ANNOTATION AND WORKING MEMORY IN SECOND LANGUAGE READING, INCIDENTAL VOCABULARY LEARNING, AND PERCEIVED COGNITIVE LOAD

BURCU VAROL

BOĞAZİÇİ UNIVERSITY

2017

ANNOTATION AND WORKING MEMORY IN SECOND LANGUAGE READING, INCIDENTAL VOCABULARY LEARNING, AND PERCEIVED COGNITIVE LOAD

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

English Language Education

by

Burcu Varol

Boğaziçi University

2017

Annotation and Working Memory in Second Language Reading, Incidental Vocabulary Learning, and Perceived Cognitive Load

The thesis of Burcu Varol

has been approved by:

Assoc. Prof. Gülcan Erçetin
(Thesis Advisor) _____

Assist. Prof. Günizi Kartal _____

Assist. Prof. Senem Yıldız _____

Assoc. Prof. H. Gülru Yüksel
(External Member) _____

Assist. Prof. Kadir Kozan
(External Member) _____

June 2017

DECLARATION OF ORIGINALITY

I, Burcu Varol, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature.....

Date .....10 June 2017.....

ABSTRACT

Annotation and Working Memory in Second Language Reading,

Incidental Vocabulary Learning, and Perceived Cognitive Load

This study attempts to explore the role of annotations and working memory capacity (WM capacity) in reading comprehension, incidental vocabulary learning, and the perceived cognitive load (CL) through a between groups design. The independent variables investigated in this study were annotation type (lexical versus topic-level) and annotation location (pop-up window versus separate window). One hundred-twenty high proficiency level second language learners were assigned to one of the four treatment conditions and were asked to read an electronic text wherein a built-in tracking software recorded their interactions. Upon reading, they were given free recall, multiple choice comprehension, vocabulary recognition, and vocabulary production tasks. Findings showed that the effects of annotation type changed according to the task used to gauge comprehension while lexical annotations were found to be the consistent determinant of short-term vocabulary learning. The effect of annotation location was less clear-cut as it produced different results in conjunction with different annotation types for different measures. WM capacity played a major role in reading comprehension, and especially when combined with the effect of pop-up window glossing conditions, it led to the best results. As for the results on CL, no significant effects were observed in the self-ratings and only a significant gloss type effect was detected in terms of the recorded annotation use. Implications drawn from this study can inform instructional designers as to the facilitative effects of annotation use that would not overload limited capacities of readers.

ÖZET

İkinci Dilde Okuma, Rastlantısal Kelime Öğrenimi, ve Bilişsel Yük Algısında Açıklayıcı Not ve İşler Bellek Kapasitesinin Rolü

Bu çalışma, açıklayıcı notların (AN) ve işler bellek (İB) kapasitesinin okuduğunu anlama, rastlantısal kelime öğrenme ve okurların bilişsel yük algısı üzerindeki etkisini gruplar arası bir araştırma deseniyle incelemeyi amaçlamaktadır. Çalışmada araştırılan bağımsız değişkenler, AN türü (kelime-seviyesinde ya da konu-seviyesinde) ve AN'in sunuluş yeridir (metin içinde açılır pencerede ya da metin dışında ayrı ekranda). Yüz yirmi ileri seviyede ikinci dil öğrencisi dört deney grubundan birine atanmış ve ortama entegre edilmiş bir izleme yazılımının olduğu elektronik bir metni okumaları istenmiştir. Okumaları bittikten sonra serbest hatırlama, çoktan seçmeli okuduğunu anlama, kelime tanıma, ve kelime türetme testlerini tamamlamışlardır. Sonuçlar AN türünün etkisinin okuduğunu anlama ölçeğine göre değiştiğini, kelime düzeyindeki AN'lerin özellikle kısa vadeli kelime öğreniminde tutarlı bir belirleyici olduğunu göstermiştir. AN'in sunuluş yerinin etkisi ise ölçek türüne ve AN türüne bağlı olarak farklılık göstermiştir. İB kapasitesi okuduğunu anlamada önemli bir rol oynamıştır ve özellikle notların açılır pencere sunulması ile birlikte en iyi sonuçları vermiştir. Bilişsel yük ile ilgili olarak, okuyucuların kendi beyanlarına göre hiçbir etmenin önemli bir etkisi olmamakla birlikte, kaydedilen verilere göre AN türünün önemli bir etkisi gözlemlenmiştir. Bu çalışmadan çıkarılan sonuçlar, okurların sınırlı kapasitelerini aşmayacak şekilde açıklayıcı notların elektronik metinlere eklenmesi konusunda öğretim tasarımcılarına faydalı bilgiler sunmaktadır.

CURRICULUM VITAE

NAME: Burcu Varol

DEGREES AWARDED

PhD in English Language Education, 2017, Boğaziçi University

MA in English Language Education, 2010, Boğaziçi University

BA in English Language Teaching, 2004, Middle East Technical University

AREAS OF SPECIAL INTEREST

Reading in L2, working memory in L2 reading, cognitive processes in language learning, technology-enhanced learning

PROFESSIONAL EXPERIENCE

Vice Head of the Department of Foreign Language Education, Yıldız Technical University, 2014 - present

Instructor, Department of Foreign Language Education, Yıldız Technical University, 2011 - present

Material Office Coordinator, Department of Modern Languages, Yıldız Technical University, 2010-2011

Instructor, Department of Modern Languages, Yıldız Technical University, 2007-2011

English Teacher, Münir Nurettin Selçuk Elementary School, Istanbul, 2005-2007

Instructor, Department of English Language and Literature, Istanbul Kültür University, 2004-2005

AWARDS AND HONORS

High Honor List, Middle East Technical University, 2003-2004

GRANTS

Boğaziçi University Research Fund grant, 2014-2017, Project No. 10901 (as PhD student)

Middle East Technical University, 2002-2004

PUBLICATIONS

Journal Articles

Kavanoz, S., Yüksel, H. G., & Varol, B. (2017). Evolvement of pre-service language teachers' beliefs through teacher education. *International Journal of Progressive Education, 13*(1), 119-135.

Varol, B. (2015). Transfer effects in compliment responses of EFL learners. *International Journal of Arts and Sciences, 8*(6), 513-522.

Book Chapters

Varol, B., & Akcan, S. (2012). EFL teachers' beliefs and practices concerning reading strategy instruction. In Bayyurt, Y. and Çetinkaya, Y. (Eds.), *Research perspectives on teaching and learning English in Turkey: Policies and practices* (235-250). Hamburg: Peter Lang.

Conference Proceedings

Varol, B., & Erçetin, G. (2016). Effects of working memory and gloss type on L2 text comprehension and incidental vocabulary learning in computer-based reading. *Procedia - Social and Behavioral Sciences, 232,* 759–768.

Varol, B., & Sinem, Y. (2010). Similarities and differences between female and male learners: Inside and outside class autonomous language learning activities. *Procedia-Social and Behavioral Sciences, 3*, 237-244.

ACKNOWLEDGMENTS

This dissertation was the result of a long and cherished journey, and I would like to express my deepest gratitude to my advisor, Assoc. Prof. Gülcan Erçetin for being with me through her constant care in every phase of it. She was truly present whenever I needed guidance and provided meticulous and priceless feedback, always on time and without which this study would not have been completed. Particularly, apart from her continual support from the beginning until the very end, she was my first and most valuable resource during the instrumentation and data analysis processes. As a conscientious and disciplined researcher, she inspired me not only in the completion of this dissertation but also in defining my path as an academician. I would also like to express my deepest thanks to the dissertation committee members, Assist. Prof. Senem Yıldız and Assist. Prof. Günizi Kartal as well as the jury members, Assoc. Prof. H. Gülru Yüksel and Assist. Prof. Kadir Kozan for their constructive feedback and enlightening comments.

I would like to thank all of my instructors at the Department of Foreign Language Education at Boğaziçi University, especially Prof. Dr. Yasemin Bayyurt and Assoc. Prof. Sumru Akcan for sharing their invaluable resources and time with me and for supporting me from the day I was accepted into the graduate program. I would also like to extend my thanks to Assist. Prof. Nalan Babür for sharing a version of a Digits Backward task with me and providing me with detailed instructions on how to conduct it.

I owe special thanks to my colleagues, Kıymet Merve Celen, Selahattin Yılmaz (a former colleague), Işıl Boy Ergül, and Gülümser Efeoğlu who helped me with their illuminating suggestions during the instrumentation process. I am also very

grateful to Assist. Prof. Suzan Kavanoz for her academic support and Ferda İlerten for encouraging her students to participate in this study. My deepest thanks go to my students at the Department of Foreign Language Education at Yıldız Technical University for taking part in this study. I wholeheartedly thank my dear friend, Elif Kemaloğlu-Er, who shared her precious comments with me during not only the dissertation writing but also the whole doctoral process.

I would like to express my immense gratitude to my family for their unconditional love and support in writing this dissertation. In particular, I am indebted to my son, Sarp Batu Varol, for filling my life with joy and being the most beautiful part of my life. I am also truly grateful to all of my friends who were with me at all times and sincerely supported me throughout the doctoral process.

Lastly, I would like to acknowledge that this dissertation could not have been completed without the financial support from the Boğaziçi University Scientific Research Project Fund (Grant No. 10901), and I express my deep gratitude to Boğaziçi University.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION 1

1.1 Background of the study 1

1.2 Purpose of the study 6

1.3 Definition of key words 8

CHAPTER 2: LITERATURE REVIEW 11

2.1 Theories and models of reading 11

2.2 Incidental vocabulary development through reading 37

2.3 Electronic reading 45

2.4 Conclusion 74

CHAPTER 3: METHODOLOGY 76

3.1 Research questions and hypotheses 76

3.2 Participants 80

3.3 Reading text 82

3.4 Piloting 83

3.5 Data collection instruments 84

3.6 Covariate 96

3.7 Semi-structured interviews 97

3.8 Procedures 98

3.9 Data analysis 99

CHAPTER 4: RESULTS 101

4.1 Effects of WM capacity, gloss type and gloss position on text recall and comprehension 101

4.2 Effects of WM capacity, gloss type, gloss position and time on incidental vocabulary learning 106

4.3 Effects of WM capacity, gloss type and gloss position on the perceived cognitive load and the recorded annotation use 117
4.4 Qualitative results 120
4.5 General findings 124
CHAPTER 5: DISCUSSION 127
5.1 Annotation use and reading 128
5.2 Annotation use and vocabulary learning 135
5.3 Annotation use and cognitive load 144
CHAPTER 6: CONCLUSION 148
6.1 Main findings 148
6.2 Pedagogical implications 152
6.3 Limitations and future research directions 155
APPENDIX A: PRIOR KNOWLEDGE QUESTIONNAIRE 157
APPENDIX B: THE READING TEXT 158
APPENDIX C: LIST OF ANNOTATED WORDS 163
APPENDIX D: TOPIC-LEVEL ANNOTATIONS 164
APPENDIX E: SCORING PROCEDURE FOR RECALL PROTOCOLS 167
APPENDIX F: COMPREHENSION TEST-KEY 169
APPENDIX G: VOCABULARY PRODUCTION TASK 172
APPENDIX H: VOCABULARY RECOGNITION TASK - KEY 174
APPENDIX I: SUBJECTIVE COGNITIVE LOAD SCALES 175
APPENDIX J: TOPIC INTEREST QUESTIONNAIRE 176
APPENDIX K: INTERVIEW QUESTIONS 177
REFERENCES 178

LIST OF TABLES

Table 1. Number of Participants in Each Cell of Analysis 82

Table 2. Descriptive Statistics for the Topic Interest Scale across the Groups..... 97

Table 3. Descriptive Statistics for the Free Recall Task 101

Table 4. ANOVA Summary of Text Recall Scores 102

Table 5. Descriptive Statistics for the Comprehension Test 104

Table 6. ANOVA Summary of Comprehension Test Scores 104

Table 7. Descriptive Statistics for Post and Delayed Matching Test 106

Table 8. ANOVA Summary of Vocabulary Recognition Scores 107

Table 9. Descriptive Statistics for Post-test and Delayed Production Test 110

Table 10. ANOVA Summary of Vocabulary Production Scores 111

Table 11. Descriptive Statistics of the Perceived Cognitive Load 117

Table 12. Descriptive Statistics of the Access to Glosses and Reading Time 118

Table 13. ANOVA Summary of Frequency of Access to Glosses 119

Table 14. ANOVA Summary of Total Reading Time 119

Table 15. Summary of Research Hypotheses, Main Findings, and Conclusions ... 125

LIST OF FIGURES

Figure 1. A model of reading (Khalifa & Weir, 2009) 14

Figure 2. Navigation map in the hypertext 83

Figure 3. Lexical pop-up window annotation condition 86

Figure 4. Topic-level pop-up window annotation condition 87

Figure 5. Lexical separate window annotation condition 88

Figure 6. Lexical annotation box viewed on a separate window 88

Figure 7. Topic-level separate window annotation condition 89

Figure 8. Topic-level annotation viewed on a separate window 89

Figure 9. A screenshot from the comprehension test 93

Figure 10. Interaction between gloss type and gloss position in terms of text recall 103

Figure 11. Interaction between WM capacity and gloss position in terms of comprehension 105

Figure 12. Interaction between gloss type and gloss position in terms of vocabulary recognition 108

Figure 13. Interaction between gloss type and time in terms of vocabulary recognition 109

Figure 14. Interaction between time and gloss type in terms of vocabulary production 112

Figure 15. Interaction between time and gloss type for pop-up window condition 113

Figure 16. Interaction between time and gloss type for separate window condition 114

Figure 17. Interaction between time and WM capacity for pop-up window condition 115

Figure 18. Interaction between time and WM capacity for separate window condition 116

ABBREVIATIONS

AWL: Academic Word List

CI: Construction-Integration

CL: Cognitive Load

CLT: Cognitive Load Theory

CTML: Cognitive Theory of Multimedia Learning

ECL: Extraneous Cognitive Load

EFL: English as a Foreign Language

ELT: English Language Teaching

ESL: English as a Second Language

GCL: Germane Cognitive Load

GP: Gloss Position

GT: Gloss Type

ICL: Intrinsic Cognitive Load

L: Lexical

L-P: Lexical Pop-up

L-S: Lexical Separate

LTM: Long Term Memory

MC: Multiple Choice

ME: Mental Effort

MM: Mental Model

QPT: Quick Placement Test

RST: Reading Span Task

STM: Short Term Memory

SW: Separate Window

T: Topic-level

T-P: Topic-level Pop-up

T-S: Topic-level Separate

WM capacity: Working Memory capacity

CHAPTER 1

INTRODUCTION

1.1 Background of the study

Rapid developments in technology and its widespread use in every sphere of life have had a substantial effect on the way learning and teaching takes place. New generations having an organic connection with technology can reach a plethora of resources through multiple means to pursue lifelong learning wherever they want. Learners can access information to accomplish tasks anytime and anywhere they want by virtue of a global network and computers. L2 learners, likewise, can reach a wide range of resources both in formal education settings and for their individual learning to improve their language skills. Electronic texts, as one of these merits, have become extensively available and turned reading into a more interactive and non-linear activity (Erçetin, 2003). Reading electronic texts compared to print reading provides flexible and multimodal reading environments through hyperlinks[1]. Readers can navigate in the text in the order they want, access the definitions of unknown words, get topic-related information, view the visual content, or listen to the audio integrated at their own pace. Instructional designers make use of a variety of tools to assist the meaning making process of L2 readers. To this end, annotations or glosses[2], as one these means, have been widely used in complementing reading comprehension (e.g., Ariew & Erçetin, 2004; Lomicka, 1998; Şakar & Erçetin, 2005) and incidental learning of vocabulary (e.g., Cheng & Good, 2009; Chun & Plass, 1996; de Ridder, 2003; Yoshii, 2006). Nation (2001) gives a detailed account of the

[1] Hyperlinks are parts in a document which are directly linked to another part of the same document or to a completely different resource ("Hyperlink", 2016).

[2] Annotations and glosses are used interchangeably in this study.

benefits of glossing. To start with, it provides an authentic reading experience as it caters to the readability of ungraded or unabridged texts for low proficiency readers. Next, it assists vocabulary learning and comprehension through supplying readers with correct information as to the meanings of unfamiliar words which would otherwise be interpreted incorrectly from the context. Third, it does not interrupt the flow of reading as much as dictionary-look-up behavior does, and is time-saving in this sense. Fourth, glossing takes attention to the form and meaning of the chosen word and thereby improves word recognition. Last, the process of looking up the glossed word and turning back to the text facilitates the retention of the word.

