocabulary retention was statistically on par with the post-vocabulary test (difference: χ^2 = 1.00, p > .05), which had the largest impact on

retention (β = .70). That is, DDL-focused strategy use was one of the most important factors in corpus-based L2 vocabulary learning.

This finding is even more meaningful in that the three DDL-focused strategies largely corresponded to the proposed learning mechanism of DDL (i.e., observe, hypothesize, and experiment stages; Johns, 1991, 1994, 1997; Lewis, 1993). The benefits of learning L2 vocabulary through DDL have been attested in various second language acquisition frameworks, and the findings of Chapter 4 showed that successful DDL learners explored and observed the concordance lines of target words and made multiple inferences about target word meaning, which led them to notice its lexical characteristics (i.e., noticing hypothesis; Schmidt, 2001). Because other concordance lines were presented, the learners actively used the additional opportunities to re-check their preliminary inferences (i.e., frequency effect; Ellis, 2002), which led them to synthesize their multiple inferences to draw conclusions, inducing them to become more involved in the task (i.e., involvement load hypothesis; Laufer & Hulstjin, 2001). Therefore, Chapter 4 sheds light on the ways in which learners use DDL-focused strategies, which have not received in-depth investigation to date; further, it confirmed that these strategies substantially influence the success of corpus-based L2 vocabulary acquisition and retention.

It is thus logical to raise the question of whether DDL-focused strategies are teachable and, if so, how to teach them effectively. Concerning strategy use in L2 vocabulary learning, Hu and Nassaji (2014) suggested that learners should be taught to use strategies appropriate to specific contexts, as there is no universal, perfect strategy. I believe that this suggestion is true regarding DDL strategy training. Moreover, considering that the KWIC format may be unfamiliar to learners, teaching these strategies prior to DDL

would be helpful (e.g., Gavioli, 2009; Chapter 2). According to Chapter 1, providing training opportunities (d = .72) had a higher average effect size than providing no training opportunities (d = .58). While the effect size difference between these two categories was not statistically significant (effect size difference: d = .14, p > .05), it does not mean that providing training opportunities has any negative effect; therefore, I encourage educators to implement necessary training opportunities to ensure successful DDL. Overall, in view of the characteristics of DDL and the KWIC format, it is ideal for DDL strategy training to be "clearly articulated and explicitly modelled by the teacher" (Macaro, 2001, p. 266).

6.2 Role of L2 Proficiency and Working Memory

Second, I found that learners' L2 proficiency and working memory were significantly correlated (r = .48, p < .01) and had significant total effects of similar magnitudes, both on vocabulary acquisition (β = .37 and .24, respectively; difference: χ^2 = .29, p > .05) and retention (β = .37 and .17, respectively; difference: χ^2 = .1.38, p > .05), confirming previous findings that learners with higher L2 proficiency and working memory benefit more from DDL than those with lower capacities (e.g., Boulton, 2009; Flowerdew, 2015; Chapter 3; Leńko-Szymańska & Boulton, 2015 for L2 proficiency; e.g., Gathercole et al., 2006; Kim, 2017; Kormos & Sáfár, 2008; Linck et al., 2014; Martin & Ellis, 2012; Williams, 2012 for working memory).

I hypothesized that L2 proficiency and working memory are general cognitive factors, and would thus have a strong influence on L2 learning in general, unlike DDL-focused strategy use, which is a task-specific and skill-based cognitive factor. For this reason, when it comes to DDL, which places more cognitive load on lexical inferencing than normal incidental vocabulary acquisition from a single context does (e.g., Allan, 2009;

Flowerdew, 2015; Lee & Lee, 2015; Chapter 2), higher levels of L2 proficiency and working memory may better manage cognitive burden during DDL performance. Previous researchers' efforts to ease the cognitive load using pre-selected and simplified concordance lines from customized or graded corpora are a similar case (e.g., Allan, 2009; Cobb, 1997; Chapter 2; Frankenberg-Garcia, 2014; Lee & Lee, 2015; Lee et al., 2015; Poole, 2012).

Although I could not determine exactly how L2 proficiency and working memory worked together to manage cognitive load, the findings led me to assume that when a learner knows more L2 vocabulary word families and has a better verbal working memory, it is much easier for them to explore concordance lines with different levels of difficulty and relevancy by storing inferred word meanings in their mind, to revisit concordance lines to check their inferences, and to synthesize multiple lexical inferences to draw a conclusive word meaning––the essential stages of DDL. If my assumption is correct, this may help to explain at least one of the controversial aspects DDL: the role of L2 proficiency. There has been pervasive concern (see Boulton, 2009a for a summary) that DDL may be ineffective for learners with lower L2 proficiency. However, I believe that learners' DDL-focused strategy use is likely to be the main reason why DDL is unsuccessful, not their L2 proficiency. Rather, it is more likely that the main role of L2 proficiency is to ease cognitive load, so that a learner can better manage the DDL task, which is what I found in Chapter 3. This idea also aligns with Boulton and other researchers' suggestions that DDL is beneficial to all learners, but benefits increase with higher proficiency (Boulton, 2009a, Flowerdew, 2015; Chapter 3; Leńko-Szymańska & Boulton, 2015).

Thus, based on the findings of Chapter 4, continuous efforts to improve learners' L2 proficiency and working memory are needed. Compared to L2 proficiency, working memory in L2 research is a relatively recent and under-researched topic. More importantly, whether verbal working memory is subject to an improvement and whether improving working memory positively influences L2 learning have not yet been investigated thoroughly according to Tsai, Au, and Jaeggi (2016). Nevertheless, I agree with Tsai et al.'s (2016) suggestion that working memory training will improve working memory, which in turn will positively affect general L2 learning, considering the empirically supported causal relationship between improved working memory and first language (L1) learning (e.g., Carretti, Caldarola, Tencati, & Cornoldi, 2014; Karbach, Strobach, & Schubert, 2015). Linch et al. (2014) confirmed the increasing number of investigations of the association between working memory and L2 learning; thus, for the next step, future research is required to explore the causal relationship between improved working memory and L2 learning, which will ultimately contribute to extending our knowledge of the DDL model and its cognitive components.

6.3 Limitations and suggestions

Chapter 4 is not without limitations. First, it was an observational study with recruited participants and no random assignment. Thus, the findings may apply to similar students in similar contexts; however, generalization of its findings to a wider population may not be possible. Second, due to time constraints, I could not assess the participants' phonological short-term memory. Thus, future studies should assess both working memory and phonological short-term memory to explore the comprehensive role of cognition in corpus-based L2 vocabulary learning. In addition, I wish to emphasize the strong need for

investigations on the causal relationship between improving working memory and L2 vocabulary learning. Finally, I also suggest that researchers include and assess motivational factors in DDL. Motivation is another dominant learner factor in L2 learning (see Tseng et al., 2006; Tseng & Schmitt, 2008) and other studies have suggested the need for further investigation of this factor to better understand successful L2 lexical inferencing (e.g., Hu & Nassaji, 2014) and DDL (e.g., Curado Fuentes, 2015; Chapter 3). In Chapter 4, I found that the participants were generally motivated during their DDL activities, which was likely due to the compensation they received or their personal interest in the study. However, this may not apply to other learners. Thus, taking learners' motivational factors into consideration may improve the model presented in this study, and thus further expand our understanding of DDL.

CHAPTER 5: Summary, Implications, and Conclusion

5.1 Summary

The primary goal of my dissertation was to understand and further explore the effects of corpus use on L2 vocabulary learning using new methodologies and theoretical perspectives. Table 5.1 represents the summary of key findings of my dissertation.

Table 5.1
Summary of Key Findings

Chapter	Key Finding
<Chapter 1> Literature Review and Meta-Analysis: Effects of Corpus Use on Second Language Vocabulary Learning	1. Overall medium-sized positive effect of corpus use on L2 vocabulary learning for both short-term and long-term outcomes 2. A large effect size for improving in-depth L2 vocabulary dimension 3. Learners' L2 proficiency, interaction types, corpus types, training, and duration influenced the magnitude of the effectiveness.
<Chapter 2> Effects of Concordance-based Electronic Glosses on L2 Vocabulary Learning	1. Adjusting methodological features enhances the effectiveness of corpus-based interventions. 2. On average, learners performed successful DDL and received benefits from the additional confirmation process. 3. Each target vocabulary may require different treatments for it to be recalled most efficiently and effectively.
<Chapter 3> Unearthing Hidden Groups of Learners in a Corpus-based L2 Vocabulary Learning Experiment	1. Identified two groups of learners (i.e., DDL-sufficient and DDL-insufficient learner types) overshadowed by the average at the group level 2. DDL can be beneficial to vocabulary learning for L2 learners in general, but the impact can be greater for those with higher L2 proficiency.
<Chapter 4> Role of Learner Factors in Corpus-based L2 Vocabulary Learning	1. Identified DDL-focused strategies, such as exploring, cross-checking/double-checking, and synthesizing, and their significant role in contributing to successful DDL 2. L2 proficiency and working memory directly and indirectly contributed to vocabulary acquisition and retention, indicating their roles to manage cognitive load in DDL.

In Chapter 1, I conducted a multilevel meta-analysis to investigate the effects of corpus use on L2 vocabulary learning as well as the influence of moderators on effectiveness. Based on 29 studies representing 38 unique samples, all of which met several criteria for inclusion (e.g., with control groups), I found an overall medium-sized positive effect of corpus use on L2 vocabulary learning for both short-term and long-term outcomes. Furthermore, large variation in adjusted mean effect sizes across moderators was revealed. Above all, for the different dimensions of L2 vocabulary knowledge, in-depth knowledge (i.e., referential meanings as well as syntactic features of vocabulary) was associated with a large effect size. Moreover, the results revealed that learners' L2 proficiency and several features of corpus use (i.e., interaction types, corpus types, training, and duration) influence the magnitude of the effectiveness of corpus use in improving L2 vocabulary learning.

In Chapter 2, I conducted an experiment to investigate the effects of two different vocabulary learning conditions in digital reading environments equipped with electronic textual glossing. The first condition presents the concordance lines of a target lexical item, thereby making learners infer its meaning by reading the referenced sentences. The second condition additionally offers the definition of a target lexical item after learners consult the concordance lines, thus enabling learners to confirm their meaning inference. Overall, the findings showed that the second condition resulted in higher vocabulary gains than both the first condition and the control condition. Yet, a closer look at the interactions of (a) the participants' clicking behaviors, (b) the difficulty of selected concordance lines, (c) the surrounding contexts around target lexical items, and (d) the participants' prior knowledge

of the target lexical items showed that each target lexical item may require different treatments for it to be recalled most efficiently and effectively.

In Chapter 3, I used a data mining approach to identify hidden groups in the experiment reported in Chapter 2. Although results of the previous chapter based on variable-centered analysis (i.e., multiple regression analysis) revealed that more glossary information could lead to better learning outcomes, using a model-based clustering technique in Chapter 3 allowed me to unearth learner types not identified in the previous analysis. Instead of the performance pattern found in the previous analysis (more glossary led to higher gains), I identified one learner group who exhibited their ability to make successful use of concordance lines (and thus are optimized for DDL), and another group who showed limited L2 vocabulary learning when exposed to concordance lines only. Further, results revealed that L2 proficiency intersects with vocabulary gains of different learner types in complex ways. That is, I found that DDL can be beneficial to vocabulary learning for L2 learners in general, but the impact can be greater for those with higher L2 proficiency.

In Chapter 4, I investigated how learner factors, such as L2 proficiency, strategy use, and working memory, are associated with successful corpus-based L2 vocabulary acquisition and retention using DDL. A mixed-method investigation identified participants' DDL-focused strategy use, such as exploring, cross-checking/double-checking, and synthesizing. This largely influenced learners' L2 vocabulary learning, highlighting the pedagogical advantages of these strategies for successful DDL. Results also revealed that participants' L2 proficiency and working memory directly and indirectly contributed to

their vocabulary acquisition and retention, indicating their roles to manage cognitive load in DDL.

Based on the findings across the four studies, I believe the dissertation provides researcher and educators with guidelines on how to effectively use corpus data and tools in their L2 classrooms. In addition, given that the use of technology in L2 learning, such as DDL, is exponentially increasing, the dissertation will provide a positive example of what we should consider when we adopt, use, and evaluate technologies in different language learning contexts. In the following, I described implications for teaching and research (see Table 5.2 for overview).

5.2 Implications for teaching

For pedagogical implications, my dissertation has focused on providing evidence-based suggestions on how to effectively implement DDL in L2 vocabulary learning. Above all, I found evidence from Chapter 1 to 4, supporting the importance of (1) the suitability of language data and (2) mastery of corpus consultation skills. First, for more effective corpus-based activities, educators are encouraged to check the suitability of language data prior to DDL in accordance with their students' L2 proficiency (Chapter 1). As part of this effort, they may provide their students with pre-selected, comprehensible, and/or finely tuned concordance lines, and learner-friendly concordancer software with less of a concern for DDL's accessibility (Chapter 1). In addition, educators may provide additional confirmation opportunities after DDL for learners in need to maximize their L2 vocabulary learning (Chapter 2); nevertheless, they should be mindful that each vocabulary may require different accommodations for it to be recalled most efficiently and effectively. Furthermore, educators should acknowledge that the size of individual learners' L2

vocabulary achievement using DDL can be varied depending on their L2 proficiency levels and working memory capacities, which are for managing cognitive load involved in language data from DDL (Chapters 3 & 4).