According to Just and Carpenter (1980) "reading can be construed as the coordinated execution of a number of processing stages such as word encoding, lexical access, assigning semantic roles, and relating the information in a given sentence to previous sentences and previous knowledge" (p. 331). It is a dynamic process where reader-related variables, like prior knowledge, language proficiency, motivation, interest, or memory capacity interact with text-related variables such as text complexity, length, or structure in order to build a mental representation or form comprehension. It is characterized by not only lower-order processing, such as distinguishing features, decoding letters, parsing sentences and moving onto further discourse, but also the higher level processing that affects lower level processing. Research has established that different levels of processing operate simultaneously and interactively which integrate both bottom-up and top-down processing of texts (Khalifa & Weir, 2011). Indeed, comprehension, as a complex cognitive functioning, is inevitably related to memory that shapes the construction of a mental model in Kintsch's (1988) construction-integration (CI) framework. The CI model which informs the current research is a cognitive architecture (Kintsch & Welsch, 1991) for

the stages of comprehension, in that it clarifies the mental operations taking place during the construction and integration phases in an exclusively symbolic way. Within this model, first the micro- and macro-propositions extracted from the text that overlay the linguistic representations are constrained. At the next level, textual information is incorporated into schemas in the light of the general knowledge (e.g., the knowledge of words, syntax, the world, and genres). In essence, the model is built on the processes of a) construction that is mainly forming a text base upon the linguistic properties of the text and the reader's knowledge base and b) integration that turns these textual knowledge bases into a coherent structure (Kinstch, 1988). For representing knowledge base, the researcher proposed an associative net which has a minimally set structure rather than semantic nets, schemata, frames, or scripts which were claimed to be too inflexible to respond to the always-changing contextual demands. Knowledge is represented in a propositional network determined by the strength of associations (i.e., positive connections among the lexical nodes). The construction of text base in the CI model is realized as follows: (a) the notions or propositions directly compatible with the linguistic input are formed, (b) the newly formed propositions are refined in the light of a small sample of closest propositions chosen from the general knowledge net, (c) other possible inferences are generated for some of the propositions, and (d) a connection strength value is allocated to all these inferences. In the end, a propositional text base that is incoherent or even contradictory is constructed. This incomplete and inconsistent network will be subjected to integration with reference to the discourse context to rule out any implausible associations represented in the linguistic base. Whereas the formation of a text base is sufficient enough to simply recall or reproduce the text, an integration of it into reader's domain-related knowledge is necessary for a deeper

understanding. The latter process is characterized as the situation model and oversees the processes of inferencing and mental model building by relating the textual information with the reader's prior knowledge as well as reorganizing and restructuring the text based on the inferences generated (McNamara, Kintsch, Songer & Kinstch, 1996).

Since comprehension requires processing symbols that are to be conceived and produced, working memory (WM) has a pivotal role in especially language comprehension (Leeser, 2007). It is claimed to affect comprehension to varying degrees that would differ across individuals. At the lexical stage, how the storage component of WM operates is more straightforward, that is the comprehender must activate the earlier representations of words or clauses to be able to link them to the new coming words or clauses. The demands the storage component constraints at other levels of comprehension are more complex though. During this stage, the comprehender has to store textual propositions from different parts of the text, the situation model that the text is referring to, the topic of the text, and the various representations of the proposition that is currently processed at the same time. Hence, language comprehension is proposed to be a good example of multilevel information processing integrating the storage of partial and eventual products at the same time (Just & Carpenter, 1992).

Cognitive load theory (CLT: Paas & van Merrienboer, 1994; Sweller, 1994) is built on the limited WM capacity of the human cognitive architecture and independent processing of visual and verbal information assumptions (Sweller, Ayres & Kalyuga, 2011). As WM is responsible for not only short-term storage but also processing, comparing, and contrasting new information and is limited in terms of the number of elements to be processed at a certain time (Cowan, 2014), schemata

construction (organization and storage of elements into the long term memory) can help reduce WM load (Kirschner, 2002). In the CLT literature, it has been documented that learning can be facilitated through presenting multiple information elements as one, automatizing the rules, and delivering the instructional material in more than one modality. Kirschner (2002), citing Paas and van Merriënboer (1994), indicates that causal as well as assessment factors are influential on cognitive load. Causal factors can be related to a) the readers, like their cognitive abilities, b) the task, like task complexity, c) the environment, like noise, and d) their interaction with each other. Assessment factors involve mental load, mental effort, and performance. Mental load is associated with the task and environmental demands. Mental effort is related to the allocation of cognitive capacities to the task. Performance incorporates mental load, mental effort and the causal factors. The measurement of these assessment factors provides information about the levels of CL. There are three types of cognitive load (CL) that affect WM processing: intrinsic load, extraneous load and germane load. Intrinsic load is the demand induced by the inherent properties of the instructional material. Extraneous load is the proportion of load related to the design of the instructional material. Germane load is the desired load associated with schema construction and automation (Kirschner, 2002).

Reading electronic texts containing hyperlinks, nodes, or navigation maps may put even more constrains on this limited capacity system. Smith and Weiss (1988) define hypertext as an information system comprised of nodes that connect interrelated documents with the help of related links. As a case in point, hypermedia annotations supply such links in various forms, ranging from definitional glosses to video animations. Yet, providing additional support to the texts may induce undue extraneous CL by occupying limited-capacity cognitive resources. Although it

enhances learning new words from the text, it could, at the same time, hinder the integration of textual information if they are placed outside the text. Readers need to quit the text to reach the gloss, retain them in memory for a short time, and return to the text to apply the word meaning to the passage. This process of leaving the text to reach the gloss and turning back to the text can create what Chandler and Sweller (1991) call *split-attention effect*. Still, ramifications of this source of cognitive overload change according to the element interactivity that the glosses possess. If the number of elements to be processed simultaneously is small (low intrinsic load), then the extraneous load stemming from disparate positioning of glosses may not matter. Quite the contrary, if the content of glosses are high in element interactivity, cognitive overload may set in resulting from mentally integrating separate sources of information. In the end, while the availability of hypermedia annotations promises enhanced learning by relating content to individual experience, the adoption of such aids also jeopardizes L2 comprehension if designed recklessly. As such, the role of WM in the personalized and autonomous meaning making process of electronic reading needs to be substantiated in the light of new research.

1.2 Purpose of the study

Investigation into annotation use, as a field, proves to be promising since researchers can invest in studying reading habits of younger generations who have immediate access to a plethora of resources at their disposal. Although hypermedia glosses have been researched in a number of studies, they have been mostly compared in terms of the modality (e.g., textual, visual, or audio) of annotations (Akbulut, 2007a; Ben Salem, 2006; Sakar &Erçetin, 2005; Yeh & Wang, 2003) or restricted to definitional glosses (AbuSeileek, 2008; Chen & Yen, 2013; Cheng & Good, 2009; Yanguas,

2009) except for a few studies (Ariew & Erçetin, 2004; Garrett-Rucks, Howles & Lake, 2015) which also provided background information in the form of cultural, demographic, or historical references. Another strand of research has been interested in the effects of gloss position on text comprehension and vocabulary learning (AbuSeileek, 2008; Morrison, 2004; Yao, 2006; Türk & Erçetin, 2014) in varying places (e.g., end of the text, marginal, or bottom of the page glosses). However, to the best of our knowledge, there are no studies that investigated the combined effects of gloss type (word-level vs. topic-level) and location (integrated format vs. split format) on comprehension, vocabulary learning through reading, and on cognitive load. Apart from such external factors, individual characteristics of readers such as WM capacity are also of concern in this design to explore their roles in learning through reading and in mental effort invested. Within the existing literature, there are studies probing the effects of WM (Alptekin & Erçetin, 2010; Daneman & Hannon, 2001; Harrington & Sawyer, 1992) yet an integrated approach to study these text-related and reader-related variables in one design is needed to explore their differing roles on reading, vocabulary learning, and CL in an L2 context.

Within this context, the purpose of this study is to explore the effects of annotation content (textual/lexical vs. contextual/topic-level[3]) and the location of annotations (pop-up-window vs. separate page) as well as WM capacity to portray how they contribute to learning through reading. Besides, this investigation is expected to shed light on how they mediate readers' perceived CL and their perceptions of electronic reading.

The remaining chapters are organized in the following way: Chapter 2 presents previous studies on theories and models of reading, electronic reading,

[3] Topic-level annotations and contextual annotations are used interchangeably in this study.

annotation use, incidental vocabulary learning, the role of working memory, topic familiarity, and topic interest on reading and the CL, and the related literature on these issues. Chapter 3 provides a detailed account of the study design including research questions and hypotheses, sampling and instrumentation processes as well as data analysis methods. Chapter 4 gives an account of the quantitative findings in the form of descriptive and inferential statistics supported by qualitative findings. Chapter 5 presents a discussion on the results with regard to research hypotheses and previous research findings. Lastly, chapter 6 conveys the main findings of the study together with their implications, limitations in the research design, and suggestions for further research.

1.3 Definition of key words

(Hypermedia) Annotation/ gloss: A technique to gloss words or necessary background information in different formats (textual, visual, or auditory) to enhance learning.

Cognitive load: Constraints on WM during the accomplishment of a particular task.

Delayed vocabulary test/ delayed vocabulary post-test: The implementation of vocabulary tests after a certain time period upon reading.

Electronic text/ Hypertext/ Digital Text/ On-screen text: Any text delivered via electronic sources such as computers, tablets, mobile phones, or e-readers.

Element interactivity: The number of elements to be processed simultaneously which has the potential to increase the perceived ICL.

Extraneous cognitive load: Unwanted mental load imposed by the design of the material.

Intrinsic cognitive load: Mental load imposed by the nature of the task.

Lexical gain: Short-term learning of new words after being exposed to them during reading or listening.

Lexical/ textual/ word-level annotations/ glosses: Verbal information in the form of definitions, synonyms, or descriptions for selected words.

Pop-up window annotations/ glosses / Integrated format: In-text annotations which can be viewed upon moving the mouse or clicking on the annotated text parts.

Print text: Traditional texts printed on paper (e.g., course books, textbooks, newspapers, magazines, etc.).

Productive vocabulary knowledge: To be able to produce L2 definitions or translation equivalents of newly learned vocabulary items.

Reading comprehension: Literal understanding and inferential comprehension of the text that go beyond simple memory instantiations and encompasses the integration of reader's pre-existing knowledge.

Recognition of words: To be able to recognize the meanings of newly learned vocabulary items.

Separate-window annotations/ glosses/ Split format: Annotations which can be viewed on a window outside the text upon moving the mouse or clicking on the annotated text parts.

Text recall: Reproduction of text from memory by including as many details as possible.

Topic-level/ contextual/ extra-textual annotations/ glosses: Verbal elaborations in the form of background information, explanations, examples, or descriptions for the selected text parts.

Vocabulary post-test/ immediate vocabulary post-test: The implementation of vocabulary tests immediately after the exposure to the text.

Vocabulary retention: Long-term integration of words into mental lexicon.

Working memory capacity: In a very brief sense, the capacity of an individual to simultaneously hold, process, and manipulate information in memory.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a detailed account of the underlying theories of the current study and the related literature. It elaborates on reading processes, reading models, the association of reading with vocabulary learning, the role of working memory, and cognitive load in electronic reading. The chapter also unfolds a selection of recent research conducted on reading and vocabulary learning in electronic learning environments.

2.1 Theories and models of reading

Reading is a language skill which includes decoding and assigning meaning to linguistic information extracted from print (Urquhart & Weir, 1998). Learning to read is an emergent process which takes place in time and which paves the way toward easy and motivated practices of reading as an expert reader in the end. It develops through the interaction of a number of factors including text-related and reader-related ones. Historically, two different approaches have been classified to explain how reading comprehension takes place: process models and componential models. Within the process models, bottom-up processing is described as using linguistic knowledge to build macro units via phonological, orthographic, lexical, and syntactic processing (Cohen & Upton, 2006). Top-down theories, on the other hand, view reading as a process in which readers constantly form guesses, test their predictions and use their background experience to attach meaning to the text (Goodman, 1967). In Stanovich (1980) and Rumelhart's (2004) interactive model of information processing, however, comprehension is depicted as the convergence of

simultaneous lexical, semantic and syntactic input altogether to interpret the message. It is obvious from the model that both lower and higher-level skills are integral to reading comprehension. Lower-level skills include feature extraction, orthographic segmentation, and phonological decoding leading to automated ability to understand the meanings of lexical items during reading. Higher-level processing, on the other hand, incorporates sentence level, and paragraph level comprehension accompanied by inferential comprehension in addition to a monitor for these comprehension processes (Koda, 2005). Since reading comprehension is a cognitively demanding activity, readers must develop automaticity in lower-order processing in order to spare sufficient cognitive resources for higher-level processes. Researchers now agree that it is neither possible nor adequate to define reading only as lower-level or higher-level processing, but instead, it is a complex process which includes the elements of both (Grabe, 1991).

Almost all models of reading encompass three vital components of reading: the reader, the text, and the interaction. Likewise, Snow (2002) describes reading comprehension as the simultaneous withdrawing and constructing meaning from the text through reader's interaction and involvement with the text. Accordingly, her three component model consisting of the reader, the text, and activity demands/purposes is in interplay with the sociocultural context where comprehension takes place. The reader is involved in the process with all his/ her "cognitive abilities (attention, memory, critical analytic ability, inferencing, visualization); motivation (a purpose for reading, interest in the content, self-efficacy as a reader); knowledge (vocabulary and topic knowledge, linguistic and discourse knowledge, knowledge of comprehension strategies); and experiences" (p. xiii- xiv). The text is represented at three levels as the reader constructs the meaning. The first

of those levels is extracting the *surface code* which is basically word by word decoding of the text. The next stage is forming the *text base* which is the representation of the idea units introduced in the text. The third one takes in the reader forming a *mental model* which is mainly about processing the text for meaning representation. The last component of her model of comprehension is the *activity* comprised of the task purposes, operations of meaning construction, and the products of these operations. The distinct feature of this model is the fact that all of these elements develop within a sociocultural context. That is, learning to read is a literacy skill which differs across contexts since the amount and type of interaction with texts depends on the reader's circle apart from the classroom. Despite the resemblance to Kintsch's (1998) *situation model* for comprehension, the integration of the activity component into the model and the emphasis on the sociocultural context where reading takes place were the departure points of this recent view from its antecedent.

An even more up-to-date model of reading which aroused from the integration of various types of reading and cognitive processes which are involved in reading was Khalifa and Weir's (2009) comprehensive model as depicted in Figure 1. In the model, all the components are related to one another wherein the importance of the goal setter on the left-hand column lies in the critical role it plays for determining the processes that will take place during reading upon setting the purposes for reading. Local level represents comprehension of sentences or clauses at the micro-level of the hierarchical structure. Global level includes comprehension of macro-propositions – that is, the propositions beyond sentence-level – how macro-propositions are connected, and the effects of micro-propositions on the construction of global understanding. Careful reading refers to the slow, incremental, and linear

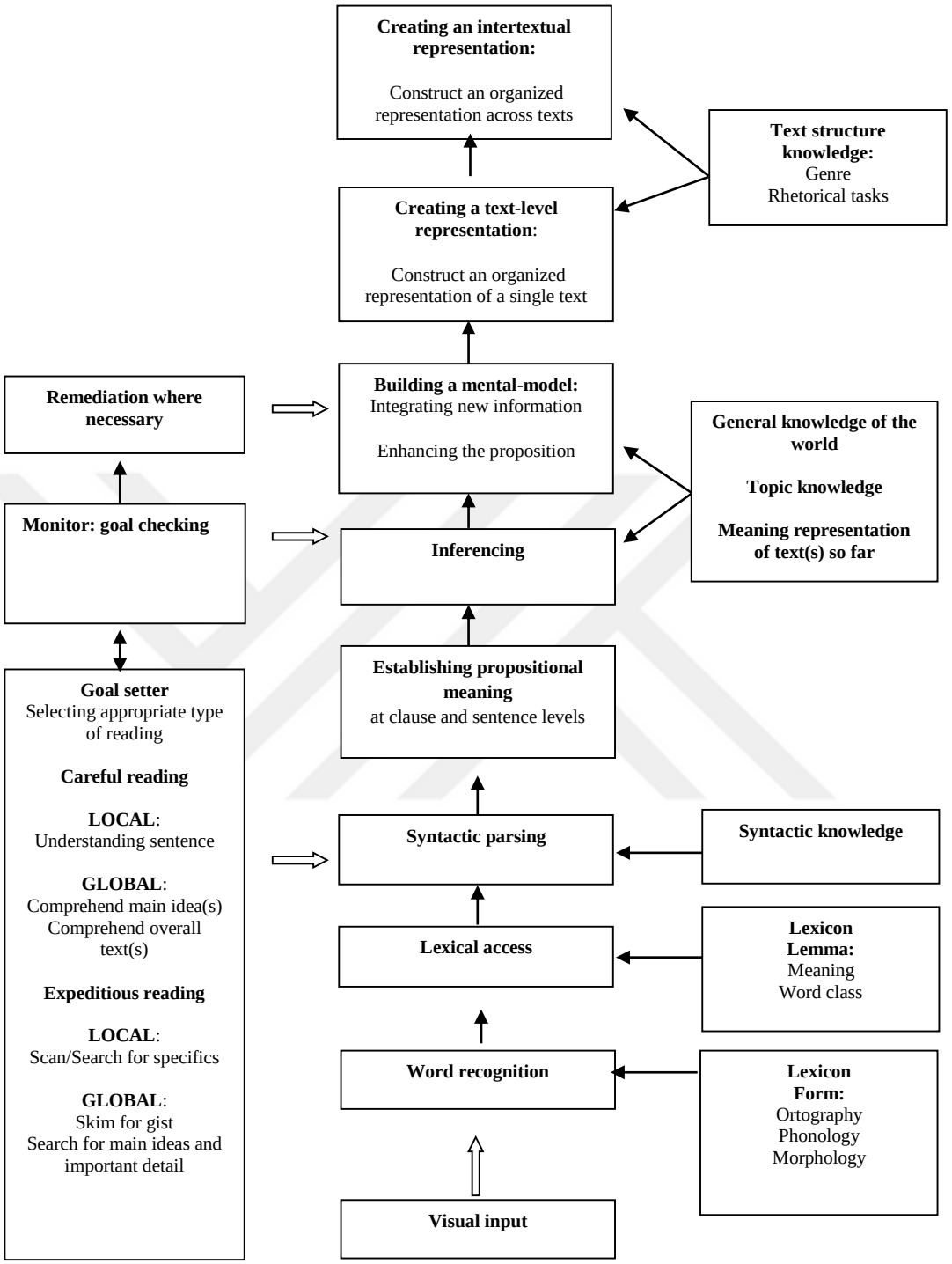


Figure 1. A model of reading (Khalifa & Weir, 2009)

processing to deduce the exact meaning from the text. It can take place both at the local and the global level. Expeditious reading is reading quickly, selectively, and efficiently to locate specific information in a text. It can take the form of skimming, scanning, and search reading. Within the central processing core in the middle, word recognition – matching the written form of a word with its orthographic representation in the mental lexicon– appears as the first component. The next element, lexical access, refers to the retrieval of all information related to word, including its phonological and orthographic representations, morphological and semantic information from the mental lexicon.

In the model, fluent syntactic parsing requires the knowledge of syntax which involves word-order, morphology, and structural constituents, such as prepositions, determiners, etc. to be able to segment the text into word units, phrases, clauses or sentences. Establishing propositional meaning at the clause or sentence level includes extracting literal meaning of the text without any extra-textual interpretation. Inferencing refers to extending the literal meaning by understanding the underlying meaning to build bridges among the implicitly stated propositions. Building a mental model (MM) occurs at the level of the integration of newly extracted information into the developing mental representation of the text. In this sense, it is continuous and subject to changes and reshaping based on the incoming information. Creating a text-level representation involves building the discourse-level structure of the whole text by recognizing its hierarchical structure and key elements for the construction of core meaning. Creating an intertextual representation, in a way synonymous to 'discourse synthesis' of Stromso and Braten (2002; as cited in Khalifa & Weir, 2009), refers to eliciting, combining, and organizing data from multiple resources via higher-order relational links to build a coherent and concise structure. Lastly, in

Khalifa and Weir's (2009) model, monitoring is omnipresent in the sense that depending on the type and purpose of reading, it can take different forms and even change the whole process.

2.1.1 The construction-integration model

Khalifa and Weir's (2009) model is an overarching account of reading representing the many interacting elements influential in comprehension. Yet, the present investigation adheres to an older but similar model because of its explanatory power in depicting the underlying processes of meaning making through a three-stage integration procedure. It is now widely accepted that for comprehension to take place a coherent situation model including prior knowledge integrated into the text representation needs to be developed by the reader. According to the construction-integration model of Kintsch (1988), there are several levels of MM building which are *surface representations*, *propositional text-base*, and *situation model*. It is assumed that these multiple processes underlying comprehension can take place both in parallel and consecutively. The reader encodes the explicit visual features of the written message and starts with letters, words, and sentences. Then, this encoded message is incorporated into micro and macro-propositions of the text to construct a propositional text base. This textual representation contains a network of information generated from the explicitly stated ideas in the text. Lastly, the propositional text-base is connected to reader's prior knowledge to form the situation model which is an amalgamation of readers' existing knowledge and the implied relationships among textual constituents. A situation model is defined as "an integrated structure of episodic information, collecting previous episodic information about some situation as well as instantiated general information from semantic memory" (van

Dijk & Kintsch, 1983, p.344). In a way, it requires a domain-specific schema as well as general background knowledge. The researchers proposed a framework for a process model of discourse comprehension which includes three major surround systems. The first of these, *sensory register*, holds the incoming perceptual information for a short time and makes it ready for the central processor. Secondly, the *long-term memory* of the reader that also includes goals, purposes, wishes, interests and emotions of the comprehender is incorporated into the meaning-making process. The last class is the memory representation that is under construction, namely the *episodic text memory* together with the *situation model*. In the center of the model is the *central processor* where all cognitive processes take place. All cognitive operations are undertaken in this unit. For instance, so as to modify any component in any of the other memory systems, that component should be brought into the central processor. Later, Kintsch (1998) elaborates on the model to which the present study subscribes as such:

> A process model of text comprehension attempts to describe the step-by-step processes by which written or spoken language is transformed into a mental representation in the reader's or listener's mind. The construction-integration (CI) model assumes that this process involves two phases: a construction phase, in which an approximate but incoherent mental model is constructed locally from the textual input and the comprehender's goals and knowledge, and an integration phase that is essentially a constrain satisfaction process that rejects inappropriate local constructions in favor of those that fit together into a coherent whole. The construction rules in this model can be relatively simple and robust because they have to take into account only the local context. The global context becomes important only in the integration phase, when the tentative, incoherent network that has been formed by the context-free construction rules settles into a stable state. (pp. 119)

According to the author, the model is different from a purely top-down, schema-controlled model in that the prerequisites of the latter would be more specific, complex, and context-sensitive construction rules. As opposed to the sophisticated and ever-changing nature of real-world, such construction rules may be rigid and

unreliable because ordinary comprehension processes do not always follow such a predetermined structure, rather they make use of simpler and more vigorous local rules in the company of a global integration process. He describes the mental models of texts as being comprised of two components: the text being read and the readers with all their background knowledge, experiences, and needs. That's why, a distinction needs to be made between a textbase that stands as a rough approximation of the text, and a situation model that incorporates reader's interpretations. In the end, the researcher states that these two components can be scaled differently by different readers at different times.

2.1.2 Reading in L2

Comprehension ability gains even more importance in a second language context as linguistic constraints such as limited vocabulary or grammar knowledge interfere with meaning. L1 literacy skills which have developed prior to the exposure to L2 reading as well as overall proficiency in L2 may pose different problems (Koda, 2005). Unlike oral communication, written language is deprived of extra linguistic information to facilitate processing of information which turns L2 reading into a more demanding activity. L1 reading competence, then, plays a crucial role in the formidable process of reading in L2. It follows that attaining a certain threshold level is also a cursor for successful comprehension as automatic word-recognition and sentence processing need to be developed in the L2.

Bernhardt (1991; as cited in Koda, 2005), one of the first writers who focused on the distinction between L1 and L2 reading puts forward the following considerations while approaching second language reading: (a) L1 literacy levels of readers, (b) previous experiences in a second language, (c) the proximity between L1

and L2, (d) shared cultural knowledge, and (e) the script types of two languages. Koda conceptualizes the characteristics of L2 reader in a similar fashion in terms of "…prior literacy experience, limited linguistic sophistication, and dual-language involvement…" (p.8). She reiterated the importance of cross-linguistic studies to reveal how L1 and L2 interplay and how this interface impacts the development of L2 reading. *Linguistic knowledge* and *language processing skills* seemingly related but actually distinct properties of reading competency contribute to comprehension process at varying degrees. While the former is a prerequisite of effective comprehension, the latter displays cross-linguistic variations which, in turn, needs to be developed in L2 reading, rather than taken for granted.

In terms of approaches to L2 reading, Koda (2005) differentiates between two major positions: early transfer paradigms and language-specific processing perspectives. The basic premise of the former (universal) framework is that there is a mutual relationship between L1 and L2 reading competence, and that there exist some habits transferred from L1 reading that facilitate or preclude L2 comprehension. With the evidence from cross-linguistic studies conferring that cognitive processing of sentence comprehension and production are language specific, cross-linguistic perspectives of language processing (such as word decoding, syntactic parsing, or discourse processing) peculiar to L1 or L2 become prevalent. The underlying principles of the language-specific view is that cognitive processes inherent in L2 reading vary according to the L1 background of the reader. To this end, the author urges for in-depth investigations into the linguistic, cognitive, and metacognitive competencies required for reading in different languages as well as the illumination of earlier literacy experiences.

Grabe and Stoller (2002) also indicate qualitative differences between L1 and L2 reading which can be listed as linguistic knowledge, transfer effects, and L2-related procedures such as using bilingual or translate dictionaries. They further categorize those variations under three subtitles: linguistic and processing differences, individual and experiential differences, and institutional differences. Under the first category lies the fact that the linguistic base of most L2 readers does not include an accumulated spoken language experience. Apart from a lack of required grammatical knowledge, they are also devoid of the discourse knowledge necessary for effective reading especially in academic English environments. Other considerations that are related to linguistic and processing features of L2 reading are described as:

- greater metalinguistic and metacognitive awareness in L2 settings
- differing amounts of exposure to L2 reading
- varying linguistic differences across any two languages
- varying L2 proficiencies as a foundation for L2 reading
- varying language transfer influences
- interacting influence of working with two languages. (p.42)

The second dimension, individual and experiential differences, refers to reader-related variables including differing levels of L1 literacy experience, differing levels of motivation toward L2 reading, differing attitudes toward different text types, and differing experiences in L2 reading strategy training. Regarding the socio-cultural and institutional differences that affect L1 and L2 reading development, different cultural or social backgrounds of L2 readers, exposure to different patterns of discourse organization, and different expectations prevalent in various L2 institutions are given as points to be taken into account (Grabe & Stoller, 2002). It follows that comprehension of expository texts is an artifact of the interaction between the individual characteristics of the reader and text features (McNamara et al., 1996; O'Reilly & McNamara, 2007).

2.1.3 The role of prior knowledge and topic interest in reading

According to Grabe (2009), reading comprehension, in simple terms, is the combination of textual input, cognitive processes, and readers' background knowledge. However, the role of background knowledge in comprehension process is rather complex, and there are multiple other factors that need to be acknowledged beforehand. Firstly, the umbrella term background knowledge should be handled under subcategories, such as 'general world knowledge, cultural knowledge, topical knowledge, and specialist expertise knowledge' (p.74). Next, attitudes, goals, motivation or language proficiency all affect the working of background knowledge in comprehension process to different degrees. Thus, the role of background knowledge would change based on the purposes for reading, the individual's reading skills, and the degree of background knowledge available. The importance of contextual/prior knowledge also lies in the role it plays in the construction of both the text-base of reading and the situation model of textual comprehension. While building the situation model, context helps integrate newly processed proposition into the developing model of the text via the inferences made between the existing parts of the text and the newly processed proposition. The reading context also determines reader attitudes, changes in goal-setting, comprehension checking, and judgments about author intentions.

Sweller (1994) puts forward that "as familiarity with a domain is gained, the need to devote attention to the required processes is reduced. Gradually, they become more automated, freeing cognitive resources for other activities" (p. 298). He sees this procedure as the second major process of learning which also affects schemas following schema construction. Across various disciplines including education, it has been shown that high levels of prior knowledge are conducive to effective

distribution of attentional resources, increased analytical abilities, and better synthesis of textual information into a situation model (Pulido, 2007). Knowledge is also heuristic during reading according to the Connectionist Principles of comprehension and constructed upon employing local-level processing of textual input (the formation of text base) in combination with the reader's global knowledge. Therefore, the presence or provision of background knowledge play an indispensable role in reading comprehension process in terms of creating an MM conforming the original text. A more comprehensive explanation of how knowledge is employed in comprehension comes from Nassaji (2007) who maintains that the processes of schemata construction which contain already existing knowledge stored in the mind, and comprehension which refers to placing new information coming from text into prior knowledge are predictive and reader-driven. In addition to schema-based models, in the construction-integration model of text comprehension of Kintsch (1998), background knowledge is assumed to be represented as associative networks of propositions which are activated at the time of reading via bottom-up interaction with text-based propositions. Koda (2007) also underscores that "successful comprehension is achieved through the integrative interaction of extracted text information and a reader's prior knowledge (p.4). In essence, then, prior knowledge is integral in every phase of meaning construction through reading, and there is a reciprocal mutual relationship between comprehension based on schemata and forming schema based on reading.

Prior research in L2 reading comprehension has depicted that the amount of recall (Alderson, 2000; Alderson & Urquhart, 1988; Carrell, 1987; Hudson, 1982) and comprehension (Bügel & Buunk, 1996; Carrell & Wise, 1998; Johnson, 1982; Lee, 1986) is increased by activating or providing necessary content-related

knowledge. Furthermore, it is documented that culturally familiar stories are read more efficiently and lead to better comprehension in relation to culturally unfamiliar stories (Alptekin & Ercetin, 2011; Hudson, 2007). Leeser (2007) by adding the element of working memory capacity in his design investigated the relationship between topic familiarity, reading comprehension and processing of future tense morphology. His findings also suggest a significant role of topic familiarity on all dependent variables. Topic familiarity has also been shown to be effective in vocabulary learning or in inferring correct meaning during reading as well as being facilitative in text comprehension. In her study looking for prior knowledge effects on comprehension and lexical gain, Pulido (2000, 2003) found out that there emerged greater vocabulary gains while reading brief narratives on familiar topics initially. In another similar study, Pulido (2007) examined the recognition of nonsense target words by intermediate Spanish learners on two conditions; in culturally familiar texts, and culturally unfamiliar texts. Upon focusing on the correlation between topic familiarity, text comprehension and form recognition, data analyses yielded results that depicted greater recognition of words in culturally familiar texts.

Cognitive abilities and prior knowledge have been extensively investigated in relation to their effects on reading comprehension. However, a number of other factors may influence the comprehension process including motivation, interest, engagement with the text, learning style, and task types. Interest, among the others, has been claimed to induce better quality performance since readers would have more resources left free as a result of the increased attention enhancing automatic processing (Hidi, 1990; Krapp, 1999). Krapp (2002) conceptualizes interest as a rather *relational construct* that includes a long-lasting relationship between the

person and the object. It is directed towards the content of the learning and always related to a specific object, subject matter or construct. Apart from influencing a person's cognitive and affective performance, interest has been defined not only as personal inclination but also as a psychological state which increases attention, cognitive and affective functioning, and trial (Ainley, Hidi & Berndorff, 2002). It has been contended that since heightened attention toward the object of interest is intentional, interest activates self-regulatory mechanisms such as goal-setting, self-efficacy, and cognitive strategies which would result in task accomplishment (Magliano, Durik & Holt, 2011). On the other hand, if heightened attention toward the object of interest is automatic, this would save resources for the fulfilment of other tasks as automatic attention underlies the benefits of interest on cognitive functioning (Hidi, 1995).