Second, educators are suggested to ensure learners' mastery of corpus consultation skills for successful DDL. For example, I believe that providing learners with pre-selected concordance lines (Chapter 2) or comprehensible, finely tuned concordance lines, and learner-friendly concordancer software (Chapter 1) did not only lower the language barrier but also contributed to making the corpus-based materials more accessible even for learners with limited mastery of corpus consultation skills. To this end, educators are encouraged to implement necessary training opportunities for successful DDL provided that DDL-focused strategies, such as exploring, cross-checking/double-checking, and synthesizing, are teachable (Chapter 4). Likewise, DDL-focused strategy use was the most dominant learner factor, demonstrating greater contributions to L2 vocabulary acquisition and retention than any other learner factors, such as L2 proficiency and working memory (Chapter 4).

Table 5.2
Implications for Teaching and Research

	Implications
For Teaching	1. *Suitability of language data* ◦ Check the suitability of language data prior to DDL in accordance with their students' L2 proficiency (Chapters 1 & 2) ◦ Be mindful that the size of individual learners' L2 vocabulary achievement using DDL can be varied depending on their L2 proficiency levels and working memory capacities (Chapters 3 & 4).
	2. *Mastery of corpus consultation skills* ◦ Implement necessary training opportunities to help learners to master DDL-focused strategies, such as exploring, cross-checking/double-checking, and synthesizing (Chapters 1& 4) ◦ Make DDL more accessible and manageable for learners with limited mastery of corpus consultation skills (Chapters 1, 2, & 3)
For Research	1. *Use of cutting-edge methodologies* ◦ A multilevel meta-analysis to capture large methodological differences within (i.e. between effect sizes within studies) and across the included studies (Chapter 1) ◦ Data mining techniques to unearth hidden groups of learners in an instructed L2 vocabulary learning context (Chapter 3)
	2. *Triangulation of findings* ◦ Run multiple statistic models to estimate causal treatment effects in a more reliable and robust way (Chapters 1, 2, & 3) ◦ Conduct a mixed-method approach or employ both variable-centered and person-centered analyses for a more comprehensive understanding of their data and findings (Chapters 2, 3, & 4)

5.3 Implications for research

For methodological implications, I would like to highlight the following two issues: (1) use of cutting-edge methodologies, and (2) triangulation of findings. First, researchers are encouraged to use cutting-edge methodologies, such as a multilevel meta-analysis and data mining techniques. For example, for their review of literature, I suggest that researchers conduct a multilevel meta-analysis (Chapter 1) instead of computing a single effect size per study, which has been a popular way of meta-analysis to avoid violating the independence assumption. By calculating multiple ESs for each study when it has multiple samples and/or measurements, researchers can fully capture large methodological differences within (i.e. between effect sizes within studies) and across the included studies.

Using a multilevel modeling approach cannot only resolve the independence issue but also ensure adequate statistical power to conduct a regression analysis (i.e., meta-regression) to statistically compare the different impacts of moderators and to compute adjusted means for each moderator, after controlling for other variables.

Along the line, researchers are encouraged to adopt data mining techniques in their research to unearth hidden groups of learners in an instructed L2 vocabulary learning context (Chapter 3). By doing so, for example, they can find any meaningful groups of learners prior to their interventions to maximize their students' L2 learning or to examine any possible interaction effects between the group membership and their interventions; therefore, it may provide valuable findings regarding personalized L2 instruction. In Chapter 3, I tried to unpack the algorithm for the analysis to be better understood by researchers seeking to apply the implications of what I have found, and I left a message to readers to refer to a guiding reference and included technical information in Appendix 3.1.

Second, I encourage researchers to employ applicable research methodologies to triangulate their findings by appropriately assessing and measuring learner data for a more accurate and comprehensive understanding of L2 research. For example, researchers may run multiple statistic models to estimate causal treatment effects in a more reliable and robust way (Chapter 2). For example, one can run a regression analysis with pre-test scores as covariate (i.e., control variable), and conduct an additional model with gain-scores (i.e., post-test scores – pre-test scores) as the dependent variable to confirm the previous inferential statistics. Further, they can run another regression model with fixed-effects adjustment to focus on within-differences for a more accurate estimate of the treatment effects.

Furthermore, researchers can conduct a mixed-method using both quantitative and qualitative methodologies to answer their research questions in a more dynamic and comprehensive way (Chapters 2 & 4). For example, researchers may conduct a randomized controlled trial and then include student interviews to understand how individual learners actually experienced the trial (Chapter 2). Also, they can simultaneously measure qualitative components (e.g., learners' strategy use) and quantitative components (e.g., L2 proficiency, working memory) and synthesize these variables to test their hypotheses (Chapter 4).

In addition, researchers can employ both variable-centered (e.g., regression analysis) and person-centered (e.g., model-based clustering) analyses for a more comprehensive understanding of their data and findings (Chapters 2 & 3). Such an implementation could either produce similar results across different analyses (and thus enhance the validity of the overall result) or present a conflicting finding that could provide useful information about hidden groups of learners with different learner types.

5.4 Conclusion

This dissertation contributes to understanding corpus use in L2 vocabulary learning and establishing a DDL model. By conducting four studies, first I found that corpus use as a learning tool was overall effective in L2 vocabulary learning (Chapter 1). Second, providing an additional confirmation process after DDL was effective on average, and yet additional qualitative analysis and data mining approach revealed that different learning patters existed beyond the average, thus requiring researchers and educators' close attention to maximize learners' L2 vocabulary learning (Chapters 2 & 3). Third, the collected data from a mixed-method approach fitted the hypothesized DDL model where learners' DDL-focused

strategy use, L2 proficiency, and working memory directly and indirectly contribute both to L2 vocabulary acquisition and retention (Chapter 4). By doing so, the findings suggest that educators consider the suitability of language data according to their teaching contexts and their students' mastery of corpus consultation skills. Also, researchers were recommended to use cutting-edge research methodologies and employ appropriate approaches to triangulate findings of their research. For future studies, first continuing meta-analytic efforts should be recommended for cumulative evidence for the effectiveness of corpus use. Second, more detailed learner data (e.g., learning analytics) should be collected and analyzed to extend our understanding of learners' DDL activities. Third, replication attempts with large and diverse samples should be encouraged to enrich DDL research. Last but not least, future efforts to expand and improve the DDL model for L2 vocabulary learning examined in this dissertation by taking other important learner factors, such as motivation, into consideration should be followed.

REFERENCES

References marked with an asterisk indicate studies included in meta-analysis (Introduction)

Abraham, L. B. (2008). Computer-mediated glosses in second language reading comprehension and vocabulary learning: A meta-analysis. *Computer Assisted Language Learning, 21*(3), 199–226. https://doi.org/10.1080/09588220802090246

AbuSeileek, A. F. (2011). Hypermedia annotation presentation: The effect of location and type on the EFL learners' achievement in reading comprehension and vocabulary acquisition. *Computers & Education, 57*(1), 1281–1291. https://doi.org/10.1016/j.compedu.2011.01.011

Acock, A. (2013). *Discovering structural equation modeling using Stata*: Revised Edition. New York, NY: Stata Press.

Adlof, S., Frishkoff, G., Dandy, J., & Perfetti, C. (2016). Effects of induced orthographic and semantic knowledge on subsequent learning: A test of the partial knowledge hypothesis. *Reading and Writing, 29*(3), 475–500. https://doi.org/10.1007/s11145-015-9612-x

Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal, 63*(1), 23–32. https://doi.org/10.1093/elt/ccn011

*Al-Mahbashi, A., Noor, N. M., & Amir, Z. (2015). The effect of data driven learning on receptive vocabulary knowledge of Yemeni university learners. *3L: Language, Linguistics, Literature®, 21*(3), 13–24. http://ejournals.ukm.my/3l/article/view/9511/3418

*Anani Sarab, M. R., & Kardoust, A. (2014). Concordance-based Data-Driven Learning activities and learning English phrasal verbs in EFL classrooms. *Issues in Language Teaching, 3*(1), 89–112. http://ilt.atu.ac.ir/mobile/article_1370.html

Anderson-Inman, L., & Horney, M. A. (2007). Supported eText: Assistive technology through text transformations. *Reading Research Quarterly, 42*(1), 153–160. https://doi.org/10.1598/RRQ.42.1.8

Angeles, G., & Mroz, T. A. (2001). *A simple guide to using multilevel models for the evaluation of program impacts*. Washington, DC: United States Agency for International Development.

Anvari, S., & Farvardin, M. T. (2016). Revisiting lexical inferencing strategies in L2 reading: A comparison of successful and less successful EFL inferencers. *The Reading Matrix: An International Online Journal, 16*(1), 63–77. http://www.readingmatrix.com/archive/16/1

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36*, 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194

Bergman, L., & Magnusson, D. (1997). A person-oriented approach in research on developmental psychopathology. *Development and Psychopathology, 9*(2), 291–319. https://doi.org/10.1017/S095457949700206X

Bernardini, S. (2004). Corpora in the classroom. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15–36). Amsterdam, The Netherlands: John Benjamins.

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical, research, and classroom perspectives*. Norwood, NJ: Ablex.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, UK: Cambridge University Press.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons Ltd.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67(*2), 348–393. https://doi.org/10.1111/lang.12224

Boulton, A. (2009a). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics, 35*(1), 81–106.

Boulton, A. (2009b). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL, 21*(1), 37–54. https://doi.org/10.1017/S0958344009000068

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning, 60*(3), 534–572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

Cameron, A. C., & Trivedi, P. K. (1990). *The information matrix test and its implied alternative hypotheses* (Working Paper 372). Davis, CA: Institute of Governmental Affairs.

Carretti, B., Caldarola, N., Tencati, C., & Cornoldi, C. (2014). Improving reading comprehension in reading and listening settings: The effect of two training programmes focusing on metacognition and working memory. *British Journal of Educational Psychology, 84*(2), 194–210. https://doi.org/10.1111/bjep.12022

*Çelik, S. (2011). Developing collocational competence through web-based concordance activities. *Novitas-ROYAL (Research on Youth and Language), 5*(2), 273–286. http://www.novitasroyal.org/Vol_5_2/CelikS.pdf

Chan, T. P., & Liou, H. C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations. *Computer Assisted Language Learning, 18*(3), 231–251. https://doi.org/10.1080/09588220500185769

Chapelle, C. A. (2003). *English language learning and technology*. Amsterdam, The Netherlands: John Benjamins.

Chen, I-J., & Yen, J-C. (2013). Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary learning in foreign languages. *Computers & Education, 63*, 416–423. https://doi.org/10.1016/j.compedu.2013.01.005

Cheng, W. (2012). *Exploring corpus linguistics: Language in action*. New York: NY: Routledge.

Chujo, K., Oghigian, K., & Akasegawa, S. (2015). A corpus and grammatical browsing system for remedial EFL learners. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 109–128). Amsterdam, The Netherlands: John Benjamins.

Chun, D. M. (2001). L2 reading on the Web: Strategies for accessing information in hypermedia. *Computer Assisted Language Learning, 14*(5), 367–403. https://doi.org/10.1076/call.14.5.367.5775

Chun, D. M. (2011). CALL technologies for L2 reading post Web 2.0. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: Theory and research to new directions in foreign language teaching* (pp. 131–170). San Marcos, TX: CALICO.

*Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System, 25*(3), 301–315. https://doi.org/10.1016/S0346-251X(97)00024-9

*Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research and Development, 47*(3), 15–31. https://doi.org/10.1007/BF02299631

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 478–497). Cambridge, UK: Cambridge University Press.

Cobb, T., Greaves, C., & Horst, M. (2001). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la didactique des langues seconds* (pp. 133–153). Montréal, Canada: Éditions logique.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Lawrence Earlbaum Associates.

Collentine, J. (2000). Insights into the construction of grammatical knowledge provided by user-behavior tracking technologies. *Language Learning & Technology, 3*(2), 44–57. https://doi.org/10125/25072

Csizer, K., & Dörnyei, Z. (2005). Language learners' motivational profiles and their motivated learning behavior. *Language Learning, 55*(4), 613–659. https://doi.org/10.1111/j.0023-8333.2005.00319.x

Cunningham, S., Moor, P., & Carr, J. C. (2003). *Cutting edge advanced with phrase builder*. Harlow, UK: Pearson Education.

Curado Fuentes, A. (2015). Exploiting keywords in a DDL approach to the comprehension of news texts by lower-level students. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 177–197). Amsterdam, The Netherlands: John Benjamins.

*Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning, 28*(2), 130–144. https://doi.org/10.1080/09588221.2013.803982

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present.* https://corpus.byu.edu/coca

De Bot, K., Paribakht, T., & Wesche, M. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition: Evidence from ESL Reading. *Studies in Second Language Acquisition, 19*(3), 309–329. https://doi.org/10.1017/S0272263197003021

Deschambault, R. (2017). Actively managed products: Think-aloud data and methods in applied linguistics research. *Applied Linguistics Review*. Advance online publication. https://doi.org/10.1515/applirev-2017-0028

Dolnicar, S. (2002, December). *A review of unquestioned standards in using cluster analysis for data-driven market segmentation*. Paper presented at the Australian and New Zealand Marketing Academy Conference, Victoria, Australia.

Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics, 70*(s1): 927–939. https://doi.org/10.1111/j.1468-0084.2008.00537.x

Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Educational Testing Service. (2016). *TOEIC® listening and reading test scored and the CEFR levels*. https://www.etsglobal.org/Tests-Preparation/The-TOEIC-Tests/TOEIC-Listening-Reading-Test/Scores-Overview

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. Cambridge, MA: MIT press.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. https://doi.org/10.3758/BRM.41.4.1149

Field, A. P. (2009). *Discovering statistics using SPSS* (3rd edition). London, UK: Sage.

Firooz, H. (2015, March 4). *When not to use Gaussian Mixture Model (EM clustering)*. [Web log post]. https://hameddaily.blogspot.com/2015/03/when-not-to-use-gaussian-mixtures-model.html

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis*. Hoboken, NJ: John Wiley & Sons Ltd.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System, 21*(2), 231–244. https://doi.org/10.1016/0346-251X(93)90044-H

Flowerdew, L. (2008, July). *Pedagogic value of corpora: A critical evaluation*. Paper presented at the 8th Teaching and Language Corpora conference, Lisbon, Portugal.