Hidi (1990) asserts that interest determines selecting and processing certain information over the others. She distinguishes two types of interest: *individual* and *situational interest*. Situational interest is a temporary and focused stimulation caused by a specific property of an object, person, or activity guided by positive emotions. Particularly, formal structural characteristics (e.g., novelty, intensity, ambiguity) and content characteristics (e.g., human activity, intensity factors, life themes) are specified as factors contributing to situational interest. Individual interest, on the other hand, is a more permanent *affective-evaluative* state towards particular objects, people, or issues accompanied by positive feelings and values (Schiefele, 2009). For instance, a reader who has individual interest on ecology and conservation would be motivated to look for other activities related and enjoy the task engagement, thus would add on to his/her existing knowledge (Ainley et al., 2002). As well as a general interest in learning, individual interest can be expressed

as domain-specific interest (e.g., interest in specific school subjects) or as activity-specific interest (e.g., sports, music, dance, etc.). General individual interest in learning which involves both acquiring new knowledge and expanding the existing knowledge has been found to be related to positive attitudes toward schooling (Ainley, 1998). Personal interest has a long lasting effect on preferences or values whereas situational interest tends to be temporary and have a short-term effect. While individual/personal interest stems from personal preferences or interests, situational interest is triggered by the environmental stimuli, such as upon reading or hearing something and can consist in some negative feelings (Hidi & Harackiewicz, 2000). In a meta-analysis of 14 studies, Schiefele (1996; as cited in Schiefele, 2012) looked for the relationship between situational interest and learning and found a correlation of 0.33. The author also notes that this relationship is not affected by factors such as text length, genre, methods of testing (e.g., recognition vs. recall tasks), age, intelligence, reading ability, importance of text, or units of analysis (e.g., sentence vs. passage). Magliano et al. (2011), on the other hand, propose that if interest turns comprehension into automatic processing, then the correlational relationship between interest and reading comprehension would be visible only for low WM capacity readers since WM can be a confounding variable for the high capacity group.

With regard to learning, topic interest has been also added as the third type of interest which can also be construed as the interaction and extension of both types (Ainley et al., 2002). It is defined as "interest elicited by a word or paragraph that presents the reader with a topic" (p. 546) and in this sense, taken as a type of situational interest (Hidi & McLaren, 1990). According to Schiefele (1996), however, topic interest is a form of individual interest standing in isolation from text-driven or situational interest since text-based interest arises from text-specific

features. Topic interest is, instead, related to the feelings associated with the topic and values attached to the topic. For the relationship between topic interest and learning from texts, in the meta-analysis mentioned above, Schiefele (1996; as cited in Schiefele, 2012) reported a correlation coefficient of 0.27. Again, this relationship was found to be independent of text length, text genre, text difficulty, age, reading ability, and methods of testing. Within the same analysis, prior knowledge was also observed to act autonomously in determining learning from texts as there found to be small to medium relationship between topic interest and prior knowledge while the effect of prior knowledge on learning was detected to be greater than that of topic interest on learning.

In addition to topic interest, Schiefele and Krapp (1996) looked for the effects of intelligence and reading process variables, such as attention, arousal, note-taking, and underlining on text recall by taking gender constant. Before reading the text, Psychology of Communication, eighty male college students were given prior knowledge, topic interest, and general intelligence tests. Topic interest was found to be significantly related to text recall although neither cognitive abilities nor prior knowledge produced similar results. This relationship between topic interest and text learning was also independent of prior knowledge or cognitive capabilities. Furthermore, topic interest was detected to be a stronger predictor of success at deeper-level text learning such as recall of main ideas, deep comprehension, and coherence of recall of main ideas than at surface-level text learning. To depict the multi-faceted nature of reading comprehension affected by various features, Bray and Barron (2004) took gender and verbal ability factors into account and looked for the relationship between interest in texts and performance in comprehension tests in a very large-scale study. With the participation of 19,735 Grades 4 to 8 students in

reading a total of 98 different texts, the researchers gauged the interplay among the aforementioned variables. Although their results depicted a statistically significant relationship between interest and text learning as measured through comprehension scores, the authors concluded that this relationship cannot be construed as large "in an absolute sense" (Bray & Barron, 2004; p. 121) nor did they find the relationship between gender and reading comprehension large. Yet, they reported a relationship between gender and interest, being that interest predicted performance on comprehension better for girls than it did for boys.

As the concluding remarks in his review paper on interest and learning, Schiefele (2012) calls for further studies to give rise to complex models to explicate the relationship between interest and text learning and to shed light on the possible mediators of this correlation.

2.1.4 The role of working memory in reading

Working memory (WM) is a multi-component system comprised of one central executive and three slave systems which are phonological loop, visuospatial sketchpad, and episodic buffer in Baddeley's (2000) framework. Basically, it refers to the cognitive systems undertaking the control, regulation, and active maintenance of information. In its simplest form, the slave systems are analogous to STM, whereas the attentional system maintains information between the storage-based systems and LTM. Within this scope, WM differs from STM whose main function is to passively keep words or digits for a short period. In contrast to the huge capacity of LTM, WM has a rather limited capacity in that adults can recall 3 to 5 chunks when presented either pairs of words or single words (Cowan, 2005).

Traditionally, there are two different subsystems of working memory; *visuospatial* WM and *verbal* WM in addition to a central executive. While the former is responsible for briefly representing, keeping and regulating information in the spatial domain, the latter is responsible for representing and processing verbally-conveyed information (Baddeley, 1986). Later, Baddeley (2000) proposed the episodic buffer which is "assumed to be a limited-capacity temporary storage system that is capable of integrating information from a variety of sources" (p.421) as the fourth component of the model. The reason why the buffer is episodic is that it keeps episodes which can be integrated across time and space. It works as a buffer as it connects different systems which have different codes. The central executive can reach the episodic buffer through conscious awareness. WM capacity is taken to be an important determiner of individual differences which influence language processing, reasoning, and learning in general (Baddeley, 2000). By ascribing to Conway et al.'s (2005) view, in this study WM is taken "as a multicomponent system responsible for active maintenance of information in the face of ongoing processing and/or distraction" (p.770). Furthermore, again following their conceptualization of domain-general executive function, WM capacity is taken to be domain-general in this study, meaning that no distinction is made between verbal WM capacity and spatial WM capacity as gauged through a backward complex digit-span task.

Reading inherently poses high demands on WM: decoding words' semantic and syntactic structures to integrate them into sentences, forming coherent relations between sentences and creating a comprehensive situation model heavily constrains working memory capacity (Alptekin & Erçetin, 2009). Just and Carpenter (1992) maintain that during reading WM keeps the newly obtained information and retrieves relevant information from LTM while processing the upcoming parts of the text at

the same time. Among the individual factors affecting information processing during reading, WM capacity plays a crucial role for it is claimed that some learners process input more effectively as they have better attentional capacity or as they carry out analytical processing within WM at a faster speed (Skehan, 1998). There is considerable evidence that WM capacity is a successful predictor of language processing, including vocabulary learning, and reading comprehension (Atkins & Baddeley, 1998; Daneman & Hannon, 2001; Friedman & Miyake, 2004). Basically, the reading skill underlies the ability to connect different ideas occurring in separate parts of the text via a laborious inference process (Daneman & Hannon, 2001). Hence, it can be taken as a basis to understand differences in text comprehension (Linck, Osthus, Koeth, & Bunting, 2013). Alptekin and Erçetin (2011) also note that cognitive operations in L2 reading are affected by a 'state-level *cognitive deficit'* (p. 236) as the required contribution from the LTM is hindered due to the processing restraints imposed upon WM. The sources of these processing demands can be linguistic (low proficiency level) or cultural (lack of culturally relevant schema). No matter what the source of difficulty is, too much dependence on text impairs comprehension as retrieving domain specific information from LTM for building situation model becomes laborious which results in "a shallow textual representation" (p. 237). It is suggested that when domain knowledge increases, the effect of WM gets more predominant leading to the *rich-get-richer* effect (Stanovich, 1986) The renowned phenomenon refers to the further development of vocabulary size, reading, verbal reasoning, and cognitive abilities of skilled readers as a result of reading more. Less skilled readers are disadvantaged, though, to the extent of missing more opportunities of learning new words via reading due to their poorer vocabulary knowledge and reading ability.

Span tasks in general target at measuring active maintenance of information during information processing and/or distraction. They work in accordance with dual-task paradigm, in that, while the participant tries to recall words, digits, etc., s/he has to process sentences for accuracy, solve mathematical operations, etc. at the same time. The reading span tasks (RST) were originally designated to tap both the processing and storage functions of working memory simultaneously (Daneman & Carpenter, 1980). The task requires the subject to recall words (one for each sentence) while at the same time reading unrelated sentences aloud as the secondary task. The sentences are given in groups including two to six sentences in total. A group is referred to as an item, and word span tasks start with the completion of an item. Remembering the sentence final words correctly is assumed to probe storage while reading the sentences aloud as the background task is believed to measure processing.

The reading span tests have been shown to be both reliable and valid for measuring the role of working memory in reading comprehension ability (Friedman & Miyake, 2004; Conway et al., 2005). In fact, most of the WM span tasks, such as operation span, reading span, and counting span, have proved to tap WM capacity, thus justifying themselves as highly reliable and valid measures. However, they are not error-free altogether in themselves. Conway et al. (2005) suggest the application of multiple measures in combination in order to validate the findings. Working specifically on RSTs, Alptekin, Erçetin and Özemir (2014) also demonstrated that the processing component of the task is language and task-dependent while the storage component is not affected by the language. They concluded that by looking at these results, the use of RST as a measure of WM "may render the outcome … in L2 research unsound" (p. 548). Another finding of that study regarding the

relationship between WM capacity and L2 reading was that L2 reading ability and the storage component of RST highly correlates. Specifically, they found out that L2 reading comprehension is related to L1 and L2 semantic resources and "to a lesser degree, to late L2 learners' level of proceduralized target language morphology and syntax" (p. 547).

In research designs, RSTs have often taken their place as a measurement of WM capacity and to assess its influence on especially reading comprehension. A pioneering work investigating the relationship between WM capacity and L2 reading skill through RST was Harrington and Sawyer's (1992) research. With the participation of advanced ESL learners, they studied the relationship between WM capacity and reading skill. They employed L2 English and L1 Japanese memory tests and a battery of L2 reading tests. Their findings depicted a strong correlation between the L2 reading span task and the performance on the Grammar and Reading sections of TOEFL. Thus, they suggested that WM capacity is an indicator of a "trade-off" between active maintenance of processing and storage functions. Chun and Payne (2004), by also following this tradition, looked at the relationship between WM and reading comprehension of L2 German readers. Through a non-word repetition and an RST, the researchers determined WM capacity of readers as high and low. While the participants were reading a German short story, they had access to the multimedia annotations of difficult words. Upon the analyses of vocabulary and comprehension tests, a relationship was observed between phonological WM and the number of look-ups. These findings implied that readers with low WM capacity use multimedia annotations to compensate for their low WM capacity. However, assessing reading performance through a multiple choice comprehension test, they were unable to report any significant effect of WM capacity on neither reading

comprehension nor vocabulary learning. Leeser (2007), on the other hand, incorporated comprehension recall protocol into his design in combination with form recognition and tense identification tasks to examine the role of individual factors (i.e., topic familiarity and WM capacity) in reading. A total of 94 learners of Spanish whose L1 was English took the computerized version of RST and were grouped as low WM, medium WM, and high WM readers based on their scores. The results revealed WM effects on text comprehension and form recognition but only with the mediation of topic familiarity.

Alptekin and Erçetin (2011) looked into a more specific dimension of L2 reading comprehension: literal understanding vs. inferential comprehension, and how WM capacity affects these separately. Another aim of the study was to detect the mediating effects of content familiarity. Sixty-two Turkish university students with advanced proficiency levels participated in their study. WM capacity was measured through a computerized RST, and content familiarity was controlled through a nativized version and an original version of a narrative. Nativization of the text was ensured by making textual and contextual adaptations reflecting the readers' own culture. After taking RST, participants were assigned either to nativized or to the original text group. Reading comprehension was assessed through a multiple-choice test comprised of literal and inferential understanding questions. Their findings added to the already existing knowledge that content familiarity improves comprehension but limited this effect only to inferential comprehension. Moreover, the results of this study delimited the relationship between reading comprehension and WM capacity only to inferential comprehension in L2 and not to literal understanding as no difference was observed between low and high span readers in their literal understanding.

Throughout correlational studies it was well evidenced that the capacity to process and store information at the same time in working memory determines the success of comprehension task (e.g., Harrington & Sawyer, 1992; Leeser, 2007). Furthermore, WM capacity was observed to be a strong determiner of different degrees of success across individuals in global scale comprehension tests prepared in multiple-choice format and in more local tests of comprehension including non-multiple choice items such as summary production (Daneman & Hannon, 2001). The researchers explain how WM capacity is related to comprehension as follows:

> According to the theory, working memory span is a good predictor of comprehension because individuals who have less capacity to simultaneously process and store verbal information in working memory are at a disadvantage when it comes to integrating successively encountered ideas in a text as they have less capacity to keep the earlier read relevant information still active in working memory. (p. 209)

One of the assumptions of the original RST was that as high capacity readers have better skills or strategies to cope with the processing demands, there would remain larger space for the storage of words. As such, it was the nature of the secondary task (i.e., a reading task) that would determine task performance based on the assumption that good readers with better skills could spare more capacity for storage (Turner & Engle, 1989). One possible conclusion to be drawn that the background processing task must be a reading task to measure the complex cognitive ability for reading comprehension. Engle (1989) brought another plausible explanation which states that it may be the larger storage capacity of high WM capacity individuals which enhances their processing capacity regardless of the secondary task type. That is, whether the background processing task is a reading or non-reading task, high WM capacity readers would outperform low WM capacity readers thanks to their "relatively stable characteristic" (p. 4). With the help of a variety of secondary tasks including sentence word task (i.e., to judge the semantic or

syntactic accuracy of the sentences while retaining the end-words in memory), digit stimuli, and operation stimuli, Turner and Engle (1989) tested the hypothesis that the relationship between the results of complex memory span tasks and the performance on reading comprehension is independent of the type of the background processing task. The first overarching finding of their study confirmed this prediction as not only word span but also digit span task determined success in reading comprehension. Additionally, it was noticed that complex span tasks correlated with reading comprehension more than simple span tasks when measured in the company of moderately difficult background tasks. Finally, the researchers cautiously maintain that "working memory may be a unitary individual characteristic, independent of the nature of the task in which the individual make use of it" (p. 150). Therefore, it is contended that process plus storage tasks (i.e., complex span tasks), either verbal or math, predict comprehension more accurately than the simple verbal span or math span tasks. The authors confer, then, that rather than the temporary storage capacity of WM, it is the combination of processing and storage capacities of WM that matters for comprehension.

Building on the related literature on the superiority of complex span tasks and secondary-task-independency of the processing component, a complex digit span task was used in the current design. Not only the individual role of WM, but also the interplay between WM capacity, and the availability topic-level information are issues that need further investigation. It is crucial to determine what the extent of an individual's WM capacity is in affecting reading comprehension and how it mediates the effects of annotation use on reading and incidental vocabulary learning. Navigating through hypermedia texts places heavy burdens on readers' working memory resources. Therefore, the availability of lexical or contextual annotations as

well as readers' complex cognitive abilities gain considerable importance in assisting them through the construction of the situation model and vocabulary acquisition.

2.1.5 Assessing reading comprehension

Generally, discrete-point items (e.g., multiple-choice) are used in the testing of reading comprehension, yet they are criticized for manipulating the interaction between the text and the reader and test performance in the end (Alderson, 2000; Bernhardt, 1991; Lee, 1986; Shohamy, 1984). Summaries or free recall tasks are seen as more integrative and non-intervening methods of assessing reading comprehension (Lee & Riley, 1990; Riley & Lee, 1996). Particularly, free recall has been supported by researchers as a "purer" method of testing comprehension which does not impose interference that would otherwise be forced by tester (Bernhardt, 1991). They are also claimed to reflect the processes taking place during information storage, organization and retrieval (Bernhardt, 1983).

The analysis of written recalls has been mostly handled quantitatively whereby text entities are analyzed as idea or pausal units (Bernhardt, 1991), and propositional units (Kintsch & van Dijk's, 1978) and the recalls produced by readers are compared to this analytical scoring scheme. Riley and Lee (1996) contend that when readers are instructed to write down whatever they remember, they could generate details as well as main ideas from the text. In fact, in their study where readers were provided with information concerning the rhetorical structure of the text, Lee and Riley (1990) revealed that task performance was affected by the knowledge of text organization. In another study, Riley and Lee (1996) utilized both quantitative and qualitative measures of summary and recall tasks. They devised a rating system which analyzed top-level idea units produced and the integration of

main ideas, supporting ideas, and details into analysis discretely. The findings depicted summary as a more appropriate measurement of global-level comprehension than free recall tasks as it yielded the generation of significantly more main ideas and a more coherent representation of text. Yet, Cohen (1994) with a detailed analysis into the processes of summary construction raised some concerns with regard to the validity of summarization technique stemming from the questions as to the reliability of raters and from the contrast between reading which is overwhelmingly bottom-up and summarizing which is overwhelmingly top-down.

More recently, Sawaki (2003) provided an in-depth analysis of summary and free recall tasks in terms of task and raters' performance. By recruiting 160 learners of Japanese, the researcher aimed at reaching quantitative and qualitative comparisons into the nature of summary and recall tasks. Summary task was conceptualized "as an activity where a test taker writes a gist of a prompt text from memory immediately after reading it" whereas free recall was conceptualized "as an activity where a test taker writes as much of the information in a prompt text as possible from memory immediately after reading it" (p.11). The analyses included data coming from summary and recall protocols as well as verbal protocols. The findings demonstrated that summaries are relatively more decent measures of global comprehension. Yet, the confirmatory factor analysis signaled that recall measured comprehension and integration of main ideas in addition to the integration of detailed information whereas summaries only gauged comprehension and integration of main ideas.

2.2 Incidental vocabulary development through reading

Traditionally vocabulary teaching has been a central concept in reading instruction for ESL learners (Bernhardt, 1991), and there is a volume of research showing that word knowledge is a predictor of reading comprehension (Bernhardt & Kamil, 1995; Nation, 2001; Nuttall, 2005) as well as general reading skills and proficiency in the language (Alderson, 2000; Grabe 2004; Perkins, Brutten, & Pohlmann 1989). Research also supports the view that learning to read in a second language is also learning new words. Grabe and Stoller (2001), for instance, report that while reading improves vocabulary development, enriching vocabulary sustains reading comprehension. This bidirectional relationship between vocabulary knowledge and reading comprehension is so strong that the correlation is reported to be around 0.90 (Stahl, 2003). That successful comprehension depends on an effective coordination of local (word-level) and global (context-level) processing entails a focus on not only discourse level elements but also word-recognition capability (Koda, 2005). Though rapid recognition of words is a good reader behavior, there is no one specific training technique to develop this ability which is assumed to evolve as an epiphenomenon of massive amounts of regular practice (Nuttall, 2005).

It is a priori that extensive reading is instrumental in vocabulary learning; however, there is still need for sound theories and frameworks into the nature of how and why it occurs. The case for incidental learning as a byproduct of reading has been substantiated by L1 studies which demonstrated that for 3.000 words learned per year between the ages of 9 and 18, intentional learning can account for only 200-300 words (Jenkins &Dixon, 1983; Nagy & Herman, 1984). It follows that vocabulary size in the L1 develops largely as a result of reading and listening when the focus is not on word learning according to the default hypothesis. By combining

data on lexical coverage of academic texts, vocabulary size, and reading comprehension scores, Laufer and Kalovski-Ravenhorst (2010) investigated the relationship between lexical coverage, vocabulary size, and reading comprehension. Their investigation led to the clarification of three important concepts as to the relationship between vocabulary knowledge and reading comprehension which are lexical coverage, sight vocabulary, and "adequate" comprehension. Accordingly, lexical coverage refers to the number/ percentage/ proportion of words in a given text that the reader successfully identifies. If, for example, the Academic Word List (AWL) integrating 570 words as proposed by Coxhead (2002) covers 10 % of an academic text, this means that learners with the knowledge of AWL can understand 10 % of that text. Sight vocabulary, on the other hand, refers to the words that the reader can easily recognize out of context and retrieve its meaning without much cognitive effort. Hence, if a word is in the reader's sight vocabulary list, the reader does not rely on the surrounding context in order to understand it. The researchers, thus, assert that the larger the sight vocabulary is, the more the lexical coverage will be which will inevitably lead to adequate comprehension (although "adequate" levels differ across contexts). Evidence is mounting that building a considerably large mental lexicon is vital. For instance, Sinatra, Brown and Reynolds (2002) emphasize that good readers save their conscious or intentional processing resources for recognizing or understanding new or difficult words since they already have an extensive amount of vocabulary knowledge to operate freely. Apart from facilitating comprehension, then, the greater the size and depth of vocabulary knowledge is, the greater the opportunities are for adding new items on it.

Whether vocabulary instruction should be explicit or implicit has taken substantial attention in the literature. Although comprehensible input, on its own, is

regarded as sufficient by some researchers (e.g., Krashen, 1995) to turn the newly-encountered words into lexical intake, within the realm of vocabulary instruction, it has been documented that extensive reading does not always guarantee incidental learning of all frequently exposed unknown words. Nagy, Herman, and Anderson (1985) in a first language context have demonstrated that school age children could answer only one in ten multiple choice vocabulary questions correctly after encountering those new words only once in passages. In Horst, Cobb and Meara's (1998) study which investigated incidental learning of vocabulary through extensive reading while the focus is on meaning construction, it has been observed that of the whole 21.000 words the participants were only able to recognize on average five new words although they encountered most of the words many times throughout the story. Moreover, it has been proven through their study that because reading in a second language is a slower and more laborious process, only a small group of learners really read in sufficient amounts for it to be a rewarding incidental learning experience. Thereby, multiple exposures to the targeted words must be assured in extensive reading if we want to develop incidental vocabulary learning. Grabe (2009), also, notes when the reader is exposed to a new word, some attention is directed to it in the process of meaning construction, but sometimes new words go unnoticed, or be even skipped during reading. As noticing contributes to the acquisition of linguistic features through conscious awareness of input, both incidental and instructed vocabulary learning are essential for the development of both vocabulary knowledge and reading comprehension due to the reciprocal causal relationship between the two. Learners need explicit vocabulary instruction targeting topical words, academic word lists or appropriate register. At the same time, they need to read extensively to reinforce their knowledge of more frequent words, to

encounter less frequent words, and to add on to the word families they learn through explicit instruction.

Incidental vocabulary learning is described as learning words from context when the attention is on the comprehension of the text rather than extending vocabulary size (Schmitt, 2000). There is aggregated evidence that incidental vocabulary learning in the context of extensive reading (i.e., Nation, 2001; Schmitt, 2000; Stahl & Nagy, 2006) enlarges learners' networks of vocabulary associations. The effects of access to word meanings during reading on text comprehension and learning new words have long been an area of interest for researchers. In one of those earlier studies, Knight (1994), for instance, investigated how the use of dictionaries during reading affected recall and incidental vocabulary learning of sophomore students learning Spanish. By determining students' verbal ability via the verbal scores of the American College Test, the researcher divided them into two groups as high and low verbal ability students. Another independent variable of the study was access to dictionaries; that is, whereas one of the groups had access to dictionaries, the other one did not while reading. The analysis of text recall and vocabulary tests displayed a significant effect of using dictionaries on both comprehension and vocabulary learning. Furthermore, dictionary use increased learning outcomes of low verbal ability students to a greater extent as they also performed similar or even better than high verbal ability students. Additionally, Rott (1999) looked for the relationship between text comprehension and vocabulary gains and retention of intermediate learners of German. Lexical gain and retention were measured by L2 to L1 translation task and a multiple choice translation recognition task. The results depicted a moderate to strong correlation between comprehension and vocabulary learning which was also observed to be strengthened over time. Lastly, Horst et al.

(1998) by keeping the books lengthy have shown that second language readers can gain a lot from incidental vocabulary learning.

Not all studies depicted a long-lasting effect of incidental vocabulary learning, though. Jacobs, Dufon, and Fong (1994), for instance, investigated whether glosses are effective for vocabulary acquisition through reading; i.e., whether text recall scores correlate with incidental vocabulary gain scores or not. The treatment involved reading an expository story and recalling it in L1 immediately after reading. The participants were also asked to translate target words into their L1. The results demonstrated that there was a modest significant correlation for intermediate Spanish learners who read with vocabulary glosses of target words and who read without glosses between their levels of text comprehension and vocabulary gain. However, this gain was not maintained 4 weeks later.

Another issue concerns how many encounters and how much involvement are required for incidental learning as the studies conducted so far have not made it explicit as to what kinds and amounts of reading facilitate incidental vocabulary learning except for a few rigorous attempts. The first of those was Craik and Lockhart's (1972) *the depth of processing* hypothesis which included two levels of processing: deeper encoding wherein a deeper and meaningful involvement in the processing of words brings on more permanent learning of those words, and shallow encoding, in which attending to surface features of words (i.e., its form) results in only short-term learning. The determining factor for the retention of new items in long-term memory was not the length of time that new information is kept in short-term memory but was the richness of processing; i.e., the spread and elaboration of encoding. Although a number of criticisms were directed toward the operationalization of constructs like the depth of processing, it is generally accepted

that the retention of new words is contingent on the quality (richness) and quantity (frequency) of encounters (Hulstijn & Laufer, 2001). In an attempt to operationalize the notion of deeper processing, Laufer and Hulstjin (2001) proposed the Involvement Load Hypothesis. They defined *involvement* as a motivational-cognitive construct including the components of *need*, *search*, and *evaluation*. The researchers explain the need component as such:

> The need component is the motivational, noncognitive dimension of involvement. Two degrees of prominence were suggested for need: moderate and strong. Need is moderate when it is imposed by an external agent. An example is the need to use a word in a sentence that the teacher has asked for. Need is strong when it is intrinsically motivated, that is, self-imposed by the learners, for instance, by the decision to look up a word in an L1–L2 dictionary when writing a composition. (Hulstjin & Laufer, 2001; p.543)

The other two components, *search* and *evaluation*, are cognitive dimensions incorporating attention allocation to the form-meaning associations. In particular, search is trying to find the meaning of a new L2 word form or trying to identify the form of a word to convey a particular meaning. Evaluation is, on the other hand, comparing the meaning of a possible word with its other meanings, or comparing the word with other words to determine whether the word is appropriate for a given context. These three components can be absent or present in natural or artificial vocabulary learning settings, rather it is the combination of these components with their relative importance which would determine the involvement load. They further exemplified the process as:

> Consider an example of two tasks that vary in involvement load. In task one, the learner is asked to write original sentences with some new words and these words are translated or explained by the teacher. The task induces a moderate need (imposed by the teacher), no search (the words are glossed) and strong evaluation because the new words are evaluated against suitable collocations in learner-generated context. If we want to describe the task in terms of an involvement index, where absence of a factor is marked as 0, a moderate presence of a factor as 1, and a strong presence as 2, then the involvement index of the task is 3 (1 + 0 + 2). In task two, the student has to read a text and to answer comprehension questions. New words, which are

> relevant to the questions, are glossed. The task will induce a moderate need to look at the glosses (moderate because it is imposed by the task), but it will induce neither search nor evaluation. Its involvement index is 1. Hence, task one induces a greater involvement load than task two. (p.544)

Hulstjin and Laufer (2001) contend that who has decided the task, whether the word has been searched, and whether it has been compared or combined with other words determine the degree of involvement load. As a result, "the greater the involvement load, the better the retention" (p. 545). With the help of two parallel studies, they set out to test the hypothesis that the retention of newly acquired words through incidental learning depends on the degree of the involvement load associated with the tasks. By employing three task types, namely (reading comprehension, comprehension plus filling in target words, and composition-writing with target words), the researchers tried to manipulate the combinations of need, search, and evaluation components with varying degrees of involvement load. The findings confirmed their prediction that the retention of words is connected to the level of task-induced involvement load since the retention of the newly encountered ten words was improved in following task order: composition task > reading plus fill-in > reading. Hence, they reached the conclusion that when a word is processed with higher involvement load, it is more likely that this word will be retained compared to a word which goes through processing with low involvement load.

To measure the outcomes of incidental learning of words through reading, a range of tests are required to portray gains and retention. As such, researchers tend to incorporate multiple tests to assess lexical intake. Waring and Takaki (2003), for instance, employed three different tasks to measure how many words were learned and retained after reading graded texts. The first test was a simple recognition task which entailed identifying whether the words took place in the text or not. The second test was a prompted recognition task which required choosing the appropriate

meaning for a given form among the options. The third test was an unprompted meaning production task which asked the participants to provide the translation equivalents of the target words. The results indicated that task type affects the scores to be obtained from a vocabulary test, and inevitably it influences the interpretation of those findings. As a consequence, the form-recognition test yielded the highest scores followed by the prompted meaning recognition which preceded the unassisted translation recall task. Based on this evidence, Waring and Nation (2004) maintain that vocabulary learning takes place over time cumulatively through repeated exposure to words. Although learners fail in terms of unassisted recall, this does not necessarily mean a total lack of knowledge. Being able to match the form with the correct meaning in multiple choice tests or to recognize the words from the texts can also show degrees of familiarity or ability to make form-meaning matches. Pulido (2003), likewise, made a distinction between sight vocabulary recognition and vocabulary production and measured the gain and retention scores of each. A multiple-choice test was used to gauge recognition of L2 to L1 translations, and a translation production test was used to test the ability to produce L2 to L1 translation. Similar results were obtained in so far as the effects of task type on the measurement were concerned. Namely, the multiple-choice meaning recognition test produced higher scores than the translation production test. The author admits that this is quite plausible considering the availability of retrieval cues which might have helped access the meaning of words as opposed to the absence of such cues for the production task. Later, Pulido (2007) included *intake* operationalized as "a measure of accuracy in memory discrimination for recently processed information" (p. 168) as an outcome of incidental vocabulary learning next to receptive vocabulary and productive vocabulary. To be able to assess intake, gain, and retention of newly

encountered words, then, more than one test type implemented more than once is needed to demonstrate the receptive and productive knowledge of vocabulary.

2.3 Electronic reading

Electronic reading, in a very broad sense, refers to reading linear or nonlinear texts on a screen (e.g., computers, tablets, mobile phones, etc.) with or without technological enhancement such as multimodal glosses or hyperlinks (Liu, 2005). By enabling noticing and sparing cognitive resources, reading on-screen can bestow learning, and the mediation of text content with readers' prior knowledge in an electronic reading environment can even facilitate MM building. Furthermore, reading in electronic environments offers individualized help to users aligned with their needs via the flexibility of paths to be followed (Reader & Hammond, 1994). By virtue of integrating technology and learner-centered contexts, it is generally believed to promote higher levels of learning than print text. That working memory is a limited-capacity system augments the importance of digital reading as it can free up cognitive resources by providing systematic and instant solutions to comprehension problems. However, in some other cases (e.g., recalling key points) print reading was found to yield better scores as opposed to digital reading (Singer & Alexender, 2017). In an attempt to describe digital reading, Liu (2005) notes an increase in "browsing and scanning, keyword spotting, one–time reading, non–linear reading, and reading more selectively" (p. 705) in addition to a decrease in careful and in-depth reading. Despite some pitfalls of electronic reading due to its discontinuous and flexible nature, it has been proposed that on-screen reading does not necessarily result in failed comprehension (Sweller, 1994; Zumbach, 2006). Similarly, Abraham (2008) notes:

> Since learners have limited cognitive resources available in working memory to manage higher- and lower level cognitive processes while reading, CALL is widely-believed to play an important role by facilitating text comprehension and retention of vocabulary with immediate and individualized support specifically designed to free-up these finite cognitive resources. (p. 200-201)

Thereby, through such benefits as interactivity, non-linearity, immediate access to relevant sources, the presence of audio-visual support, reading electronic texts can assist the multilevel process of reading comprehension.