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). Amsterdam, The Netherlands: John Benjamins.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal, 41*(8), 578–588. https://doi.org/10.1093/comjnl/41.8.578

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611–631. https://doi.org/10.1198/016214502760047131

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., & Fop, M. (2017). *mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation* (R package version 5.3). https://CRAN.R-project.org/package=mclust

*Frankenberg-Garcia, A. (2012). Learners' use of corpus examples. *International Journal of Lexicography, 25*(3), 273–296. https://doi.org/10.1093/ijl/ecs011

*Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL, 26*(2), 128–146. https://doi.org/10.1017/S0958344014000093

Fraser, C. A. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition, 21*, 225–241. https://doi.org/10.1017/S0272263199002041

Freebody, P., & Anderson, R. C. (1983). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Literacy Research, 15*(3), 19–39. https://doi.org/10.1080/10862968309547487

*Gan, S. L., Low, F., & bte Yaakub, N. F. (1996). Modeling teaching with a computer-based concordancer in a TESL preservice teacher education program. *Journal of Computing in Teacher Education, 12*(4), 28–32. https://doi.org/10.1080/10402454.1996.10784301

Gass, S. M., Behney, J., & Plonsky, L. (2013). *Second language acquisition: An introductory course* (4th ed.). New York, NY: Routledge/Taylor & Francis.

Gathercole, S. E., Alloway, T. P., Willis, C. S., & Adams, A. M. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology, 93*, 265–281. https://doi.org/10.1016/j.jecp.2005.08.003

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*(5), 887–898. https://doi.org/10.1037/0012-1649.28.5.887

Gavioli, L. (2005). *Exploring corpora for ESP learning*. Amsterdam, The Netherlands: John Benjamins.

Gavioli, L. (2009). Corpus analysis and the achievement of learner autonomy in interaction. In L. Lombardo (Ed.), *Using corpora to learn about language and discourse* (pp. 39–71). Bern, Switzerland: Peter Lang.

Godwin-Jones, R. (2001a). Language testing tools and technologies. *Language Learning & Technology, 5*(2), 8–13. https://doi.org/10125/25121

Godwin-Jones, R. (2001b). Tools and trends in corpora use for teaching and learning. *Language Learning & Technology, 5*(3), 7–12. https://doi.org/10125/44559

*Gordani, Y. (2013). The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners' vocabulary development. *Computer Assisted Language Learning, 26*(5), 430–445. https://doi.org/10.1080/09588221.2012.685078

Graves, M. F. (2006). *The vocabulary book*. New York, NY: Teachers College Press.

Guo, Y. (2008). *The role of vocabulary knowledge, syntactic awareness and metacognitive awareness in reading comprehension of adult English language learners*. (Unpublished doctoral dissertation). Tallahassee, FL: Florida State University.

Gyllstad, H. (2013). Looking at L2 vocabulary knowledge dimensions from an assessment perspective—Challenges and potential solutions. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 11–28). Amsterdam, The Netherlands: Eurosla. http://www.eurosla.org/

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press. https://doi.org/10.1016/C2009-0-03396-0

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition, 21*(2), 303–317. https://doi.org/10.1017/s0272263199002089

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics – Theory and Methods, 19*(10), 3595–3617. https://doi.org/10.1080/03610929008830400

Hermagustiana, I. (2018). Exploring English lexical inferencing strategies performed by EFL university students. In S. Madya, F. A. Hamied, W. A. Renandya, C. Coombe, & Y. Basthomi (Eds.), *ELT in Asia in the digital era: Global citizenship and identity* (pp. 73–80). New York, NY: Routledge.

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocation: Further development in the lexical approach* (pp. 47–69). Oxford, UK: Oxford University Press.

*Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning & Technology, 9*(2), 90–110. https://doi.org/10125/44021

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Hu, H. C. M., & Nassaji, H. (2012). Ease of inferencing, learner inferential strategies, and their relationship with the retention of word meanings inferred from context. *Canadian Modern Language Review, 68*(1), 54–77. https://doi.org/10.3138/cmlr.68.1.054

Hu, H. C. M., & Nassaji, H. (2014). Lexical inferencing strategies: The case of successful versus less successful inferencers. *System, 45*, 27–38. https://doi.org/10.1016/j.system.2014.04.004

Huang, L. -S. (2011). Language learners as language researchers: The acquisition of English grammar through a corpus-aided discovery learning approach mediated by intra and interpersonal dialogues. In J. Newman, S. Rice, & H. Baayen (Eds.), *Corpus-based studies in language documentation, use, and learning* (pp. 91–112). Amsterdam, The Netherlands: Rodopi.

Huber, C. (2013, September 5). *Measures of effect size in Stata 13* [Web log post]. http://blog.stata.com/tag/hedgess-g/

Hummel, K. M., & French, L. M. (2016). Phonological memory and aptitude components: Contributions to second language proficiency. *Learning and Individual Differences, 51*, 249–255. https://doi.org/10.1016/j.lindif.2016.08.016

Johns, T. (1986). Micro-concord: A language learner's research tool. *System, 14*(2), 151–162. https://doi.org/10.1016/0346-251X(86)90004-7

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. In T. Johns & P. King (Eds.), *Classroom concordancing* (pp. 1–16). Birmingham, UK: English Language Research Journal.

Johns, T. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 293–313). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139524605.014

Johns, T. (1997). Contexts: The background, development and trialing of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). Harlow, UK: Addison Wesley Longman.

Jung, Y. G., Kang, M. S., & Heo, J. (2014). Clustering performance comparison using K-means and expectation maximization algorithms. *Biotechnology & Biotechnological Equipment, 28*(sup1): S44–S48. https://doi.org/10.1080/13102818.2014.949045

Karbach, J., Strobach, T., & Schubert, T. (2015). Adaptive working-memory training benefits reading, but not mathematics in middle childhood. *Child Neuropsychology, 21*(3), 285–301. https://doi.org/10.1080/09297049.2014.899336

*Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL, 28*(2), 166–186. https://doi.org/10.1017/S0958344015000154

Kaufman, D. (2004). Constructivist issues in language learning and teaching. *Annual Review of Applied Linguistics, 24*, 303–319. https://doi.org/10.1017/S0267190504000121

*Kaur, J., & Hegelheimer, V. (2005). ESL students' use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning, 18*(4), 287–310. https://doi.org/10.1080/09588220500280412

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, O., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

*Kim, E. (2015). Enhancing a corpus-based approach to teach English phrasal verbs to Korean learners. *The Journal of Studies in Language, 31*(2), 295–312.

Kim, Y. S. G. (2017). Multicomponent view of vocabulary acquisition: An investigation with primary grade children. *Journal of Experimental Child Psychology, 162*, 120–133. https://doi.org/10.1016/j.jecp.2017.05.004

Kita, K., & Ogata, H. (1997). Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning, 10*(3), 229–238. https://doi.org/10.1080/0958822970100303

Kline, R. B. (2012). Assumptions in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 111–125). New York, NY: Guilford Press.

*Koosha, M., & Jafarpour, A. A. (2006). Data-driven learning and teaching collocation of prepositions: The case of Iranian EFL adult learners. *Asian EFL journal, 8*(4), 192–209. https://www.asian-efl-journal.com/journal-2006/

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and cognition, 11*(2), 261–271. https://doi.org/10.1017/S1366728908003416

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. New York, NY: Longman.

Lai, C., & Zhao, Y. (2005). Introduction: The importance of input and the potential of technology for enhancing input. In Y. Zhao (Ed.), *Research in technology and second language learning: Developments and directions* (pp. 95–101). Charlotte, NC: Information Age.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75*(4), 440–448. https://doi.org/10.1111/j.1540-4781.1991.tb05380.x

Laufer, B. (1993). The effect of dictionary definitions and examples on the use comprehension of new L2 words. *Cahiers de Lexicologie, 63*(2), 131–142. http://www.classiques-garnier.com/editions-tabmats/LexMS63_tabmat.pdf

Laufer, B. (2000). Electronic dictionaries and incidental vocabulary acquisition: Does technology make a difference? In U. Heid, S. Evert, E. Lehmann, & C. Rohrer (Eds.), *Proceedings of the ninth EURALEX international congress* (pp. 849–854). Stuttgart, Germany: Stuttgart University.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*, 1–26. https://doi.org/10.1093/applin/22.1.1

Laufer, B., & Paribakht, T. S. (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning, 48*(3), 365–391. https://doi.org/10.1111/0023-8333.00046

Laufer, B., & Yano, Y. (2001). Understanding unfamiliar words in a text: Do L2 learners understand how much they don't understand? *Reading in a Foreign Language, 13*(2), 549–566. http://nflrc.hawaii.edu/rfl/PastIssues/rfl132laufer.pdf

Lawrence, J. F., Crosson, A. C., Paré-Blagoev, E. J., & Snow, C. E. (2015). Word generation randomized trial: Discussion mediates the impact of program treatment on academic word learning. *American Educational Research Journal, 52*(4), 750–786. https://doi.org/10.3102/0002831215579485la

Lee, H., & Lee, J. H. (2013). Implementing glossing in mobile-assisted language learning environments: Directions and outlook. *Language Learning & Technology, 17*(3), 6–22. https://doi.org/10125/44334

Lee, H., & Lee, J. H. (2015). The effects of electronic glossing types on foreign language vocabulary learning: Different types of format and glossary information. *The Asia-Pacific Education Researcher, 24*(4), 591–601. https://doi.org/10.1007/s40299-014-0204-3

Lee, H., Warschauer, M., & Lee, J. H. (2017). The effects of concordance-based electronic glosses on L2 vocabulary learning. *Language Learning & Technology, 21*(2), 32–51. https://doi.org/10125/44610

Lee, H., Warschauer, M., & Lee, J. H. (2018a). Advancing CALL research via data mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344018000162

Lee, H., Warschauer, M., & Lee, J. H. (2018b). The effects of corpus use on second language vocabulary learning: A multilevel meta-analysis. *Applied Linguistics*. Advance online publication. https://doi.org/10.1093/applin/amy012

Lee, J. H., Lee, H., & Sert, C. (2015). A corpus approach for autonomous teachers and learners: Implementing an on-line concordancer on teachers' laptops. *Language Learning & Technology, 19*(2), 1–15. https://doi.org/10125/44411

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, A. M. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 11–23). London, UK: Longman.

Leńko-Szymańska, A., & Boulton, A. (2015). Introduction: Data-driven learning in language pedagogy. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 1–14). Amsterdam, The Netherlands: John Benjamins.

Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2013). New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (pp. 1150–1181). Newark, Delaware: International Reading Association.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove, UK: Language Teaching Publications.

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*, 861–883. https://doi.org/10.3758/s13423-013-0565-2

Lomicka, L. L. (1998). 'To gloss or not to gloss': An investigation of reading comprehension online. *Language Learning & Technology, 1*(2), 41–50. https://doi.org/10125/25020

Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102*(3), 701–711. https://doi.org/10.1037/a0019135

Maris, E. (1998). Covariance adjustment versus gain scores – revisited. *Psychological Methods, 3*(3), 309–327. https://doi.org/10.1037/1082-989X.3.3.309

Martin, K. I., & Ellis, N. C. (2012). The role of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition, 34*, 379–413. https://doi.org/10.1017/S0272263112000125

Meila, M., & Heckerman, D. (2001). An experimental comparison of model-based clustering methods. *Machine Learning, 42*(1/2), 9–29. https://doi.org/10.1023/A:1007648401407

Min, H. T. (2008). EFL vocabulary acquisition and retention: Reading plus vocabulary enhancement activities and narrow reading. *Language Learning, 58*(1), 73–115. https://doi.org/10.1111/j.1467-9922.2007.00435.x

*Mirzaei, A., Domakani, M. R., & Rahimi, S. (2016). Computerized lexis-based instruction in EFL classrooms: Using multi-purpose LexisBOARD to teach L2 vocabulary. *ReCALL, 28*(1), 22–43. https://doi.org/10.1017/S0958344015000129

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies, 22*, 1–18. http://mizumot.com/files/ecs2015.pdf

Mondria, J. A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings. *Studies in Second Language Acquisition, 25*(4), 473–499. https://doi.org/10.1017/S0272263103000202

Morris, S. B. (2008). Estimating effect sizes from the pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Mun, E. Y., von Eye, A., Bates, M. E., & Vaschillo, E. G. (2008). Finding groups using model-based cluster analysis: Heterogeneous emotional self-regulatory processes and

heavy alcohol use risk. *Developmental Psychology, 44*(2), 481–495. https://doi.org/10.1037/0012-1649.44.2.481

Nassaji, H. (2003). L2 vocabulary learning from context: Strategies, knowledge sources, and their relationship with success in L2 lexical inferencing. *TESOL Quarterly, 37*(4), 645–670. https://doi.org/10.2307/3588216

Nation, I. S. P. (2001). *Learning vocabulary in another language*. New York, NY: Cambridge University Press.