Hyperlinks within on-screen reading environment provides the reader with tools such as glosses or annotations that assist a range of reading processes and enhance comprehension. The integration of annotations into electronic text allows readers to read flexibly and get immediate help without having to leave the page that is being read. Bell (2005) added other forms of aids, such as translations, further explanations, and visuals into the electronic reading environment and studied their effects as well as reader-related variables such as L2 instruction history. Specifically, the researcher analyzed how adult L2 learners of Spanish employed reading comprehension aids, how and whether their level of experience in L2 instruction influences their use of comprehension aids, and how tracking this interaction benefits the analysis of reading performance data. An authentic text, namely a Spanish short story was given through a computer network accompanied with aids in one of the following formats: 1) English translations of certain terms, 2) L2 definitions of the same terms, 3) historical information in the form of an essay written in L2, 4) an English translation of the same historical information written in essay format, 5) literary commentary in the form of an essay written in L2, 6) an English translation of the same literary commentary written in essay format, and 7) visuals pertaining to certain words or historical items in the text. The researcher was primarily interested in the number of consultation to particular resource materials which were designed as

comprehension aids. It was assumed that a 'successful L2 reader' would use those resources which cater to bottom-up processing and those which assist top-down processing in an interactive manner. The analysis of the number of times each resource was consulted showed that lexical items, especially English translations, (subservient to bottom-up processing) were the most frequently accessed comprehension aid. In terms of the relationship between the level of instructed experience in L2 and the use of comprehension aids, the findings conflicted with the researcher's predictions partially. Although it was expected that more experienced learners with better developed word-recognition abilities would consult lexical items less and non-lexical, global resources more, they exhibited a pattern which was just the opposite. Participants with less Spanish experience, on the other hand, were unable to apply successful bottom-up processing strategies to comprehend the story. The researcher concluded that readers in the upper group managed to construct a mental model of the text by applying bottom-up and top-down processing strategies simultaneously without the need to consult an external global resource.

In a meta-analysis of the use L1 glosses during L2 reading, Taylor (2006) compared computerized text to print text, and revealed that readers benefited more from access to computer-assisted L1 glosses like Abraham (2008) whose meta-analysis also indicated that readers generally benefit from reading computer-assisted texts in comprehension measures. It makes much sense because it creates a "tailored" reading experience specifically geared towards readers' background or interests (Foltz, 1996). But, this idiosyncrasy does not always guarantee comprehension or a coherent situation model building since flexibility may sometimes impair successful integration of textual information (Glanzer, Fischer, & Dorfman, 1984; Miall & Dobson, 2001). As Snow (2002) maintains "electronic texts that incorporate

hyperlinks and hypermedia introduce some complications in defining comprehension because they require skills and abilities beyond those required for the comprehension of conventional, linear print" (p. 14). When it is presented without any adjustments or modifications, it can lead to failures rather than success for some learner groups (Shapiro & Niederhouser, 2004). By looking at the demands it creates, instructional material designers should take necessary precautions in relation to the facilitative and adverse impacts of reading electronic texts.

2.3.1 Annotation use in electronic reading

Following Roby's (1999) taxonomy of glosses, a gloss is defined as a set of resources providing information which is not part of the reader's declarative or procedural knowledge. Following Yao's (2006) conceptualization, annotations in electronic texts refer to the glossaries, explanatory notes or additional information provided through hyperlinks. Apparently, the two concepts, namely annotation and gloss, refer to more or less similar things though the latter is more specific in description in terms of its plausible forms. That is why no distinction was made between them in this study, and they were used interchangeably. Annotations are intended to fill in the lack in the reader's available resources, and thus they should answer the immediate needs of readers without interrupting the flow of reading (AbuSeileek, 2008). Likewise, glosses "provide fast and easy access to the meaning of unknown words" and "compensate for insufficiently automatic lower-level processes" which allows the reader to attend to higher-level reading processes (Chun, 2006, p. 70).

The use of annotations/glosses is supported in order to help readers when text puts demands beyond their capability (Widdowson, 1978), to improve

comprehension, to enhance vocabulary learning, to address learning styles, and to enable more frequent use of authentic texts (Jacobs et al., 1994). Furthermore, Nation (2001) notes that first they enhance text processing for those readers who would not comprehend otherwise. Second, they supply the explanations of those words whose meaning cannot be inferred from the text. Next, they enable an uninterrupted reading context while reading especially in L2. Last, by getting the readers' attention to new words, they can facilitate vocabulary acquisition. Likewise, Abraham (2008) upon analyzing 11 studies on multimedia glosses, concludes that online glossing has a medium effect on reading comprehension and a larger effect on incidental vocabulary learning.

Studies investigating the use of annotations or glosses mainly cluster around two groups: the format of annotations and the place of annotations. Generally speaking, the format of glosses refers to the information presented within a gloss (e.g., definition, translation, background information, etc.) whereas the location of glosses designates where such information is placed in computer-assisted reading (Abraham, 2008) although it is difficult to make a strict distinction between the two in terms of terminology. For instance, while gloss format refers to the options of basic dictionaries or translated sentences in one study (Gettys, Imho & Kautz, 2001), it denotes the modes of presentation (e.g., in-text gloss vs. marginal gloss) in another (Chen, 2016). In some others, the glossing conditions which are framed as in-text, marginal, single-translation, or multiple-choice gloss are referred to as glossing types (Yoshii, 2013). In a rather different framework, Lomicka (1998) devised traditional gloss and multimedia extended gloss as the glossing formats and investigated their effectiveness over no gloss condition.

Another line of research has been established on the effects of gloss type which ranges from textual information to animated visuals on comprehension and learning new words through reading. In the literature, different presentation modalities of hypermedia information have been consistently termed as types of glosses. Yun (2011) builds a meta-analytic study on this conceptualization of hypertext gloss types which are described as "short definitions or explanations with nonlinearly linked data associated with graphics, audios, and videos in computerized texts" (p.41). In a pre and post-test design, Akbulut (2007a), also took text-only or text with visuals (i.e., text-plus-picture and text-plus-video) annotation forms as annotation types and studied their immediate and delayed effects on vocabulary learning and reading comprehension. In his study, groups receiving textual, textual plus pictorial, and textual plus video annotations were tested in terms of incidental vocabulary learning and text comprehension. The results demonstrated an interaction between vocabulary learning and annotation type since visual groups performed better than definition only group. Still, reading comprehension was not affected by annotation type. Another study on the type/format of glosses was is the frequently cited work of Chun and Plass (1996), one of the earliest studies, that investigated the effects of multimedia glossing. Specifically, they explored the extent of incidental vocabulary learning when the primary aim was text comprehension, the contribution of multimedia annotations to incidental vocabulary learning, and the correlation between look-up behavior and vocabulary test results. One-hundred-sixty learners of German read a short story presented in a special computer program, *CyberBuch*, designed by the researchers and completed vocabulary and reading tasks in the end. They conducted a series of three studies within the program which allowed readers to access to annotations in the form of text, picture, and video. For the results of the

second study, the authors concluded that especially picture+text annotations catered for vocabulary learning both in the short-term and in the long-term which suggests that "visual imagery aids in the learning of foreign words" (p. 194). However, no meaningful relationship was observed between look-up behavior and vocabulary test performance from which the authors drew the conclusion that other factors may influence readers' look-up behavior. In Lomicka's (1998) pilot study exploring if glossing helps L2 reading comprehension, or if it inhibits fluency in the second language, the results indicated no meaningful effect of glossing on inference construction as well. The researcher noted an effect of multimedia glosses on text comprehension and situation model building despite the lack of statistical significance due to the small sample size.

In another earlier work, Ariew and Erçetin (2004) explored the role of various hypermedia annotation formats and proficiency level in aiding text comprehension. Different forms of digital media including text, graphics, audio, and video annotations were used to annotate an expository text. Regarding the content of annotations, two types were included: textual (word-level) and contextual (topic-level) annotations. A total of 84 intermediate and advanced level ESL learners read the hypertext during which the type of annotations chosen and the frequency of access to annotations were recorded. Data analyses yielded three main findings. First, annotation use did not improve reading comprehension; on the contrary, more frequent access to audio and video annotations decreased comprehension for the intermediate group. Second, prior knowledge was found to be a strong contributor to comprehension for both intermediate and advanced learners although it was more so for intermediate learners. Third, the analysis of qualitative data depicted that

hypermedia reading experience gave rise to positive attitudes towards computerized reading.

The prevalent idea that the easy accessibility of glosses boosts look-up behavior (de Ridder, 2002; 1992; Hulstijn et al., 1996; Roby, 1999) has raised questions on the quality and quantity of incidental vocabulary learning on one hand, and building an MM on the other. For one thing, previous research suggests that although immediate access to word meaning augments vocabulary learning in the short-term, its effect is not long-lasting (Coady, 1997; Hulstijn, 1997; Jacobs et al., 1994; Nagy et al., 1985). For another, no significant effects on reading comprehension were reported in some studies (Akbulut, 2007a; Bell & LeBlanc, 2000; Cheng & Good, 2009); in some others, it was even found to decrease global passage comprehension (Ariew & Erçetin, 2004; Chen & Yen, 2013; Şakar & Erçetin, 2005). To further investigate this effect, de Ridder (2003) designed a study where she highlighted in-text glosses by font-colors and underlining and made them "marked". By exposing students to both the marked text (with visible links) and the unmarked text (with invisible links), she measured their interaction with the text (through log files), incidental vocabulary learning (both in the short term and in the long term), and text comprehension. The findings of the study were mostly confirmatory of the research hypotheses, especially of the one which maintained that readers would consult glosses more frequently if they are visible. Confirming the allegations of the previous research mentioned above, this frequent look-up behavior triggered by marked glosses led to an advantage in terms of immediate vocabulary gains; however, this short-term advantage disappeared over time.

As the popularity and availability of electronic texts have expanded, learners' preferences in line with their learning styles were also explored as regards the effects

of glosses on learning. By combining the influence of learning preferences with annotation type, Yeh and Wang (2003) designed an experiment in which they gave vocabulary annotations in three different formats, which are text-only, text plus picture, and text plus picture and sound annotations. The perceptual learning styles of learners are grouped as auditory, visual-verbal (with text), visual-nonverbal (with pictures), and mixed preferences. The results indicated while perceptual learning styles did not affect vocabulary learning, providing annotations with pictures led to the best vocabulary acquisition. In a way, pictorial annotations led to the greatest gains regardless of learning styles. Yanguas (2009) added noticing dimension into the research design and investigated the effects of gloss format on reading comprehension and vocabulary acquisition. While thinking aloud, ninety-four participants read a text in one of the four conditions (i.e., no gloss, textual glosses, pictorial glosses, and textual-pictorial glosses) and completed pre and post vocabulary and comprehension tests. This study targeted to explore whether any of the experimental conditions led to noticing or not and whether this noticing facilitated comprehension or vocabulary learning. The findings showed that all of the study groups outperformed the control group in vocabulary recognition whereas the textual-pictorial gloss group performed better than all the other groups in reading comprehension. Furthermore, all the glossed groups reported noticing the target words more than the control group. Lastly, it was concluded that the effects of multimedia glosses changed according to task types. Similarly, Garrett-Rucks, Howles and Lake (2015) investigated students' perceptions of reading texts with audio-visual aids. Seventy French language learners read traditional printed texts as well as hypermedia texts assisted with contextualized images, roll-over translations, cultural information, audio explanations and comprehension check exercises. Upon

reading, participants were given a format preference survey. The results depicted a preference for hypermedia texts in addition to the assumption that audio visually enhanced texts have facilitative effects on reading comprehension.

Türk and Erçetin (2014) examined the effects of control over choosing multimedia annotation type on incidental vocabulary learning and passage recall. Both textual and visual annotations were provided either interactively or simultaneously. In the interactive group, readers chose from a menu of visual and verbal information. In the simultaneous condition, readers were presented with both visual and verbal information at the same time on the same page. They were, then, tested on a number of vocabulary and comprehension measures. Confirming the spatial contiguity principle of the Generative Theory of Multimedia Learning (Mayer, 2001), simultaneous presentation condition yielded better results across tests.

Another strand of studies was clustered around the location of annotations, and the effects of gloss position on reading and word learning. AbuSeileek (2008), for instance, designed a study where the relationship between the place of glosses and vocabulary learning as well as reading comprehension was investigated. Hypermedia annotations were placed at the end of the text, in the margin, at the end of the screen, or in a pop-up window. Eighty intermediate level EFL learners were randomly distributed to one of the annotation groups. Overall, participants receiving hypermedia annotations outperformed their peers who had access to traditional glosses. Furthermore, annotations located in the margins were both influential in learning and preferable for learners. In a more recent research, Abuseileek (2011) studied the effects of the location and the type of annotations on beginner-level English learners' reading comprehension and vocabulary learning. The locations of

annotations were a) after the target word, b) bottom, c) margin, d) pop-up window compared to a control group who received no annotations. The types of annotations included either one-word synonyms or two to seven-word definitions of the glossed word. The overall comparison of hypermedia annotation groups with no annotation group portrayed that they significantly assisted reading comprehension and vocabulary learning. As to the location of annotations, glosses placed right after the word led to the best results. For the type of the annotations, readers benefited most from the glosses providing definitions in three to five words.

The investigation into the effects of the location of annotations on vocabulary learning and reading comprehension has also been conceptualized around the term gloss format. As a case in point, Chen and Yen (2013) measured both the readers' attitudes toward annotation use and the effects of annotation format over a four-week period. Participants read four versions of hypertext passages each week. One of those passages had no annotations, while three of them were annotated, each in a different format, namely in-text annotations, glossary annotations, and pop-up annotations. Although reading passages with in-text annotations yielded the lowest results in reading comprehension tests, pop-up window annotations led to significantly better comprehension supporting the view that glossaries presented in inappropriate formats may hinder rather than enhance comprehension. No significant differences were observed among the three annotation types for vocabulary acquisition while providing annotations as opposed to no annotations resulted in overall short-term and long-term benefits in terms of vocabulary learning. In relation to readers' attitudes toward annotations, they reported positive attitudes toward annotations and the glossary annotations were the least preferred annotation format. More recently, Chen (2016) compared the effects of different modes of gloss presentation (i.e., in-text,

marginal, and pop-up) on reading comprehension and vocabulary acquisition. One-hundred ten elementary level college students whose L1 was Mandarin read a computerized text in one of the following modes: a) in-text mode which granted verbal explanations within the text; b) marginal gloss mode which included lists of definitions in text margins; and c) pop-up mode which integrated verbal notes that could be viewed upon a mouse-click. Reading comprehension was measured through summary writing and a multiple-choice comprehension test while vocabulary acquisition was assessed through an L2-L1 translation test and an L2-L1 word-matching test. The results depicted that in-text gloss condition yielded the best results in all measures except for the multiple-choice comprehension test in which marginal gloss mode produced the best results. Pop-up gloss condition resulted in the lowest performance in this study. The author thinks this last finding might have stemmed from the inability of these lower level learners to take the advantage of this gloss format. Further, Yao (2006) studied the effect of hypertext gloss presentation format on cognitive load (CL), learning, and learner control in reading a web-delivered text. Five versions of a web-based text on educational psychology were generated which included annotations in one of the following formats: a) embedded annotation format (additional explanatory notes integrated into the text); b) online glossary (the list of definitions or explanations given in a hidden pop-up window); c) roll-over annotations (hidden pop-up window annotations which appear by rolling over the mouse on the glossed term); d) multiple annotations (annotations which are constantly provided to learners in embedded annotation, online glossary, or roll-over annotation formats); and e) no annotation. The results revealed no significant effect of annotation format on CL although an interaction was detected between glossing format and readers' prior knowledge. The researcher explained this interaction effect

in relation to Yeung, Jin, and Sweller's (1997) and Yeung's (1999) studies that demonstrated an interaction between gloss format and prior reading experience. In line with the expertise reversal effect proposed by Yeung et al., the gloss formats yielding the highest CL for the lowest prior knowledge group were the multiple annotation and no annotation formats while the embedded format produced the lowest CL. It was concluded that aligned with the principles of CL theory, the multiple annotation format must have created a redundancy effect by presenting annotations repeatedly and the absence of annotations must have turned reading into a more cognitively demanding activity for these novice readers.