Nation, I. S. P. (2009). New roles for FL vocabulary? In L. Wei & V. Cook (Eds.), *Contemporary applied linguistics volume 1: Language teaching and learning* (pp. 99–116). London, UK: Continuum.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13. https://www.victoria.ac.nz/lals/about/staff/paul-nation

Papi, M., & Teimouri, Y. (2014). Language learner motivational types: A cluster analysis study. *Language Learning, 64*(3), 493–525. https://doi.org/10.1111/lang.12065

Pires, A. M., & Branco, J. A. (2010). Projection-pursuit approach to robust linear discriminant analysis. *Journal of Multivariate Analysis, 101*(10), 2464–2485. https://doi.org/10.1016/j.jmva.2010.06.017

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second language multimedia learning environment. *Journal of Educational Psychology, 90*(1), 25–36. https://doi.org/10.1037/0022-0663.90.1.25

Plonsky, L., & Oswald, F. L. (2012). How to do a meta-analysis. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 275–295). London, UK: Basil Blackwell.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Poole, R. (2012). Concordance-based glosses for academic vocabulary acquisition. *CALICO Journal, 29*(4), 679–693. https://doi.org/10.11139/cj.29.4.679-693

Pratt, T. C., Turanovic, J. J., & Cullen, F. T. (2016). Revisiting the criminological consequences of exposure to fetal testosterone: A meta-analysis of the 2D: 4D digit ratio. *Criminology, 54*(4), 587–620. https://doi.org/10.1111/1745-9125.12115

*Rahimi, M., & Momeni, G. (2012). The effect of teaching collocations on English language proficiency. *Procedia-Social and Behavioral Sciences, 31*, 37–42. https://doi.org/10.1016/j.sbspro.2011.12.013

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge, UK: Cambridge University Press.

*Rezaee, A. A., Marefat, H., & Saeedakhtar, A. (2015). Symmetrical and asymmetrical scaffolding of L2 collocations in the context of concordancing. *Computer Assisted Language Learning, 28*(6), 532–549. https://doi.org/10.1080/09588221.2014.889712

Royston, J. P. (1991). sg3.5: Comment on sg3.4 and an improved D'Agostino test. *Stata Technical Bulletin, 3*, 23–24. https://stata-press.com/journals/stbcontents/stb3.pdf

Rüschoff, B., & Ritter, M. (2001). Technology-enhanced language learning: Construction of knowledge and template-based learning in the foreign language classroom. *Computer Assisted Language Learning, 14*(3-4), 219–232. https://doi.org/10.1076/call.14.3.219.5789

Saldaña, J. (2016). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.

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

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

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal, 8*(1), 289–317. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental & quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (Complete samples), *Biometrika, 52*(3-4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shen, M. (2018). The role of text type and strategy use in L2 lexical inferencing. *International Review of Applied Linguistics in Language Teaching, 56*(2), 231–252. https://doi.org/10.1515/iral-2015-0054

Shin, Y., & Kim, Y. (2012). Assessing the relative roles of vocabulary and syntactic knowledge in reading comprehension. *Korean Journal of Applied Linguistics, 28*(2), 169–198. https://doi.org/10.17154/kjal.2012.06.28.2.169

Sideridis, G., Simos, P., Papanicolaou, A., & Fletcher, J. (2014). Using structural equation modeling to assess functional connectivity in the brain power and sample size considerations. *Educational and Psychological Measurement, 74*(5), 733–758. https://doi.org/10.1177/0013164414525397

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Sinclair, J. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 27–39). London, UK: Longman.

Sinclair, J. (2003). *Reading concordances: An introduction*. London, UK: Pearson Longman.

Sinclair, J. (Ed.). (2004). *How to use corpora in language teaching* (vol. 12). Amsterdam, Netherlands: John Benjamins Publishing.

Skehan, P. (1986). Cluster analysis and the identification of learner types. In V. Cook (Ed.), *Experimental approaches to second language acquisition* (pp. 81–94). Oxford, UK: Pergamon.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

*Sripicharn, P. (2003). Evaluating classroom concordancing: the use of concordance-based materials by a group of Thai students. *Thammasat Review, 8*(1), 203-236. https://www.tci-thaijo.org/index.php/tureview/article/view/40909

Stahl, S., & Nagy, W. (2006). *Teaching word meanings*. Mahwah, NJ: Lawrence Erlbaum Associates.

Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 243–274). New York, NY: Routledge.

*Stevens, V. (1991). Concordance-based vocabulary exercises: A viable alternative to gap-fillers. *Classroom Concordancing: English Language Research Journal, 4*, 47–61.

Summers, D. (1988). The role of dictionaries in language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 111–125). New York, NY: Routledge.

Sun, Y. C. (2003). Learning process, strategies and web-based concordancers: A case study. *British Journal of Educational Technology, 34*(5), 601–613. https://doi.org/10.1046/j.0007-1013.2003.00353.x

*Sun, Y. C., & Wang, L. Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning, 16*(1), 83–94. https://doi.org/10.1076/call.16.1.83.15528

*Supatranont, P. (2005). *A comparison of the effects of the concordance-based and the conventional teaching methods on engineering students' English vocabulary learning* (Unpublished doctoral dissertation). Chulalongkorn University, Bangkok, Thailand.

Tacq, J. (2010). Multivariate normal distribution. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (pp. 332–338). Oxford, UK: Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.01351-8

*Tongpoon, A. (2009). *The enhancement of EFL learners' receptive and productive vocabulary knowledge through concordance-based methods* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, AZ.

Tsai, N., Au, J., & Jaeggi, S. M. (2016). Working memory, language processing, and implications of malleability for second language acquisition. In G. Granera, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 69–88). Amsterdam, Netherlands: John Benjamins Publishing.

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

Tseng, W. T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics, 27*(1), 78–102. https://doi.org/10.1093/applin/ami046

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods, 45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710–718. https://doi.org/10.1093/aje/kwk052

*Vyatkina, N. (2016). Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL, 28*(2), 207–226. https://doi.org/10.1017/S0958344015000269

Weekes, B. (2018). Learning written word vocabulary in a second language: Theoretical and practical implications. *Bilingualism: Language and Cognition, 21*(3), 585–597. https://doi.org/10.1017/S1366728917000141

Williams, J. N. (2012). Working memory and SLA. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 427–441). London, UK: Routledge.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques.* Cambridge, UK: Morgan Kaufmann.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models an evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913–934. https://doi.org/10.1177/0013164413495237F

Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics, 27*(4), 741–747. https://doi.org/10.1093/applin/aml036

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52–69. https://doi.org/10.20982/tqmp.08.1.p052

Wong, W. (2005). *Input enhancement: From theory and research to the classroom*. Boston, MA: McGraw-Hill.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT press.

Yamamori, K., Isoda, T., Hiromori, T., & Oxford, R. L. (2003). Using cluster analysis to uncover L2 learner differences in strategy use, will to learn, and achievement over time. *International Review of Applied Linguistics, 41*(4), 381–410. https://doi.org/10.1515/iral.2003.017

*Yang, J. (2015). Effects of collaborative corpus-based learning on the acquisition and retention of delexical verb collocation. *The Journal of Studies in Language, 31*(1), 67–94.

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology, 13*(2), 48–67. https://doi.org/10125/44180

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., & Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics, 17*(10), 977–987. https://doi.org/10.1093/bioinformatics/17.10.977

*Yılmaz, E., & Soruç, A. (2014). The use of concordance for teaching vocabulary: A data-driven learning approach. *Journal of Psycholinguistic Research, 45*(2), 275–285. https://doi.org/10.1007/s10936-014-9344-0

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10*(3), 85–101. https://doi.org/10125/44076

*Yunus, K., & Awab, S. A. (2012). The effects of the use of module-based concordance materials and data-driven learning (DDL) approach in enhancing the knowledge of collocations of prepositions among Malaysian undergraduate law students. *International Journal of Learning, 18*(9), 165–181.

*Yunxia, S., Min, Y., & Zhuo, S. (2009, August). *An empirical study on a computer-based corpus approach to English vocabulary teaching and learning*. Paper presented at the 2nd International Conference on Computer Science and Information Technology, Beijing, China. https://doi.org/10.1109/ICCSIT.2009.5234423

APPENDIX 1.1: Effect Size Calculation

For the effect size calculation, I used the equations ([1] through [5]) described below. Post-test effect sizes were computed from post-test scores of treatment and control groups (i.e., post-test effect sizes), while additional follow-up effect sizes were computed only when follow-up test (i.e., delayed post-test) results were available in a selected study.

First, each effect size was calculated according to unbiased *d* or Hedges' g (see equations [1] through [3]; Hedges & Olkin, 1985), which provides more conservative calculations than Cohen's *d* does, particularly for small samples ($n < 50$; Hedges & Olkin, 1985; Huber, 2013). It is calculated by multiplying the so-called bias correction factor (*J*; equation [3]) by Cohen's *d*, as shown below:

$$[1]\ ES_n = J_{Correction\ Factor} \times \frac{Mean_{treated} - Mean_{control}}{\sqrt{\left[\frac{(n_T - 1)SD^2_{treated} + (n_C - 1)SD^2_{control}}{n_{treated} + n_{control} - 2}\right]}}$$

$$[2]\ SE_n = J_{Correction\ Factor} \times \sqrt{\frac{1}{n_{treated}} + \frac{1}{n_{control}} + \frac{\text{Cohen's } d^2}{2 \times (n_{treated} + n_{control})}}$$

$$[3]\ J_{Correction\ Factor} = 1 - \frac{3}{\{4 \times (n_{treated} + n_{control} - 2) - 1\}}$$

It should be noted here that the bias correction factor (J) is always smaller than one. It was suggested by Hedges (1981) based on his gamma function calculation to eliminate the upward bias in Cohen's d estimates (also see Hedges & Olkin, 1985).

After calculating multiple effect sizes for a study, I estimated a single average effect size of a unique sample by combining the computed effect sizes. Since each effect size has its own standard error which indicates how precise the estimate is, instead of using a normal arithmetic approach (i.e., the sum of the effect sizes for a unique sample divided by

the number of the effect sizes; a simple mean), I opted to compute a weighted mean with more weight on effect sizes with higher precision (see Borenstein, Hedges, & Rothstein, 2007). In particular, since multiple effect sizes in a unique sample came from the same sample, I assumed that there was only random variation within each effect size (i.e., measurement error variance) when combining these effect sizes. To this end, I used the inverse-variance weighting to assign the weight to each effect size; thus, the weighted average effect size per study ($ES_{unique\ sample}$) is

$$[4]\ ES_{unique\ sample} = \frac{\sum(w_n \times ES_{n(level-1)})}{\sum w_n}$$

$$[5]\ SE_{unique\ sample} = \sqrt{\frac{1}{\sum w_n}}$$

, where ES_n is the value of effect size n (i.e., ES for effect sizes of a unique sample), w_n is the inverse-variance weight ($1/SE^2$) for effect size n, and n is equal to the number of effect sizes for a unique sample.

After the calculation of the average effect size for unique samples, I computed the overall average effect size across the collected studies for the first research question. Given that different studies did not come from the same population, I assumed that there was random variation both within each study (i.e., measurement error variance) and between studies (i.e., sampling error variance) when combining the overall average effect size at level 2. By doing so, the weighted mean effect size across all the unique samples is

$$[6]\ ES_{overall\ average} = \frac{\sum(w_k^* \times ES_{k(level-2)})}{\sum w_k^*}$$

, where ES_k is the value of effect size k (i.e., ES for unique samples), w^*_k is the inverse-variance weight for effect size k, and k is equal to the number of unique samples included in

the meta-analysis. Given that the weights are computed from the total variance of each unique sample level (i.e., weight = 1 / total variance; total variance = within-unique sample variance + between-unique sample variance), the within-study variance was computed previously (i.e., $SE_{unique\ sample}^2$), and the between-study variance (τ^2) requires a series of calculations,

$$[7]\ \tau^2_{between-unique\ sample\ variance} = \frac{\sum w_k \times (ES_{k(level-2)})^2 - \frac{(\sum w_k \times ES_{k(level-2)})^2}{\sum w_k} - (k-1)}{\sum w_k - \frac{\sum w_k^2}{\sum w_k}}$$

, where w_k is the inverse of within-study variance for effect size k (see Borenstein et al., 2007 for more information about the calculations).

References

Borenstein M, Hedges, L., & Rothstein, H. (2007). *Meta-analysis fixed effect vs. random effects* (Unpublished manuscript). https://www.meta-analysis.com

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. https://doi.org/10.3102/10769986006002107

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press. https://doi.org/10.1016/C2009-0-03396-0

Huber, C. (2013, September 5). *Measures of effect size in Stata 13* [Web log post]. http://blog.stata.com/tag/hedgess-g/

APPENDIX 1.2: Gain-Score Effect Size Calculation

One of the common meta-analysis issues that Oswald and Plsonky (2010) point out is the need to estimate gain-score effect sizes by taking into account pre-test variances in the effect size calculation, particularly for primary studies with pre-test-post-test designs. Although this idea is valid, it has not frequently been discussed or documented. Here, I describe three reliable approaches to resolve this issue. I check these approaches step by step, not only to identify the most precise method but also to find the best option that fits the data. First, the CMA program (Borenstein, Hedges, Higgins, & Rothstein, 2009) introduced equations for computing effect sizes from studies that used pre-test and post-test scores, which are as follows:

$$[1]\ \text{Hedges}'g = J_{Correction\ Factor} \times Cohen's\ d$$

$$[2]\ SE_g = J_{Correction\ Factor} \times SE_d$$

$$[3]\ J_{Correction\ Factor} = 1 - \frac{3}{\{4 \times (n_{treated} + n_{control} - 2) - 1\}}$$

$$[4]\ Cohen's\ d = \frac{MeanChange_{Treated} - MeanChange_{Control}}{\sqrt{[\frac{(n_T - 1)SD^2_{T,Change} + (n_C - 1)SD^2_{C,Change}}{n_T + n_c - 2}]}}$$

$$[5]\ SE_d = \sqrt{\frac{1}{n_{treated}} + \frac{1}{n_{control}} + \frac{Cohen's\ d^2}{2 \times (n_{treated} + n_{control})}}$$

$$[6]\ SD_{Gain} = \sqrt{SD^2_{Pre} + SD^2_{Post} - 2 \times Correlation_{Pre\&Post} \times SD_{Pre} \times SD_{Post}}$$

Given that these equations take into account statistics from pre-test as well as post-test scores, this approach is believed to provide the most precise and robust results. However, this approach has not gained much attention because most studies do not report a correlation coefficient between pre-test and post-test scores (Plonsky & Oswald, 2012).