A third line of research dealt with the medium of language used in glossing: L1 or L2. Bell and LeBlanc (2000), for instance, explored the preferences of English learners of Spanish with regard to L1 or L2 glosses. They employed a language experience survey, a vocabulary knowledge survey, a gloss-tracking device, a post-reading comprehension test, and an exit survey. The results showed that though readers preferred L1 glosses over L2 glosses, there was not a significant difference in the post-reading comprehension test. Recorded data revealed that while the Spanish gloss group accessed half as many word definitions as the English group, they did not report lower levels of perceived help from those glosses and they performed equally well in comparison to the English gloss group on the comprehension test. Another study that depicted no difference in terms of the effectiveness of L1 or L2 glosses is Yoshii's (2006) experiment conducted with 195 Japanese students learning English. The researcher added a visual dimension into the design and presented pictorial cues as well while glossing. The glosses were provided in one of these four conditions: 1) L1 text-only, 2) L2 text only, 3) L1 text plus picture, and 4) L2 text plus picture. Mixed-design repeated measures analyses revealed that there was not a

significant difference between L1 and L2 glosses in definition-supply or recognition tasks. Nevertheless, a significant difference was observed between no-picture and picture groups in definition-supply test. It was concluded both L1 glosses and L2 glosses are influential in incidental vocabulary learning, but their effects can change based on time. Complying with Yeh and Wang's (2003) findings, the availability of a visual presentation mode determined success for this study.

More recently, Fang (2009) explored the effects of L1 or L2 glosses for incidental vocabulary learning when students were reading for comprehension. Fifteen low proficiency level Chinese learners of English participated in the study. The findings pointed to the benefits of both types of glosses on account of their facilitative effects on noticing new words and directing attention toward them. Unlike Yoshii's (2006) findings, there emerged a difference between L1 and L2 glosses in terms of their effects over time. While L1 glosses were found to be more effective for short-term vocabulary learning, L2 glosses were more influential for long-term vocabulary retention. The author ascribes the short-term effects of Chinese glosses to shallow mental processing words due to quick access to its meaning. To the contrary, higher mental effort invested to derive meaning out of English (L2) explanations ensures deeper word processing which in turn leads to longer retention of words. Another published research in the same year was Cheng and Good's study which again focused on the effects of L1 glosses as well as of language proficiency. They recruited 135 business administration and engineering students who were randomly assigned to one of the four groups: no gloss, L1 (Chinese) gloss with L2 (English) example sentences, L1 in-text glosses, and L1 marginal glosses. Before the treatment, the participants were given a vocabulary pre-test containing 16 items, and they took different formats of the test containing the same target words as post-test,

and as the 1st and the 2nd delayed test. Other instruments included in the study were a five-itemed multiple choice reading comprehension test and a 24-item questionnaire on the use of glossaries. The findings showed that all glossing conditions facilitated vocabulary acquisition although they did not seem to affect reading comprehension. More specifically, L1 glosses were helpful for immediate recall of words while the acquisition of words regressed between the immediate and the 1st delayed test. Yet, interestingly, there was again a modest increase in their vocabulary retention between the 1st and the 2nd delayed test. Additionally, they noted that proficiency level did not mediate the effects of glossing for reading comprehension, but it was found to be effective for vocabulary acquisition. Finally, the findings of the questionnaire data revealed positive attitudes of students toward the use of glosses.

Ben Salem (2006) also investigated the combined effects of the language of glosses and type of glosses on reading comprehension and vocabulary retention. Ninety-three intermediate learners of Spanish (L2) were given a text with 25 annotated words in one of five conditions: 1) no gloss, 2) L1 (English) text only, 3) text and audio (pronunciation of word both in English and in Spanish), 4) text, audio, and picture, and 5) text, audio, picture, and writing (noting the gloss consulted on a piece of paper). Performance measures included reading comprehension and vocabulary acquisition tests as well as their interaction with the software as recorded by a tracking device. Overall, gloss use was found to be effective for both reading comprehension and vocabulary learning, yet no significant difference was observed as to the type of glosses that yielded better results. Finally, the author concluded that as the frequency of access to glosses increases, the possibility of text comprehension and word learning augments.

Following a rather novel approach, Wallen (2002) registered to Bloom's (1956) taxonomy of learning and organized hypertext annotations around selection, organization, and integration levels for a scientific text. In other words, annotations were designed to facilitate knowledge level, comprehension level, and analysis plus synthesis levels of learning. The results depicted that for recognition, selection-level annotations yielded the highest results. In the second experiment, he compared no access to glosses with glosses given in combination such as selection plus organization, organization plus integration, and selection plus integration. The results did not depict any significant differences across the groups. Later, based on a CLT framework, Wallen, Plass and Brünken (2005) investigated the functions of annotations in supporting cognitive processes in the comprehension of scientific texts. The main focus of the study was the contributions of text annotations (rather than picture annotations) to the processes of selecting relevant information, organizing the information in memory, or integrating information with prior knowledge. Ninety-eight college students were randomly assigned to the treatment conditions which included three types of text annotations designed in accordance with Mayer's (2001) CTML. Accordingly, selection-type annotations were constructed to aid the selection of relevant information and involved the definitions of the glossed term without contextual information. Organization-level annotations were generated to aid the connection of words into ideas and incorporated words or explanations within a specific context. Lastly, integration-level annotations were designed to aid the construction of nodes among different text parts as well as the integration of this information with prior knowledge. Annotations of such nature were in the form of links showing the directions of relationship within the text. As expected, the results depicted that different types of annotations were associated with

different learning outcomes. Besides, simultaneous presentation of multiple types of annotations led to the greatest levels of CL, especially for low verbal-ability students.

It is apparent from the above review of the relevant research that although the terminology referring to the information in the glosses and the one to the presentation forms of glosses differ, they typically underlie more or less the same concepts. As such, gloss types or formats can denote textual or contextual gloss content (verbal; e.g., Getz et al., 2001) delivered in various media forms (pictorial, animated, etc.; e.g., Chun & Plass, 1996; Akbulut, 2007a; Yun, 2011), and at the same time they can refer to the location of annotations (embedded/in-text, pop-up/roll-over, marginal/separated, etc.; e.g., AbuSeileek, 2011; Chen, 2016; Yoshii, 2013) in computer-assisted reading. Taken altogether, these studies suggest that the findings are inconsistent as to the role of (multimedia) glosses in improving reading comprehension and vocabulary learning. Additionally, the comparative effects of word-level versus topic-level glosses on learning through reading is a neglected area except for a few studies (e.g., Erçetin, 2003; Ariew & Erçetin, 2004; Wallen et al., 2005). Consequently, there is still a lot to be discerned in terms of the other factors affecting this process such as individual differences (e.g., working memory capacity, interest, motivation, etc.) as well as the combined effects of gloss type and position on these processes.

2.3.2 Cognitive load in electronic reading environments

In digital reading environments, the availability of additional information in the form of annotations or hyperlinks distributed across different nodes allows the users to retrieve information in an adaptable and self-paced manner. The possibility of creating a flexible text representation based on the route the user follows makes

electronic texts a favorite learning medium by the supporters of constructivist views (Zumbach, 2006). Yet, the flexibility of the navigational routes or the presence of a number of additional notes in multiple modes may constrain the readers' limited cognitive resources. Building mainly on the limitations of working memory capacity assumption and cognitive operations activated during multimedia learning, Sweller (1994) and Paas, Renkl, and Sweller (2004) proposed Cognitive Load Theory (CLT) to study learning processes with linear and nonlinear digital media. Cognitive load (CL), as a multidimensional construct, refers to the load induced by performing a specific task on the limited capacity cognitive system (Paas & van Merrienboer, 1994). Limited capacity assumption is one of the three underlying premises of CTML in addition to the separate channels for processing visual and verbal information and the active processing assumptions (Mayer, 2009). That is to say, WM has a limited capacity divided between the storage and processing components (Baddeley, 2000). Accordingly, learning occurs best under the conditions which are suitable for human cognitive architecture.

Over the last 25 years, CLT has been influential in explaining instructional mediums that help or hinder learning. In essence, what the theory puts forward is "that instructional techniques that require students to engage in activities that are not directed at schema acquisition and automation, frequently assume a processing capacity greater than our limits and so are likely to be defective" (Sweller, 1994, p.299). This framework describes CL imposed on the learner under three types which are intrinsic cognitive load (ICL), extraneous cognitive load (ECL), and germane cognitive load (GCL) (Sweller, van Merrienboer, &Paas, 1998). ICL stems from the inherent complexity of learning materials and depends on the prior knowledge of learners. In other words, it is the load an instructional material induces resulting from

the element interactivity of the material. According to Sweller (1994), "the level of element interactivity or connectedness refers to the extent to which the elements of a task can be meaningfully learned without having to learn the relations between any other elements" (p. 304). If the information elements are related, they interact in order for learning to take place. If, on the other hand, the elements are unrelated and can be learned consecutively, then interaction is not necessary. While the former scenario imposes high levels of ICL, the latter does not evoke such an overload. ECL is associated with the instructional design rather than the topic to be covered and it can also be the source of difficulty in certain contexts. It results from the ineffective material design and is not directly conducive to learning, so it should be kept at a minimum. However, when high element interactivity is not the case for a given material or ICL is already low, it may not be necessary to reduce ECL (Sweller, 1994). GCL refers to the integration of knowledge extracted from the resource into schemas. It is considered to lead to schema formation via the integration of learning content with learners' prior knowledge (Sweller et al., 1998). To this end, the aim of any instructional material design should be to minimize or eliminate any extraneous load and to create opportunities to maximize germane load so as to facilitate knowledge construction and integration process (Brünken, Plass & Leutner, 2004).

These three types of load were previously conceptualized to be additive to each other, therefore exceeding working memory resources results in decrease in information processing causing lapses in knowledge acquisition. In the previous model, learning is enhanced when ECL is minimized to spare resources for ICL and GCL which promote learning (Mayer, 2008). However, recently Sweller (2010) and Kalyuga (2011) questioned the value of GCL on the grounds that it is difficult to distinguish it from ICL. It was claimed that if the germane load is essential for

learning based on the processes of schema construction and/ or activation, then it is not much different from intrinsic load that is related to the number of information elements that need to be integrated for learning to occur (Kalyuga, 2011). Within this mapping, the author suggests that the compartmentalization as ICL and ECL are sufficient to be able to elaborate on the demands of instruction and design. In this reframing, GCL plays a rather different role wherein it stands as the actual WM processing responsible for dealing with intrinsic, not extraneous, load.

Kalyuga, Ayres, Chandler, and Sweller (2003) maintain that learners' level of expertise is another important aspect that instructional designers should pay attention since it may facilitate or worsen cognitive functioning. They contend that learning can help bypassing WM limitations by virtue of improving information processing with the help of activation of schemas reserved in long-term memory. For novice learners, devoid of the necessary schemas, there are no existing resources to assist them in cognitive processing. Instructional design can aid such learners via providing replacement for the missing schemas or even in the formation of the relevant schemas. In the absence of instructional guidance, they may opt for using compensatory strategies which would further limit the capacities of WM. Expert learners, on the other hand, equipped with the necessary schemas to construct a proper mental representation can effortlessly bring their stored knowledge into processing. Yet, the instructional assistance that is hard to ignore or avoid for such expert learners may hamper processing too. That being so, information that is conducive to schema formation for a beginner-level learner may become redundant and distract an advanced-level learner which is also called the expertise reversal effect (Kalyuga et al., 2003). In other words, information that would be beneficial for a novel learner can be ineffective or redundant for an expert learner. As a case in

point, McNamara, Kintsch, Songer and Kintsch (1996) showed that instructional aids for high school biology texts to increase coherence benefited only low prior knowledge readers whereas the original, not-enhanced text was more beneficial for high prior knowledge readers.

ICL also defined as the element interactivity of the material relative to the expertise level of the learner can determine total CL evoked by a learning material. For example, when the element interactivity is low compared to the current level of expertise of the learner, then the extraneous (or inefficient) CL caused by poor instructional design may not create an overload in WM resources (Yeung et al., 1997). By contrast, if the learning task is characterized by high element interactivity, then any redundant or ineffective instructional design feature may impose CL since it will be already high in terms of the intrinsic load. In such cases, Kalyuga (2008) suggests that reducing or eliminating ECL can be vital for meaningful learning. In line with the previously mentioned problems related to the ECL, the researcher offers several steps to follow in designing multimedia materials which are:

- enrich printed text with visual representations
- present visualizations and corresponding textual explanations simultaneously rather than successively to avoid temporal split-attention
- present related sources of information close to one another on screen (e.g., embed the text into graphic, avoid covering or separating information that must be mentally integrated for learning, design space for guidance or feedback close to problem statements)
- avoid irrelevant graphics, stories, interesting but irrelevant details, irrelevant sounds and music, nonessential words and lengthy text
- use visual representations explained by audio narration rather than onscreen text
- use animated visualizations with brief audio narrations rather than onscreen textual explanations
- present static or animated visualizations with narration-only instead of duplicating the narration with onscreen text. (Kalyuga, 2008, p. 52-53)

Complying with these principles, Mayer and Moreno's study (1998) demonstrated an advantage for the concurrent presentation of verbal and visual information in different modalities. In their two experiments based on the theory of dual-processing systems of WM, students viewed an animation depicting lightning formation and car braking system respectively accompanied by either audio narration or on-screen text. The results of both experiments displayed that the group receiving concurrent narration performed better in all test conditions. The researchers concluded that in multimedia learning, students can integrate words and pictures better when the verbal information is presented auditory rather than visually. This modality finding further confirms the dual-processing model of WM reiterating the idea of separate visual and auditory channels. Pass, et al. (2004) assert that it is not the level of the load that interferes with learning, but the source of the load. Thus, CL does not need to be kept low for every learning condition especially if WM limits are not exceeded. If the load is caused by mental practices that impede with schema construction or automation, then it is ineffective or extraneous. But, if it is induced by activities that are meaningful for schema construction and organization, like effective or germane load, then it will be beneficial for learning. Accordingly, instruction should facilitate learning by the activation of existing relevant schemata, and construction and automation of new ones (Kirschner, 2002; Sweller, 1994, 2005; Sweller et al., 1998). CLT puts forward that for cognitive load to be effective, extraneous/ ineffective load should be kept at a minimum by optimizing the level of intrinsic load. For instance, extraneous processing is imposed when learners have to split their visual attention between two places concurrently (split-attention effect) or if the instructional aid is redundant (redundancy effect). Paas and van Merriënboer's (1994) study, for example, showed that learners could invest germane load and work

on different practice problems only when they were provided with worked-out examples, a condition that reduces ECL.

CLT asserts that WM capacity is limited in the sense that in a particular learning situation we have limited cognitive resources to meet the instructional demands. Meaningful learning necessitates considerable cognitive processing, but a person's cognitive capability is highly limited (Mayer & Moreno, 2003). Cognitive overload is in action when processing demands surpass the available processing capacity (Sweller et al., 1998). Plass, Chun, Mayer, and Leutner (2003) investigated whether the participants' cognitive capacities mediate the effects of multiple representations of information on learning new words from reading. Learners divided according to their cognitive abilities read a German text with visual or verbal annotations. Their vocabulary gains and text recall scores were computed. Low verbal and low spatial ability students were found to be worse at vocabulary translation tasks than high verbal and high spatial ability students in visual annotations condition. Furthermore, all learners performed worst in recall tasks when they received visual annotations. The researchers concluded that multiple representations of information do not always facilitate learning. To the opposite, they may sometimes hinder meaningful learning by creating ECL for low-ability learners as a result of the processing demands of visual information.

WM as the set of cognitive resources briefly processing and storing information while dealing with other mental activities (Baddeley, 2003) is assumed to allow limited amount of information to be processed simultaneously, and this assumption corresponds well to the theory of cognitive load. This limited capacity WM can accommodate better processing and storing though while dealing with familiar information already stored in LTM since both schema construction and

automation can free WM resources (Pass et al., 2004). Multiple information sources can be recognized as a single element when knowledge is organized around schemas, and this will free up WM capacity. Learners' background knowledge, namely domain knowledge, is an important factor in comprehension as the interaction of text content and domain knowledge can turn L2 comprehension into an L1-like efficient cognitive process (Alptekin & Erçetin, 2011). As such, individual predisposition of readers (i.e., WM capacity) as well as the contribution of contextual information provided through annotations will be investigated in the current study as to their putative roles in the perception of cognitive load during electronic reading.