As an alternative, Hofmann et al. (2010) used this approach by adopting a conservative correlation estimate of .70, following Rosenthal's (1993) suggestion.

Second, as a response to the absence of a standard deviation of gain score (i.e., equation [6]), researchers, (e.g., Morris, 2008), suggest using the standard deviation at baseline (i.e., pre-test scores) instead. The rationale behind this argument is that this statistic is not influenced by treatments; therefore, it is considered consistent across different studies. However, it was found that some of the selected primary studies could not satisfy this assumption. In particular, studies that measured participants' prior knowledge of specific target vocabulary before the interventions tended to have pre-test variances of nearly zero, making the results of this approach biased.

Third, in line with the idea that standard deviation for gain score differences should not be affected by treatment, I considered using the standard deviation from the post-test control group instead of that from the gain score (i.e., equation [6]). The logic behind this approach is based on the assumption that when a study has zero or nearly-zero pre-test scores at baseline, it is highly likely to have baseline standard deviations that are different from the population standard deviation.

Overall, I found that the third approach was the most plausible option for my meta-analysis. First, using the correlation coefficient of .7 between pre-test and post-test scores may not reflect the various contexts of empirical studies in the field of language learning. Most primary studies used assessments that were developed by researchers themselves and customized in numerous ways. Second, the second approach was not applicable because of the aforementioned limitations. Besides a few primary studies that measured participants' absence of prior knowledge of the target vocabulary, the field of applied

linguistics is filled with empirical studies that have similar research questions and objectives. Taken together, the first and second approaches do not fulfill my original intent of reporting replicable and robust steps for meta-analysis. The results are displayed in a bar graph, from which one can see that the newly generated gain-score effect sizes were slightly larger than the post-test effect sizes, though the difference was not statistically significant.

References

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons Ltd.

Hofmann, S. G., Sawyer, A. T., Witt, A. A., & Oh, D. (2010). The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review. *Journal of Consulting and Clinical Psychology, 78*(2), 169–183. https://doi.org/10.1037/a0018555

Morris, S. B. (2008). Estimating effect sizes from the pre-test-post-test-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85–110. https://doi.org/10.1017/S0267190510000115

Plonsky, L., & Oswald, F. L. (2012). How to do a meta-analysis. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 275–295). London, UK: Basil Blackwell.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.

APPENDIX 1.3: Publication Bias

To check any possible publication bias in the effect size calculation, I first examined publication bias using a funnel plot, which is standard errors (y-axis) against effect sizes (x-axis) and includes an inverted cone with the overall average effect size across all the unique samples as the center. The left graph in Figure A.1 shows the funnel plot; each dot in the figure represents the computed post-test effect sizes. Most of the dots are within the cone, and seemed to be evenly distributed on both sides of the centered line, which represents the overall average effect sizes across the all unique samples (on the assumption that the studies come from a single population; no between-study variance was added; *mean* = .70, *SE* = .04, $p < .001$); therefore, I could assume that the scatter plot is symmetric, indicating a possible absence of publication bias.

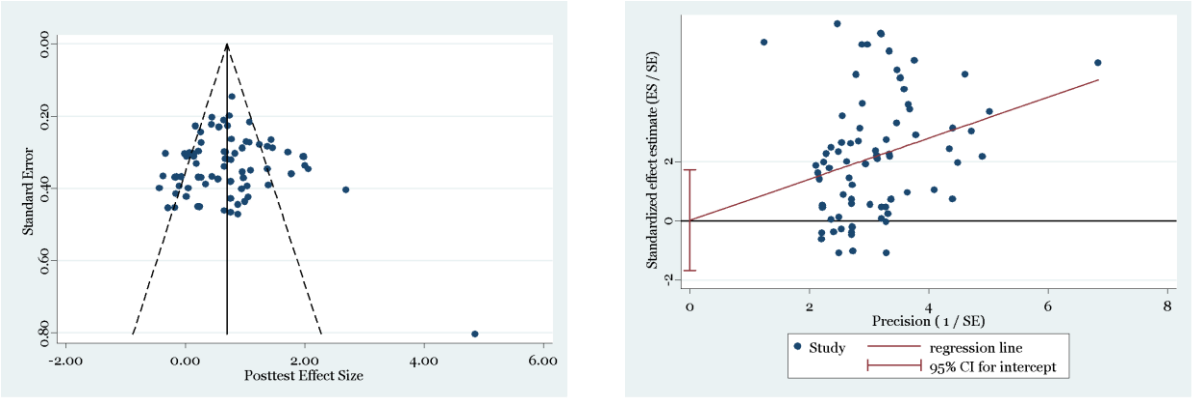

Figure A.1. Funnel Plot (left) and Egger's Test (right) for Post-test Effect Sizes.
Note. Each dot indicates a post-test ES. The dotted lines of the cone in the left graph represent 95% confidence intervals. The intercept of the regression line in the right graph indicates a standard normal deviate with expected value of zero for no publication bias in effect size calculation.

In addition to this visual judgment, I ran Egger's test to evaluate the statistical significance of asymmetry of the funnel plot by checking whether the intercept in a linear regression of standardized effect estimates (effect size / standard error) on precision (1 /

standard error; see the right graph in Figure A.1) is significantly different from zero (Egger et al., 1997). The results of the test confirmed that my funnel plot's asymmetry is not different from zero (*intercept* = .02, *SE* = .86, *p* = .978 [95% Conf. Interval: -1.68 ~ 1.73]), indicating no evidence of asymmetry in the funnel plot. Taken together, the results of the robustness and sensitivity checks revealed that there was no significant publication bias among the calculated effect sizes.

APPENDIX 1.4: Equations of Multilevel Meta-Analysis

An example equation of a general multilevel model regression analysis is as follows:

$$[1]\ Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$$

, where Y_{ij} is the observed outcome (e.g., a calculated ES) i in study j, γ_{00} is an estimated mean of the regression line (e.g., mean ES estimate), u_{0j} is level 2 variance, and r_{ij} is level 1 variance. In particular, the level 2 variance in the equation shifts the regression line for the level 1 units (i.e., effect sizes) up or down by level 2 units (i.e., unique samples).

The equations of unconditional (i.e., no covariate) multilevel models are as follows:

$$[2]\ \textit{Post-test ES}_{ij} = \gamma_{00} + u_{0j}$$

$$[3]\ \textit{Follow-up ES}_{ij} = \gamma_{00} + u_{0j}$$

, where γ_{00} is the grand mean of the equation (i.e., mean ES estimates in this case) and u_{0j} is level 2 variance. Our multilevel models differ from a common multilevel modeling (e.g., equation [1]) in that level 1 variance (r_{ij}) is omitted, because my model is a variance-known model. In other words, I already know the level 1 variance (r_{ij}) because the dependent variable consists of the computed ESs, and we know their sampling errors: the variance (the square of the standard errors) of the ESs.

In addition, for the moderator analysis, I ran two more regression analyses with the variables of interest included on the right-hand side of the equations, one for the publication and population data (equation [4]) and the other for the treatment data (equation [5]). Example equations of this analysis are as follows:

$$[4]\ \textit{Post-test ES}_{ij} = \gamma_{00} + \gamma_{10}\times\textit{Publication type}_{ij} + \gamma_{20}\times\textit{Region}_{ij} + \gamma_{30}\times\textit{Proficiency}_{ij} + \gamma_{40}\times\textit{Specialty}_{ij} + u_{0j}$$

[5] $Post\text{-}test\ ES_{ij} = \gamma_{00} + \gamma_{10} \times Interaction\ type_{ij} + \gamma_{20} \times Corpus\ type_{ij} + \gamma_{30} \times L2\ vocabulary\ dimension_{ij} + \gamma_{40} \times Training_{ij} + \gamma_{50} \times Duration_{ij} + u_{0j}$

, where γ_{00} is the grand mean of the equation when all the moderators have the reference values, γ_{10} in equation [4] is the mean ES difference between the values of the publication, γ_{50} in equation [5] is the mean ES difference between the values of the duration variable, and the like.

For the follow-up ESs, however, I did not choose certain variables, but found that the variables excluded in equations [6] and [7] could not be included in the model because the variables had either a multicollinearity issue or small samples.

[6] $Follow\text{-}up\ ES_{ij} = \gamma_{00} + \gamma_{10} \times Publication\ type_{ij} + \gamma_{20} \times Proficiency_{ij} + \gamma_{30} \times Specialty_{ij} + u_{0j}$

[7] $Follow\text{-}up\ ES_{ij} = \gamma_{00} + \gamma_{10} \times Interaction\ type_{ij} + \gamma_{20} \times L2\ vocabulary\ dimension_{ij} + \gamma_{30} \times Duration_{ij} + u_{0j}$

APPENDIX 1.5: Forest Plots for Single Effect Size Approach

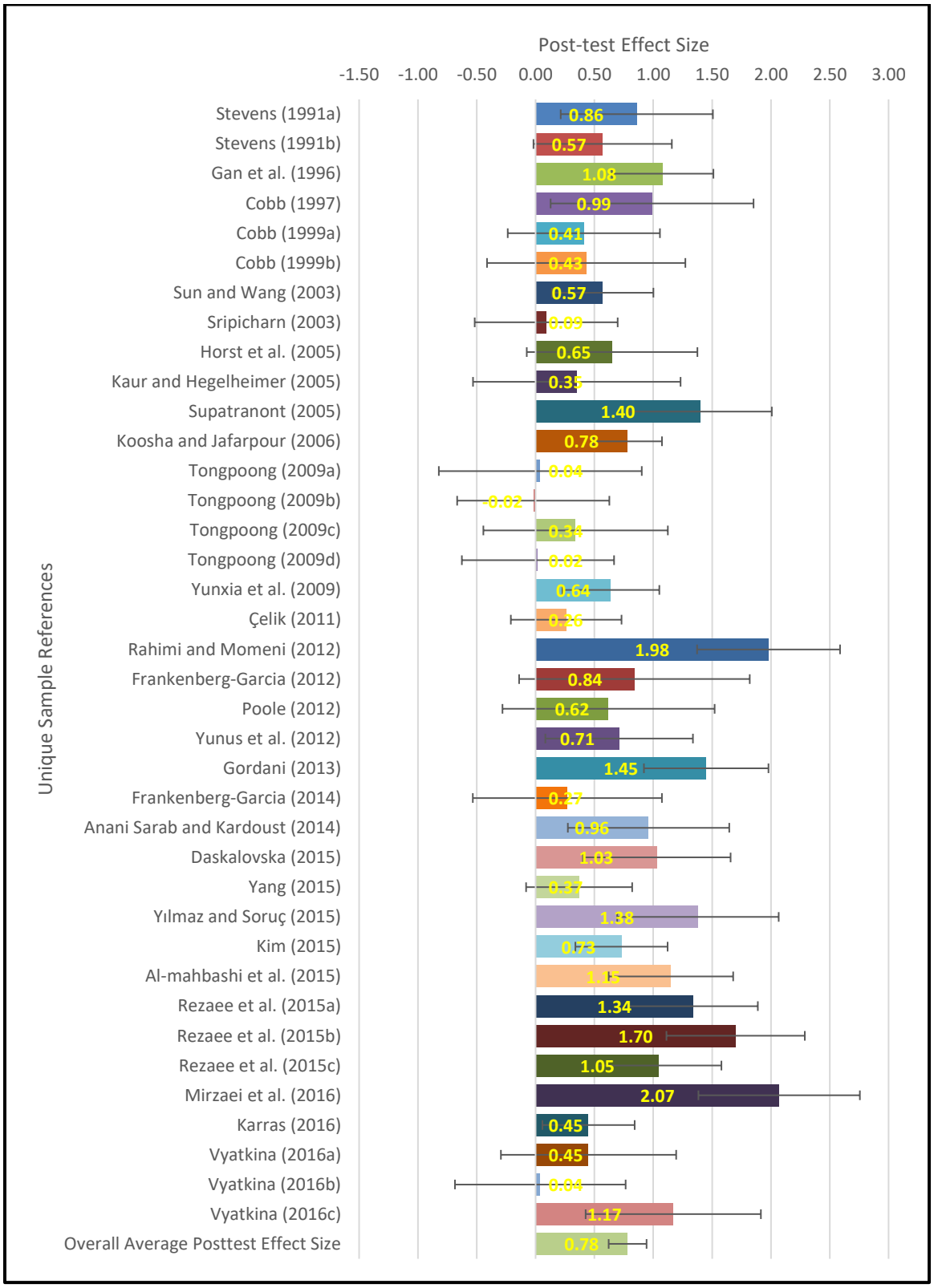

Figure A.2. Forest Plot for Single Post-test Effect Sizes for Each Unique Sample

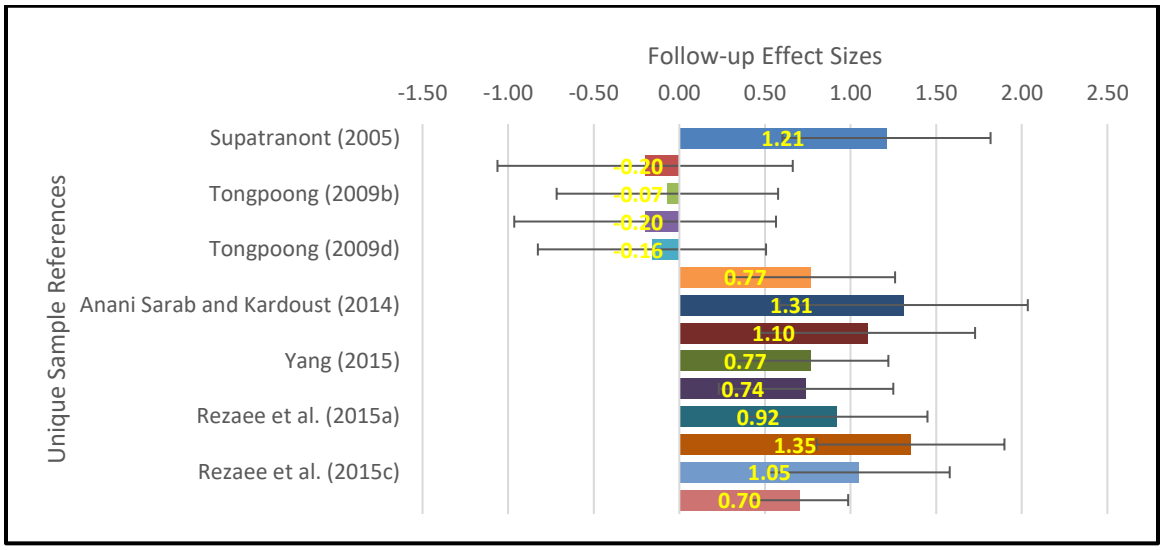

Figure A.3. Forest Plot for Single Follow-up Effect Sizes for Each Unique Sample

APPENDIX 1.6: Multilevel and Clustered Regression Models

Table A.1

Two Regression Analyses for Publication & Population Data Moderators

Independent Variables of Interest	Post-test Effect Sizes		Follow-up Effect Sizes	
	Multilevel Regression	Clustered Regression	Multilevel Regression	Clustered Regression
	Adjusted Means (Predicted margins)	Adjusted Means (Predicted margins)	Adjusted Means (Predicted margins)	Adjusted Means (Predicted margins)
1. Publication & Population Data				
(1) Publication type				
A. Journal article	0.74*** (0.08)	0.72*** (0.06)	0.99*** (0.14)	0.95*** (0.06)
B. PhD dissertation	0.42* (0.18)	0.46* (0.21)	0.08 (0.16)	0.09 (0.24)
C. Conference paper / Book chapter	1.08* (0.45)	1.15** (0.38)	-	-
(2) Region				
A. Asia	0.53*** (0.11)	0.53*** (0.09)		
B. Middle East	1.05*** (0.13)	1.05*** (0.13)	Not estimable	
C. Other (e.g., Europe and US)	0.53*** (0.15)	0.51*** (0.14)		
(3) Proficiency				
A. Low	0.47*** (0.13)	0.49** (0.14)	0.29 (0.18)	0.34** (0.10)
B. Intermediate	0.69*** (0.09)	0.73*** (0.10)	0.57*** (0.13)	0.69** (0.17)
C. High	1.27*** (0.31)	1.23*** (0.20)	0.79* (0.37)	0.93*** (0.20)
D. Mixed	0.74* (0.35)	0.62* (0.27)	-	-
(4) Specialty				
A. Languages	0.62** (0.20)	0.67*** (0.19)	0.36 (0.29)	0.43** (0.14)
B. Other	0.54*** (0.14)	0.53** (0.15)	0.23 (0.49)	0.35*** (0.05)
C. Mixed	0.75*** (0.09)	0.78*** (0.09)	0.56*** (0.12)	0.68*** (0.12)
Number of ES (n)	77	77	34	34
Number of Unique Sample (k)	38	38	13	13

Table A.2

Two Regression Analyses for Treatment Data Moderators

Independent Variables of Interest	Post-test Effect Sizes		Follow-up Effect Sizes	
	Multilevel Regression	Clustered Regression	Multilevel Regression	Clustered Regression
	Adjusted Means (Predicted margins)	Adjusted Means (Predicted margins)	Adjusted Means (Predicted margins)	Adjusted Means (Predicted margins)
2. Treatment Data				
(1) Interaction type				
A. Paper-based	0.55*** (0.14)	0.53* (0.20)	0.70*** (0.19)	0.72* (0.28)
B. CALL program	0.70** (0.23)	0.73*** (0.17)	-	-
C. Concordancer	0.72*** (0.14)	0.75*** (0.14)	0.39*** (0.09)	0.48*** (0.08)
D. Mixed (e.g., paper-based + concordancer)	1.30*** (0.32)	1.23*** (0.30)	1.11*** (0.27)	1.18*** (0.13)
(2) Corpus type				
A. Public corpus (e.g., Brown, BNC, OANC)	0.65*** (0.11)	0.67*** (0.12)		
B. Local corpus (e.g., own, specialized, graded)	0.59** (0.20)	0.63*** (0.17)	Not estimable	
C. Pre-selected concordance lines	0.98*** (0.28)	0.99** (0.32)		
(3) L2 vocabulary dimension				
A. Precise knowledge	0.40** (0.13)	0.42* (0.16)	0.29 (0.17)	0.37 (0.21)
B. In-depth knowledge	0.91*** (0.10)	0.87*** (0.09)	0.77*** (0.13)	0.86*** (0.12)
C. Productive use ability	0.55** (0.18)	0.43*** (0.11)	0.18 (0.17)	0.21 (0.14)
(4) Training				
A. Not received	0.58** (0.22)	0.61** (0.19)	Not estimable	
B. Received	0.72*** (0.08)	0.72*** (0.10)		
(5) Duration				
A. Short (> 2 hours in total or only 1 session)	0.68** (0.21)	0.64* (0.25)	1.24*** (0.22)	1.28*** (0.06)
B. Medium (about 3 to 8 sessions)	0.55*** (0.13)	0.56*** (0.13)	0.23* (0.11)	0.32* (0.11)
C. Long (< 10 sessions in total)	0.90*** (0.14)	0.86*** (0.14)	0.75*** (0.16)	0.77*** (0.18)
Number of ES (*n*)	77	77	34	34
Number of Unique Sample (*k*)	38	38	13	13

APPENDIX 2.1: Process of Selecting Example Concordance Lines

The following illustrates the process of selecting example concordance lines for one of the target vocabulary item, "inflection," which has the meaning of "a change in the pitch or tone of a person's voice" in the target context. Below are sample concordance lines selected from BNC, OANC, and Brown corpus.

i. It was purely to bring his ear reverentially into line with the mouth of whoever was speaking. "Exactly," he murmured. "Exactly." And Dyson knew from the depth of humility and reverence in his "INFLECTIONS" that he was getting a larger fee than even Lord Boddy (from the BNC).

ii. When you deal with customers over the phone, you have a whole new set of etiquette rules. The minute you pick up the phone, body language disappears, and your "INFLECTIONS" and tone of voice, and the words you use become the entire story (from the OANC).

iii. Godunov, it is the consistency with which every person on the stage—including the chorus—comes alive in the music. Much of this lifelike quality results from Mussorgsky's care in basing his vocal line on natural speech "INFLECTIONS" (from the Brown Corpus).

I made the following decision in terms of selecting concordance lines for the target word "inflection."

(1) The concordance line (i) was excluded. Its surrounding context requires further information to be comprehended, and there are many unfamiliar words and phrases, such as "the depth of humility" and "reverence," along with the key

word. Furthermore, there is no obvious clue for inferring the meaning of "inflection" from the context.

(2) The concordance line (ii) was selected because the surrounding words and structures are not only comprehensible to the participants, but also clearly indicate the meaning of "inflection" as a modulation of intonation in the voice.

(3) The concordance line (iii) was not selected, although there are some words that allow for the meaning inference, such as "music," "vocal line," or "speech." The reason is that the given clues are not strong enough to provide the aforementioned definition of "inflection," but may induce faulty meaning inferences.

APPENDIX 2.2: List of Target Vocabulary and Their Definitions

1. First reading text

(1) endowments: an attribute of the mind or body; natural talents or qualities.

(2) idiot savants: a mentally defective person with an exceptional skill or talent in a special field.

(3) fib: a small or trivial lie; minor falsehood.

(4) corroborated: being supported to be more certain; be confirmed.

(5) are beset with: being harassed by something; being attacked on all sides.

(6) misnomer : a misapplied or inappropriate name or designation.

(7) a vestige of: visible evidence of something that is no longer present or in existence.

(8) dismayed: being loss of courage completely, disheartened thoroughly

(9) nipping at the heels of: trying to be almost as good as someone that you are competing with.

(10) a plethora of: overabundance; excess; too many; a lot of.

2. Second reading text

(1) dodgy: untruthfully tricky; uncertain or unreliable.

(2) tuck: to put into a small, close, or concealing place.

(3) lumbering: moving clumsily or heavily.

(4) get on with: to proceed with; to begin or continue; to work with.

(5) cracked: pass through (a barrier); break through.

(6) deluded: deceived; misguided; the mind or judgment is misled.

(7) mucky: filthy, dirty, or slimy.

(8) traipse: to walk or go aimlessly or idly.

(9) grannies: a grandmother; an elderly woman.

(10) double-glazing: two panes of glass in a window.

3. Third reading text

(1) interrogators: one who asks questions of (someone, especially a suspect or prisoner) closely, aggressively or formally.

(2) polygraphs: a machine designed to detect and record changes in physiological characteristics, such as a person's pulse breathing rates; used as a lied detector.

(3) rationale: a strong reason to support for something.

(4) under duress: under pressure; forcibly restraint or restricted.

(5) electrodermal: related to electrical properties of the skin.

(6) plea: serious and emotional request for something.

(7) in the vicinity of: the area around or near a particular place.

(8) inflections: a change in the pitch or tone of a person's voice.

(9) latency: the state of being inactive or late.

(10) map out: to plan or sketch.

APPENDIX 2.3: Example Texts and Hyperlinks

1. First reading text (Hyperlinks to the texts: CONC / CODI / CTRL)

Education: fact or myth?

Do you think all of us have the endowments possessed by so-called "idiot savants" -- as depicted by Dustin Hoffman in the film Rain Man? This passage will talk about few stories which sometimes are regarded as a fib.

First. To give your children a head start in life, sit them in front of the television. A study of 200 American kids has showed that babies who watch TV for two hours a day develop more quickly than those who do without. Also this argument has been corroborated by the fact that on average, the two- and three-year-olds who watched TV scored 10 percent higher in English and Mathematics. However, the programs have to be directed towards their age group because it turned out that children derive no benefits from watching TV designed for adults. But the positive impact of TV dwindles with age, reports The Sunday Times. Older children who watch more than 16 hours of TV a week carry out worse than their peers.

Second. Actually most of us are beset with a belief that the early bird catches the worm. But humans would be a misnomer in this phrase if you try to apply this saying to them blindly. According to recent research, however, people who live around in bed in the morning and work into the evening are more intelligent. The scientists asked 400 volunteers to fill in questionnaires to work out if they considered themselves early-rising 'morning types' or late-working 'evening types'. Each was then subjected to mental ability and memory tests. The researchers discovered that the evening types had significantly better mental speed and memory. The results indicate that evening types are more likely to have higher intelligence scores, contrary to conventional folk wisdom. He also argued that the link between intelligence and working late may be a vestige of prehistoric times, when those who were still alert after dark would be more likely to survive attacks by night-time predators.

Third. Pushy parents may be doing their children more harm than good. Professional parents frequently over-stimulate babies and youngsters and buy them educational toys that are too old for them in the belief that they are improving their prospects. In fact, faced with such demands, the children may become dismayed completely. Worse still, the children recognize that they are disappointing their parents and this sense of failure will be nipping at the heels of their self-esteem. The warning comes as an ever-increasing range of educational material is being produced for the very young. In the US, hyper-parenting is common. Expectant mothers are pressured into buying classic music CDs in that they think this music would help build their babies' brain. By the age of one, enrollment in a plethora of classes, including English and Mathematics, is obligatory to babies.

2. Second reading text (Hyperlinks to the texts: CONC / CODI / CTRL)

What didn't come to pass

Forecasting what life is going to be like years down the line is a dodgy business. Even the experts don't always get it right. Take Bill Gates, for example. In 1981, he firmly stated that '640K of memory ought to be enough for anyone.' So it's more than a bit embarrassing for him now that, even a standard issue home PC, you need 200 times that amount of memory just to run his own company's software. Fortunately for Bill, other predicted that the technological future would involve giant computers that were the size of cities, whereas what we actually have are ever-shrinking models that you can tuck neatly into your pocket, which are hundreds of times more powerful than their lumbering old computers.

They imagined the robots of the future would not only be able to think for themselves, but get on with the housework too. Now what have we got? Absolutely no sign of a helpful house robot to mix a perfect beverage at the end of a hard day. Face it. I haven't even cracked the level of robotic vacuum cleaners yet. In the same sense, there has been famous cryogenics super-salesmen who have persuaded some people to part with vast sums of money on a promise that they will ice their customers and will defrost them when 'the time is right' may be 2052. But since we have not experienced perfect freezing strawberries yet, these poor deluded people may be nothing more than mucky water puddles by 2052.

Another two pieces of idea: One, nutritionally-perfect pills to replace all our food! Second, Only online shopping will be available, so there's no need to traipse around the shops! Both have met with a resounding thumbs-down from the public. I simply refuse to give up eating our nutritionally nightmarish fish and chips. And we show absolutely no inclination to forego and the pleasure of touching, examining and trying the purchases we make. I love our food and our shopping, thank you very much.

Next concern is our reproductive function. For instance we worry that come 2052, it will be increasingly normal for grannies to be giving birth, or that male pregnancy will be possible. It's my bet that if you asked 100 women in their sixties, now or even in 2052, if they wanted a test-tube baby or double-glazing windows at their home, 99 percent would opt for the windows. As for male pregnancy, I have it filed under 'o' as in 'Only for the lunatic', along with human cloning and genetic engineering. Yes, it might all be technically possible, and you might well see genetic engineering for very specific and well-defined medical reasons, but it will remain risky for the baby. It's an unchangeable part of human nature that what we really want, above everything else, is the best for our future generations.

3. Third reading text (Hyperlinks to the texts: CONC / CODI / CTRL)

How do you know when someone is lying?

The Korean used rice. An examination for truthfulness might go something like this: "Is your surname Kim?"" (They know the guy's surname is, in fact, Kim.) When Mr. Kim answers correctly, then the interrogators hand Mr Kim some rice. They have already counted the number of rice grains. Mr. Kim put the handful of rice in his mouth and spit it out after holding it for three seconds. Then they count how many rice grains come out. After that, they ask another question to check if he is telling a lie or not: "Did you steal the chicken?" After Mr. Kim responds, he again would put rice grains in his mouth and spit it out after three seconds. Again, they knew how many grains went in, and they count how many come out. If more grains come out after the question about the stolen chicken than came out after the "easy" question, where the suspect truthfully gave his name, they know he's lying. How? The stress of being caught lying makes the suspect's mouth drier. Fewer grains stick. More come out. Mr Kim stole the chicken.

Modern lie detectors - also known as "polygraphs" rely on the same basic rationale - that lying causes bodily changes, which can be detected and measured. Having agreed to do the test (if the test is done under duress, the extra stress caused makes the test unreliable), the suspect is connected to three devices measuring blood pressure, breathing rate and electrodermal response. Increased activity in these areas suggests increased stress -- which means the subject might be lying. Lie detectors have been widely used in the US since the 1950s but they remain controversial and their results are not always accepted by courts. For example, the results of a test taken by the British babysitter Louise Woodward to support her plea of not guilty to killing a child in her care were not admitted as evidence at her trial in the vicinity of Massachusetts.

Cheaper and faster than a polygraph, the voice stress analyser, or VSA is based on the premise that our voice changes when we are under stress -- when we're lying for example. The VSA detects the changes, and will work on a telephone, tape recording or from the next room via a wireless mic or bug. The analyzer monitors the subject's voice patterns and inflections, and electronically evaluates their relative stress patterns to determine if they are lying or not.

The period of time between the last word of an investigator's question and the first word of the subject's response is known as "Response latency". Research tells us that the average response latency for subjects who are telling the truth is 0.5 seconds, whereas the average latency for liars is 1.5seconds. This is because the subject is mentally considering whether to tell the truth, part of the truth, or a complete lie. Also the subject needs more time to map out an escape route! Latencies of two or three seconds should be regarded as highly suspicious. In other words, he who hesitates is probably lying!!

APPENDIX 2.4: Equations for Regression Models

The first residualized change model (Model 1 in Table 2.3) included variables for treatment conditions. The independent variables are the participants' pre-test results, prior English proficiency (participants' official TOEIC scores, as developed by Educational Testing Service), and gender (male = 0 and female = 1), whereas their post-test results are the dependent variable. In addition, to detect any difference in the treatment effects between the three conditions, three dummy variables were generated that identified different conditions ("CONC" is the reference group among the "CTRL," "CONC," and "CODI" conditions), in the second regression equation (i.e., Model 2 in Table 2.3). The third equation includes trials (the first = 1, the second =2, and the third = 3, among three different trials), and order effect (the interaction effect of delivering different conditions in different orders/trials) as additional independent variables, in order to detect any order effects (i.e., Model 3 in Table 2.3). The equation of Model 3 is as follows.

$$(1)\ (Post\text{-}test)_{ij} = a + b_1(CTRL)_{ij} + b_2(CODI)_{ij} + b_3(Trial)_{ij} + b_4(Order\ effect)_{ij}$$
$$+ b_5(Pre\text{-}test)_{ij} + b_6(Eng\_proficiency)_{ij} + b_7(Gender)_{ij} + \varepsilon_{ij.}$$

Second, an additional regression analysis, including the classroom fixed-effects ($\delta_{classroom}$), was conducted to only capitalize on within-classroom differences after removing between-classroom differences that could bias the estimation of the treatment effects. This approach was part of an effort to eliminate any possible discrepancies between the participants' intact classrooms. The equation of Model 4 is as follows.

$$(2)\ (Post\text{-}test)_{ij} = a + b_1(CTRL)_{ij} + b_2(CODI)_{ij} + b_3(Trial)_{ij} + b_4(Order\ effect)_{ij} +$$
$$b_5(Pre\text{-}test)_{ij} + b_6(Eng\_proficiency)_{ij} + b_7(Gender)_{ij} + \delta_{classroom} + \varepsilon_{ij.}$$

Third, I employed simple change models in order to check the robustness of the results of the aforementioned residualized change models. These equations considered the participants' meaning-recall knowledge gains per each condition (calculated by subtracting pre-test scores from post-test scores) as the dependent variable, and included all the independent variables, except for

the pre-test results variable (i.e., Models 1, 2, and 3 in Table 2.4). The equation of Model 3 is as follows.

$$(3)\ (\Delta Score;\ Post\text{-}test - Pre\text{-}test)_{ij} = \Delta a + b_1(CTRL)_{ij} + b_2(CODI)_{ij} + b_3(Trial)_{ij} + b_4(Order\ effect)_{ij} + b_5(Eng\_proficiency)_{ij} + b_6(Gender)_{ij} + \delta_{classroom} + \Delta\varepsilon_{ij.}$$

APPENDIX 2.5: Classroom Fixed-Effects in Tables 2.3 and 2.4

To find a more accurate and robust estimation of the effects of different conditions on the participants' meaning-recall test scores, I conducted an additional regression analysis with classroom fixed-effects (see Model 4 in Table 2.3), in addition to the residualized change models (Models 1, 2, and 3 in Table 2.3). The results showed that coefficients did change but by a very small amount relative to their standard errors. In other words, it appeared that the fixed effects adjustments produced small changes in the coefficients.

As a part of the efforts to check the robustness of the findings, I additionally conducted simple change regression analyses. Although participants received, on average, nearly zero for their pre-tests (see Table 2.2), it should be noted that everyone may have experienced different amounts of gains (i.e., post-test – pre-test) throughout the experiment. Since the residualized change models only focus on the within-group mean, which may imply a regression toward the mean between groups, simple change models make more sense in this case by focusing on the participants' individual gains across the experiment.

Models 1, 2, and 3 in Table 2.4 included change scores (i.e., "post-test – pre-test" per each condition) as the dependent variable, instead of post-test scores, without having the pre-test variable as one of the independent variables in its regression equation. The results revealed that different treatment conditions are still significant predictors of the participants' vocabulary gains ($b = 2.02$, $p < .001$; see Model 1 in Table 2.4). In particular, CTRL would, on average, lead a participant to gain a 1.83 lower vocabulary score than CONC ($b = -1.83$, $p < .01$), and participants who were given CODI, on average, would gain a

2.20 higher score than those given CONC (b = 2.20, p < .01). Furthermore, it was re-confirmed that there was no order effect (b = .25, p > .05), in accordance with receiving different conditions in different orders (see Model 2 in Table 2.4).

Finally, I conducted an additional regression analysis with classroom fixed-effects (see Model 3 in Table 2.4). The results showed a similar pattern to that of the fourth model (i.e., Model 4 in Table 2.3), in which most of the standard errors got larger and the coefficients on the key predictors increased as well, i.e., the main effect of providing different conditions became even larger after removing variations across classrooms.

APPENDIX 3.1: GMMs in the mclust Package

Rather than identifying clusters by physical distance from centroids, as traditional clustering techniques do (e.g., hierarchical clustering, partitioning clustering), the *mclust* package allows us to use the EM algorithm to find the most likely set of clusters based on Gaussian mixture models. As its name implies, these models are mixtures of Gaussian probability distributions for each cluster, and the most recent *mclust* package has 14 underlying Gaussian models (see Table A.3). Each of these models takes the following form (adapted from Scrucca et al., 2016; Soto-Valero, 2017):

$$f(x) = \sum_{k=1}^{G} \pi_k f_k(x)$$

wherein x is the sample of observations (x_1, …, x_n), G is the number of clusters (mixture components), π_k is the probability that any observation belongs to cluster k (such that $0 < \pi_k < 1$; so, the sum of π_k for all k [1, …, G] is 1), and $f_k(x)$ is the Gaussian probability density function of the observation x in cluster k.

Table A.3
General Characteristics of 14 GMM Models Included in mclust Version 5.3

Models	Distribution	Volume	Shape	Orientation	Reference
EII	Spherical	Equal	Equal	-	Banfield & Raftery (1993)
VII	Spherical	Variable	Equal	-	Banfield & Raftery (1993)
EEI	Diagonal	Equal	Equal	Coordinate axes	Banfield & Raftery (1993)
VEI	Diagonal	Variable	Equal	Coordinate axes	Banfield & Raftery (1993)
EVI	Diagonal	Equal	Variable	Coordinate axes	Banfield & Raftery (1993)
VVI	Diagonal	Variable	Equal	Coordinate axes	Banfield & Raftery (1993)
EEE	Ellipsoidal	Equal	Equal	Equal	Banfield & Raftery (1993)
EVE	Ellipsoidal	Equal	Variable	Equal	Browne & McNicholas (2014)
VEE	Ellipsoidal	Variable	Equal	Equal	Browne & McNicholas (2014)
VVE	Ellipsoidal	Variable	Variable	Equal	Celeux & Govaert (1995)
EEV	Ellipsoidal	Equal	Equal	Variable	Banfield & Raftery (1993)
VEV	Ellipsoidal	Variable	Equal	Variable	Banfield & Raftery (1993)
EVV	Ellipsoidal	Equal	Variable	Variable	Celeux & Govaert (1995)
VVV	Ellipsoidal	Variable	Variable	Variable	Banfield & Raftery (1993)

The goal of this data mining technique is thus to find the optimal number of clusters (G) on the assumption that G is fixed and to assign the sample of observations (x) to each cluster, using its model-based maximum likelihood method. For this reason, the optimal model can be determined by the Bayesian information criterion (BIC), widely used for model selection (such that the highest BIC is preferred in the case of *mclust*). As part of the goal, a finite mixture model can be obtained via the EM algorithm. For example, the model goes through the following three steps: (1) in the initial step, the model allows the cluster memberships to be hidden variables and randomly assigns cluster memberships to each data point; (2) in the expectation (E) step, the model estimates the probability that each cluster includes or does not include each data point to the current cluster membership; and (3) in the maximization (M) step, the model modifies the cluster memberships to maximize the likelihood of the model that was generated in the previous expectation step.

In this way, the algorithm repeats the second and third steps (i.e., expectation and maximization steps – EM algorithm) until it reaches the maximum likelihood fit, and *mclust*, a model-based clustering technique, simplifies the maximum likelihood method by using the previously defined parsimonious covariances matrices – 14 Gaussian models, that have been proposed and studied in previous literature. Assuming a Gaussian distribution for each cluster, clusters are predicted to be ellipsoidal, centred at the mean of their values; therefore, their geometric characteristics, such as distribution, volume, shape, and orientation, can be pre-defined for a more accurate and precise prediction, and the 14 Gaussian models are composed of different combinations of these geometric characteristics (Table A). For more information about this data mining technique, I encourage readers to refer to Scrucca et al. (2016).

References

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*(3), 803–821. https://doi.org/10.2307/2532201

Browne, R. P., & McNicholas, P. D. (2014). Estimating common principal components in high dimensions. *Advances in Data Analysis and Classification*, *8*(2), 217–226. https://doi.org/10.1007/s11634-013-0139-1

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793. https://doi.org/10.1016/0031-3203(94)00125-6

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *The R Journal, 8*(1), 289–317. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/

Soto-Valero, C. (2017). A Gaussian mixture clustering model for characterizing football players using the EA Sports' FIFA video game system. *RICYDE. Revista Internacional de Ciencias del Deporte, 13*(49), 244–259. https://doi.org/10.5232/ricyde2017.04904

APPENDIX 3.2: Complete Results of Regression Analyses

Table A.4

Results of Multiple Regression Analysis (for Table 3.2)

DV: SQRT of Post-test	Total Sample (N = 132)				DDL-sufficient learners (n = 82)				DDL-insufficient learners (n = 50)			
	Coefficients (SE)		Predicted values		Coefficients (SE)		Predicted values		Coefficients (SE)		Predicted values	
			Mean	SE			Mean	SE			Mean	SE
Condition												
CTRL	-0.547*	(0.207)	1.755***	(0.142)	-0.608*	(0.216)	2.164***	(0.121)	-0.441	(0.291)	1.091***	(0.168)
CONC	(reference)		2.302***	(0.111)	(reference)		2.772***	(0.132)	(reference)		1.533***	(0.139)
CODI	0.556*	(0.170)	2.8658**	(0.102)	0.404	(0.198)	3.176***	(0.096)	0.797**	(0.196)	2.329***	(0.183)
TOEIC	0.003***	(0.000)			0.002**	(0.000)			0.001	(0.001)		
SQRT of Pre-test	0.241	(0.147)			0.162	(0.095)			0.203	(0.339)		
Female	-0.268	(0.511)			0.146	(0.163)			-0.772	(0.746)		
Constant	-0.147	(0.368)			1.479*	(0.449)			0.855	(0.627)		

Note. DV = dependent variable. SQRT = square root. SE = standard error.

* p < .05, ** p < .01, *** p < .001

Table A.5
Results of Multiple Regression Analysis (for Table 3.4)

※ Before clustering

DV: SQRT of Post-test	Total Sample (N = 132)					
	CTRL		CONC		CODI	
	Coef. (SE)		Coef. (SE)		Coef. (SE)	
TOEIC	0.002*	(0.001)	0.004***	(0.001)	0.003***	(0.001)
SQRT of Pre-test	0.202	(0.157)	0.317	(0.191)	0.286	(0.207)
Female	0.360	(0.532)	-0.530	(0.583)	-0.666	(0.553)
Constant	0.227	(0.625)	-0.944	(0.649)	0.339	(0.612)

※ After clustering

DV: SQRT of Post-test	DDL-sufficient learners (n = 82)						DDL-insufficient learners (n = 50)					
	CTRL		CTRL		CTRL		CTRL		CONC		CODI	
	Coef. (SE)		Coef. (SE)		Coef. (SE)		Coef. (SE)		Coef. (SE)		Coef. (SE)	
TOEIC	0.001	(0.001)	0.002*	(0.001)	0.002*	(0.001)	-0.002	(0.002)	0.003*	(0.001)	0.002	(0.002)
SQRT of Pre-test	0.126	(0.145)	0.270	(0.179)	0.136	(0.171)	0.117	(0.392)	0.191	(0.338)	0.487	(0.601)
Female	0.400	(0.270)	0.268	(0.268)	-0.279	(0.348)	0.334	(0.691)	-1.551*	(0.599)	-1.096	(0.719)
Constant	1.495*	(0.613)	1.397*	(0.587)	1.400*	(0.659)	2.101	(1.107)	-0.381	(0.957)	1.233	(1.152)

Note. DV = dependent variable. SQRT = square root. Coef. = coefficient. SE = standard error.
* p < .05, *** p < .001

APPENDIX 4.1: Reading Passage and Target Vocabulary

What didn't come to pass

Forecasting what life is going to be like years down the line is a ***dodgy*** business. Even the experts don't always get it right. Take Bill Gates, for example. In 1981, he firmly stated that "640K of memory ought to be enough for anyone." So, it's more than a bit embarrassing for him now that, even a standard issue home PC, you need 200 times that amount of memory just to run his own company's software. Fortunately for Bill, other predicted that the technological future would involve giant computers that were the size of cities, whereas what we actually have are ever-shrinking models that you can ***tuck*** neatly into your pocket, which are hundreds of times more powerful than their ***lumbering*** old computers.

They imagined the robots of the future would not only be able to think for themselves, but get on with the housework too. Now what have we got? Absolutely no sign of a helpful house robot to mix a perfect beverage at the end of a hard day. Face it, we haven't even ***cracked*** the level of robotic vacuum cleaners yet. In the same sense, there has been famous ***cryogenics*** super-salesmen who have persuaded some people to part with vast sums of money on a promise that they will ice their customers and will defrost them when "the time is right" may be 2052. But since we have not experienced perfect freezing strawberries yet, these poor deluded people may be nothing more than ***mucky*** water puddles by 2052.

Another two pieces of idea: One, nutritionally-perfect pills to replace all our food! Second, only online shopping will be available, so there's no need to ***traipse*** around the shops! Both have met with a resounding thumbs-down from the public. We simply refuse to give up eating our nutritionally nightmarish fish and chips. And we show absolutely no inclination to forego and the pleasure of touching, examining and trying the purchases we make. We love our food and our shopping, thank you very much.

Next concern is our reproductive function. For instance, we worry that come 2052, it will be increasingly normal for ***grannies*** to be giving birth, or that male pregnancy will be possible. It's my bet that if you asked 100 women in their sixties, now or even in 2052, if they wanted a test-tube baby or ***double-glazed windows*** at their home, 99 percent would opt for the windows. As for male pregnancy, I have it filed under 'o' as in "Only for the lunatic", along with human cloning and genetic engineering. Yes, it might all be technically possible, and you might well see genetic engineering for very specific and well-defined medical reasons, but it will remain risky for the baby. It's an unchangeable part of human nature that what we really want, above everything else, is the best for our future generations.

Reference

Cunningham, S., Moor, P., & Carr, J. C. (2003). *Cutting edge advanced with phrase builder*.

Harlow, UK: Pearson Education.

APPENDIX 4.2: DDL Materials for Target Vocabulary

dodgy [adjective]

1. From outrageous ticket price markups of up to 327% to denying refunds and selling fake tickets, the time has come to say fair's fair and stop these … dodgy business practices of ticket resale sites.

2. As I watched the story move around the web, I saw how the worlds of fake websites and fake news exist to reinforce one another and give falsehood credence. Many of the web sites quoted not the original, dodgy source, but one another.

3. There's a couple of things. Running a good anti-malware, antivirus program will help catch many of these. You know, don't download dodgy files on peer-to-peer networks. Don't accept, you know, files from people you don't know over instant messaging.

4. There are many more high-profile cases of dodgy business behavior such as the 2015 Volkswagen emissions scandal. There are also smaller examples of people such as accountants, financial planners and lawyers ripping off clients.

5. There is a reason why there has been radio silence on her claims from all major publications - her approach is very dodgy. She is using an anonymous account and refuses to make her identity known to the very journalists she wants to publish her claims.

tuck [verb]

1. Quickly, he examined the other one. It had held blue paint. He set them down and straightened up. He was about to tuck his hand back into his sleeve when he noticed a glint of color from several particles that had adhered to his fingers.

2. She watched Jen's Nikhil tuck his white uniform shirt into his white pants and walk out of the clinic. He always seemed to move as if an invisible crane were pulling him forward, always against his will.

3. Rory is yelling at Two, Liv is still crying, and I hear something crash to the floor. I tuck my phone in my shirt pocket and head toward the litany of tears. My daily mantra, "All is well. The Universe supports me," is on replay in my head.

4. We make plans to go to lunch the next day. He tells me Lily has a daughter--four years old. That she's cute as a button. He gives me a picture of the girl, and I tuck it into my wallet. I have one more drink with my father, and he falls asleep before it's finished.

5. Dr. Rabin advises people to tuck three to five funny thoughts into the mind, then recall those thoughts to extinguish the flames of stress. He said it's difficult to experience stress when recalling a happy or funny moment and smiling.

lumbering [adjective]

1. For the fourth straight day, I walked through the winding corridors of Piedmont Hospital, heels clicking on the tile floor. I had grown accustomed to the smell of antiseptics and the slow, lumbering elevators that carried me to the third floor.

2. Then there is the whole separate category of acerbity directed at William Jefferson Clinton. Mr. Clinton tends to am to poundage-his slow and lumbering morning run seems an act of contrition rather than of grace.

3. Even a $7,000 SA7, and these things, as I said, are very easy to buy in the black market, they pose a lethal threat to all civilian airliners because civilian airlines do not carry countermeasures. They're slow, lumbering planes. They're an easy target for even a 30 or 40-year-old surface to air missile.

4. You've got a lot of space out there over the South China Sea. What is their aircraft doing so close to ours? We had a slow, lumbering, relatively un-maneuverable aircraft; they had a fighter plane.

5. Then he dropped to the middle level, pulled up at the gate, and exited onto the street. From there, he took a slow, lumbering bus to his own neighborhood. It was the fashionable district for indigo bachelors.

crack [verb]

1. It debuted at number one on the Digital Songs chart and reached number two in the U.S. on Billboard's Hot 100, US Adult Top 40 and US Mainstream Top 40 lists. And it was Swift's 20th song to crack the top 10 on the Hot 100 list, making her just the sixth female music artist ever to achieve that feat.

2. All played within the past three seasons, including the past three Charlotte teams -- the product of Clifford's abdication of fast-breaking in favor of having all five guys protect the glass. (Last season's Pistons snared 81.2 percent of opponent misses, the only team ever to crack the 80 percent barrier.)

3. With his 3,465 hits, 14 All-Star appearances and five championship rings, Jeter isn't just a lock for Cooperstown, he is another candidate who will likely crack the top 10 in shares of the vote -- that's upwards of 97.2%, at this writing -- even with his defensive shortcomings.

4. The Frogs received the most votes of the unranked teams in each poll, but failed to crack the Top 25. # Arkansas rose to No. 24 in both polls on the strength of their win on Saturday. The Razorbacks had been unranked for the previous two polls.

5. As usual, Penske didn't rely on conventional wisdom. Penske's three drivers Briscoe, three-time Indy winner Helio Castroneves and Power, the points leader spent most of this week just trying to crack the top 10 of the speed charts.

cryogenics [noun]

1. You remember cryogenics? At the outset, people dismissed it as a rich man's folly. Eccentric millionaires freezing their brains, hoping to wake up in a new body. Even when cloning showed signs of making it feasible, it wasn't moral affront that caused the backlash.

2. In “Forever Young,” the actor plays an Army test pilot in 1939 who, crushed by the apparent accidental death of his beloved, insists that a scientist friend use him as a human guinea pig in an early cryogenics experiment, freezing his body for a year of forgetfulness.

3. What I was really fascinated with in making Death Warmed Up was the whole area of cryogenics, which is freezing people's brains and bodies to try and bring them back in the future when medical science is at a better level.

4. In the last half century, the science of cryogenics, or freezing humans to preserve them for reanimation, has had some spectacular failures when people inadequately frozen have simply started to decompose.

5. People who turn to cryogenics are usually captivated by the possibility of having their body preserved until some indeterminate future time when it is imagined that science and technology will be capable of curing any cause of death, repairing damaged tissues and, most importantly, bringing them back to life.

mucky [adjective]

1. One long heel sinks into the mud. The past days have brought late-summer rains to New Hampshire, and although the air is now dry, the grass between the parking areas and the dormitories is soft and mucky. This is a girl used to walking on city pavement, concrete. She laughs and pulls herself out.

2. The next day the sun was out, and while it was chilly, no rain made a big difference. The trail was still mucky, though, and that meant slow going. We adjusted our expectations and our attitudes and went at it, the sunlight filtering through the tree canopy, the fall leaves a Berber carpet of red, orange, yellow and burgundy.

3. No rushing off before the sun rises and I got to decide when to go to bed. Camp bed, that is. The weather was beautiful. I was secretly happy for a short break that our vehicles were not going through mucky trails during this sector of our journey.

4. The onset of monsoon brings along wet shoes and feet which in turn translate into fungal infections, athlete's foot and other diseases. Not to forget, those long hours you spend at work or elsewhere in those wet and mucky socks and shoes.

5. If you didn't have grass planted around and if it was raining and a little bit mucky, it would at least save your shoes or boots a little bit of wear and tear too by being able to step onto a solid stone step.

traipse [verb]

1. "There was nothing illegal about it," says archaeologist Jerry Spangler, "but it reflects the growing conflict between private landowners and tourists who traipse over their property without permission." So far, the sites at the Wilcox ranch are unscarred. The question now is how to keep them that way.

2. Many inns and hotels have begun to recruit travelers during the sugar season by offering them the chance to traipse through the woods, collect sap, and participate in making maple syrup. To contact the Trapp Family Lodge about sugar-season rates, call (802) 253-8511.

3. On days when I was free sometimes I would drive out to that splendid knoll he lived on and I would park my car in the communal parking area and I would stroll or traipse about, hunting the inhabitants of wildness with my camera which I had found one day in an intersection while delivering lumber to Stinson Beach.

4. At hall level, the two reception rooms are served by a new butler's pantry which now sits in what was a bathroom when the property was divided into flats. "I did this, so owners could have a glass of wine or a cup of coffee without having to traipse all the way to the kitchen downstairs," the owner says.

5. We checked into a hotel, and the room was very, very far from reception. My daughter said she needed a wee – I said to wait till we got to the room and walked past the public loo. But the key didn't work so we had to traipse back to reception, with the child asking where the toilet was, followed by a dramatic, "Oh, it just comed Mam, it's okay don't worry."

grannies [noun]

1. In Chinese, "dama" is a colloquial term used to describe rambunctious elderly women -- also called "aunties" or "grannies" -- who congregate in loud groups, dance in public squares and mind other people's business.

2. When Nick was born, eleven years later, I was back in the playground with the stroller, but by then, most of the mothers had gone off to work and the women pushing the swings were grannies and nannies. Today, I'm still in the park, but now I'm pushing a stroller with my grandson in it.

3. Not everybody got to actually talk to the queen, but she did wander along the streets very, very slowly, particularly talking to lots and lots of the children who had come today. They come in their school uniform, they come with their moms and dads, they come with their grannies and their great grannies, all turned out for a really special day marking a landmark in the queen's life.

4. The e-mail messages began as a trickle, carrying intriguing subject headings such as "Tired grandmother," "Feeling needed again!" and "Grannies are so precious." Within days, the trickle had grown to a small flood. Grandparents - make that grandmothers - were weighing in with personal experiences and perspectives on afront-page story.

5. With assistance from Door of Hope and social workers some babies are returned to the care of their mothers while others find a home and family at the Door of Hope Village – a community of moms, dad, uncles, aunties, uncles, grannies, grandpas and cousins to love and cherish them.

double-glazed windows [noun]

1. Although the estimated outdoor noise levels in our study were moderate, with the highest level of exposure comparable to the noise from loud conversation, indoor noise levels would have been further diminished because all residential buildings in Finland have good insulation and triple-glazed windows (minimum standard is double-glazed windows) against the harsh climate, which reduce the levels of traffic noise indoors.

2. It replaces the double-glazed windows normally used in apartments and offices with a complex structure that looks like a normal pane but has internal membranes and other devices that almost totally block the transmission of heat.

3. Chinese buildings use three times as much energy for heating as comparable U.S. ones, even though inside temperatures remain colder. By making boiler improvements and using insulation and double-glazed windows, the Chinese could raise average building temperatures from 11deg Celsius to 18deg -- while consuming 40% less coal.

4. Citadines also makes good use of natural lighting and ventilation to minimize environmental impact. More importantly, the apartments have double-glazed windows that prevent unwanted heat from coming in and which can help reduce medium to high frequency noise from outside.

5. I commented on how stunningly quiet the train was. He explained that was partly due to its double-glazed windows, which also prevented stones from coming in. I raised an eyebrow. "Yes, sometimes children throw them. They may break the outside window, but they won't break the inside."