Particularly, the effect of the disparate positioning of information sources on knowledge integration and schema automation has attracted much attention as a source of extraneous load. Ayres and Sweller (2005) define the split-attention principle as designing instructional materials which "avoid formats that require learners to split their attention between, and mentally integrate, multiple sources of information" (p. 135). In designs which forces learners to divide their attention and mentally integrate information coming from sources which are physically and temporally separate, the mental integration of information which is essential to understanding induces ECL. Hence, a restructuring to spatially and temporally integrate disparate information sources is assumed to enhance learning, as it will cancel out the need for mental integration. In computerized learning, multimedia materials can enhance learning through cognitive processes which entail determining the relevant information to be extracted from the text, organizing them into mental representations, and integrating them into appropriate schemas, in a form ready to be recalled from LTM later (Moreno & Valdez, 2005). Still, poorly-designed materials may deteriorate the processing of available sources due to overloading the limits of

cognitive capacities leading to failure in constructing textually appropriate representations. Split-attention is one such source of a difficulty stemming from the presentation format of multimedia aids. When different information sources are spatially or temporally separated, the process of information integration may load limited WM resources with undue strain (Kalyuga et al., 2003). CLT stresses that for effective learning environments, novice learners should be given full support to integrate information mentally from the task. When novice learners face problems while working on a learning task, they should be supported by facilitative designs that will not exceed their already high CL. To prevent split-attention-effects that will overload the capacities of novice learners, a support system in the learning tasks should help them process and mentally integrate additional information (van Merrienboer, Kirschner, & Kester, 2003).

Two pioneering studies on the comparison of the traditionally designed instruction and integrated designs were Sweller and Chandler's (1994) and Chandler and Sweller's (1996) studies. In the former, participants in the traditional design had to consult the manual for instructions while simultaneously coordinating the use of the screen, keyboard or displays on the computer. Participants in the integrated format took all the information from one source (a modified manual) where diagrams and the text were also integrated to reduce the split-attention. The results indicated that the conventional group, despite having worked with the hardware, performed worse than the modified manual group in terms of demonstrating how the hardware worked due to the split-attention effect stemming from consulting the manual and the computer at the same time. Similar results were observed in the second study, in which the effect of the practice on the computers was also investigated with reference to the use of integrated format. Again, paper-only instruction presenting

disparate sources of information in a physically integrated format was found to be superior to the simultaneous but redundant use of the hardware and the manual. They, exponentially, added a new dimension to the investigation of split-attention effect which is element interactivity, the level of interdependence between the consequential parts of information to be learned, and it can be marked as low or high. Ayres and Sweller (2005) summarized the relationship between element interactivity and split-attention as follows:

> An inadequate instructional format may not overload working memory if the intrinsic cognitive load associated with the task is low. For tasks high in element interactivity, such as completing a complicated spreadsheet formula, a split- attention format will have a negative impact on learning. The addition of a heavy working memory load due to high element interactivity and due to split-attention may be overwhelming. (p. 142)

Proficiency level or domain-specific knowledge of learners also determine whether the physically integrated format facilitates mental integration or creates a redundancy effect because an information source which is necessary for a novice learner may be ineffective for an expert learner. When various information sources are externally integrated in this way, avoiding unnecessary information during processing becomes impossible, and the subsequent integration of this redundant information with learners' schemas may cause retarded learning for those learners who do not need such instruction. As a consequence, "elimination rather than integration" of redundant information is essential for learning with more experienced learners (Kalyuga et al., 2003, p. 25). Despite the common association of split-attention effect with multimedia instruction[4], the sources of this effect can also be text and their presentation modes (Ayres & Sweller, 2005). In terms of electronic reading, providing glossaries on a separate page may overload WM and impede learning since learner attention will be divided between the text and the annotation.

[4] "the presentation of materials using both words and pictures" (Mayer, 2001, p.1)

On the other hand, the integrated formats, i.e., in-text annotations, may create undue cognitive load and impede learning for learners who do not need such information. As Spiro et al. (1992) reiterate whether the same format of annotation will facilitate or complicate learning either through splitting attention or presenting redundant information is determined by learner needs such as their topic-related background or vocabulary knowledge.

To investigate the effects of split-attention and redundancy, Yeung et al. (1997) conducted a series of five experiments with the participation of learners from different age groups in both L1 and L2 contexts. Their first assumption was that young or low-ability learners can benefit more from an integrated glossary than a separate one regarding text comprehension since they would not have to divide their attention between the text and the words. Yet, for vocabulary learning, a separate format was predicted to yield better learning, as it would enhance learning words, and the necessity to devote attention to the text would be eliminated. On the other hand, adults or higher-ability learners would perceive the integrated word meanings as redundant and processing them may hinder comprehension. Nevertheless, the opposite would be the case in vocabulary learning for expert readers in an integrated format. Since they would not ignore the word meanings given within the text, they would be more likely to supply their definitions. The overall results emerging from the five experiments were as follows: The results of the first experiment showed that the use of explanatory notes increased the text comprehension of 5th graders while this effect was not maintained for vocabulary learning. The second experiment revealed that the integration of explanatory notes within the text rather than the separate vocabulary list improved the comprehension of 5th graders but decreased vocabulary learning. Reversely, in Experiment 3, it was seen that the integrated word

meanings hampered comprehension but enhanced vocabulary learning. The low-ability ESL group in Experiment 4 was found to behave similar to the 5th graders in Experiment 2. Likewise, in Experiment 5 expert ESL readers were found to act similar to the adult learners of Experiment 3. Altogether these findings corroborated their initial assumption that the facilitative effects of explanatory notes depend on the expertise of the reader and task type that would create either the split-attention or redundancy effects. This changing effect of instruction based on the learner is called *the expertise reversal effect* by the researchers.

Morrison (2004) specifically explored the split-attention and redundancy effects in electronic reading environments enhanced with glossaries in an L1 context. One-hundred eleven 5th grade students were assigned to one of the following conditions: a) no definitions, b) an online glossary, c) in-text definitions, and d) hypermedia rollover definitions. Upon reading two stories at two different levels, participants completed comprehension and vocabulary quizzes. Analyses showed that performance on reading and vocabulary measures was enhanced if the cognitive load imposed by the design was kept at a minimum, and if readers were provided with hover definitions[5] in electronic reading environments.

More recently, Al-Shehri and Gitsaki (2010) investigated the combined effects of gloss location and access to vocabulary glosses on reading comprehension and vocabulary learning to explore the split-attention effects. Twenty learners of English at an Australian language institution were randomly assigned to one of the following conditions: SAND (Split-Attention No Dictionary), SAOD (Split-Attention with Online Dictionary), IFND (Integrated Format No Dictionary), and IFOD (Integrated Format with Online Dictionary). The effect of split-attention was

[5] a type of hypermedia rollover allowing the reader to view the hidden content with the movement of the cursor

manipulated through either integrating the comprehension questions within the text or separating them. The results showed that the integrated format enhanced reading comprehension while access to glosses improved performance on the vocabulary tests. Also, SAOD and IFOD groups were found to spend more time on reading while SAOD group looked up more words during reading. As a result, the authors put forward that reading comprehension is facilitated when extraneous cognitive load is reduced.

In the general model of CLT proposed by Paas and van Merrienboer (1994), the assessment dimension apart from learner and task characteristics was also elaborated under three elements which are *mental load, mental effort,* and *performance*. Mental load is described as a division of CL "imposed by the task or environmental demands" (p. 122). It is believed to provide a priori information on the expected cognitive demands. Mental effort is another division of CL which "refers to the amount of capacity resources that is actually allocated to accommodate the task demands" (p. 122). Performance, on the other hand, refers to the subject's accomplishments, like the score on a test or the time spent on a particular test part, and it can be measured either during or after the task engagement. As for the measurement of CL, the researchers suggest the assessment of both mental effort and performance, and think that performance measures as indicators of cognitive load may be deficient in providing evidence when additional demands are enforced. For instance, on the condition that WM is not overloaded, the subject may overcome the mental load by investing more effort on a particular task and still succeed it which may not be revealed by looking at their performance results only. As such, measuring the mental effort as well can provide vital aspects which cannot be detected through performance-only measures. Furthermore, they assert that a combination of mental

effort with performance measures can make the computation of *mental efficiency* index possible. To calculate the reliability and sensitivity indices of the subjective scale of mental effort, the researchers compared the results with Paas' (1992) study whose scales were used to measure the perceived CL in the current design. The reliability coefficient was .90 for the former and .82 for the latter study. The sensitivity analyses yielded that the mental effort rating scale was sensitive to the fluctuations in the cognitive structure originating from different training strategies and from task complexity.

2.4 Conclusion

To be able to comprehend authentic L2 texts either in print or electronic formats requires an amalgamation of lower-order processes such as word recognition and higher-order processes such as the integration of macro-propositions of texts into already existing schemata and the interaction of these processes with reader-related factors such as motivation, strategy use, or interest (Abraham, 2008). Progressively, there have been several experiments conducted on annotation usage and their relative effects on readers' representation of text structure and content. Furthermore, across a number of studies the role of reader characteristics such as WM has been investigated with regard to their effects on reading performance (e.g., Just & Carpenter, 1992; Leeser, 2007; McNamara & Kintsch, 1996). Given the inconsistent results as to the role of WM in L2 processing (Daneman & Hannon, 2001;Harrington & Sawyer,1992), the effects of glosses on reading comprehension (Akbulut, 2007a; Chun & Plass, 1996; Yanguas, 2009) and vocabulary learning (Ben Salem, 2006; Chun & Payne, 2004) and the interplay between topic interest and text recall and comprehension (Ainley et al., 2002; Bray & Barron, 2004; Schiefele, 2009; Schiefele

& Krapp, 1996), there is still need for extensive research investigating these issues further. It seems that mostly one dimension of digital text design (e.g., only gloss format or only gloss modes) were analyzed in relation to comprehension, vocabulary learning, or CL leaving no room for other factors influential in reading process and vocabulary development, such as WM capacity which has the risk of overestimating the role of annotations in text comprehension and lexical development. Consequently, there is still a gap in the literature in terms of combining the influences of annotation content and location of annotations within the same research and measuring their combined effects on CL with the help of navigation tracking data. Therefore, this study will fill this gap in the literature and provide deeper insights into the cognitive processes of electronic text readers. Given the increasing popularity and availability of electronic texts and the possible facilitative or hindering (i.e., split-attention) effects of glossaries and WM capacity on text comprehension, word learning, and the CL, the present study targets to provide a richer understanding of the nature of these effects by adopting a factorial design. More specifically, this study will lend support to language teaching and learning practices by showing what kinds of annotations assist comprehension and vocabulary learning.

CHAPTER 3

METHODOLOGY

This section presents the methods and instruments recruited within the current study. First, research questions and hypotheses are listed to make the purpose of the study explicit. Next, the participant group and research instruments are introduced followed by a detailed explanation of the research procedure. The experimental conditions designed in this study are also described in this chapter supported with the screenshots from each of them. The chapter finalizes with the explication of data analysis processes implemented in this study.

3.1 Research questions and hypotheses

The main purpose of this study is to investigate how the type (lexical vs. topic-level) and the position (pop-up window vs. separate window) of glosses affect text recall and comprehension, word learning, and the perceived cognitive load (CL). Additionally, the study examines whether individual differences in working memory capacity (WM capacity) moderate the effects of gloss type and gloss position. Both quantitative and qualitative data were collected in order to answer three research questions:

1. What are the effects of gloss type (lexical versus topic-level), gloss position (pop-up window versus separate window), and WM capacity (low vs. high) on readers' recall, comprehension, incidental vocabulary learning, and the CL?

2. What are the effects of gloss type (lexical versus topic-level), gloss position (pop-up window versus separate window), and WM capacity (low vs. high) on frequency of access to glosses and time spent on reading?
3. Are frequency of access to annotations and time spent on reading related to reading comprehension, incidental vocabulary learning, and the perceived CL?
4. What are the participants' perceptions of the usefulness of glosses in terms of text comprehension, word learning, and the mental effort invested?

With regard to the effect of gloss type, it was presumed that access to the meanings of unknown words would facilitate rapid word recognition, which would in turn facilitate the construction of text base, and thus would contribute to building a micro-structure and a macro-structure of the text. In the case of access to topic-level annotations, readers were expected to rely on the strategy of guessing the meaning of words from context since lexical glosses were not available. This was not expected to hinder them from forming a coherent text base because advanced learners of English were observed to use this strategy efficiently instead of using definitional glosses (Ariew & Ercetin, 2004). In this group, topic-level annotations were hypothesized to facilitate building a situation model by providing learners with sufficient background knowledge about the topic and helping them make inferences required for situation model formation. In other words, providing readers with the necessary extra-textual information essential for meaning construction was assumed to help learners go beyond forming of a text model and allow better integration of textual nodes into the situation model of the text. As such, access to lexical versus topic-level annotations was not expected to make any difference on recall task performance that requires the formation of a text base (Hypothesis 1). On the other hand, access to topic-level

annotations which were not directly to any of the questions on the comprehension test was expected to facilitate performance on the comprehension questions (Hypothesis 2) since it requires a more detailed reading of the text and combining disparate parts of the text to be able to locate specific information or to make necessary inferences (Alderson, 2000; DeStefano & LeFevre, 2007).

Regarding vocabulary learning, readers with access to lexical annotations were hypothesized to outperform those who have access to topic-level annotations in both vocabulary measures, namely vocabulary matching and vocabulary production, and in both times of testing, i.e., immediate and delayed post-tests, regardless of gloss position (Hypothesis 3). In the literature, glosses have been consistently found to be conducive to incidental vocabulary learning during reading in comparison to reading without access to glosses (Nation, 2001). A number of studies conducted in both print and electronic reading environments (e.g., Chun & Plass, 1997; Grabe & Stoller, 2002; Laufer, 1997; Yoshii, 2006) have shown that providing readers with immediate access to unknown words will cater to the immediate intake of new words from the context and better vocabulary learning.

As for the effect of gloss position, separate window annotations were thought to induce extraneous CL by splitting the attention of learners between the text and the annotation content and prevent the efficient mental integration of text and annotation content (Sweller & Chandler, 1991). Split attention occurs as a result of disparate information sources' being used for mental integration. Accordingly, the reader's leaving the recently read page to reach the annotation content would increase the perceived CL. The integration of annotation content within the same window, on the other hand, would decrease the level of perceived CL reported by the participants as the time needed for temporary storage of information would decrease.

On the other hand, the effect of glossing position was hypothesized to depend on the amount of information provided by glosses. With short pieces of information such as word definitions in lexical glosses, the reader can more easily keep the information in working memory. As such, the SW condition should affect performance to a greater extent in the case of topic-level annotations due to the amount of information provided in these annotations. Accordingly, it was hypothesized that the participants accessing topic-level annotations in a separate window should have lower recall, comprehension and vocabulary scores (Hypothesis 4) and should experience and report higher levels of CL compared to the other three conditions (Hypothesis 5).

As for the effect of WM, in general, WM capacity is assumed to aid the performance of the participants in each level of processing. More specifically, in an attempt to answer how WM capacity affects reading, vocabulary learning, and the CL, it was hypothesized that high WM participants would be able to handle all CL conditions more easily. Thus, regardless of gloss type and gloss position they were expected to have higher recall, comprehension, and vocabulary scores (Hypothesis 6) and experience and report less CL (Hypothesis 7).

A final hypothesis pertains to the participants' use of the annotations as recorded through the tracking software. It was hypothesized that the participants would access lexical glosses more frequently compared to topic-level glosses due to the intrinsic load associated with the latter (Hypothesis 8). In addition, the pop-up condition was hypothesized to lead to frequent access to glosses compared to the separate window condition since the latter is hypothesized to cause attention split between the text and glosses (Hypothesis 9). As for reading time, it was expected that frequent access to lexical annotations would result in longer reading time (Hypothesis 10) in this condition. It was also expected that the pop-up condition

would yield less reading time thanks to the ease of access to glosses in this condition (Hypothesis 11).

Finally, as the last research question pertaining to readers' perceptions of hypertext reading experience is explanatory in nature, no hypotheses were generated on it.

3.2 Participants

The participants of this study are undergraduate students majoring in the Department of Foreign Languages Education at Yıldız Technical University, Turkey. All of the participants were native speakers of Turkish. As students of the ELT department, they were thought to be proficient users of English as they had to either report a TOEFL score of at least 96 or an equivalent score from a different exam (i.e., IELTS, CAE, or PTE), or take the English Language Proficiency exam given by the department to certify an equivalent level. Still, before being admitted to the study, they were given a version of Oxford Quick Placement Test (QPT) to unravel their proficiency level and to make sure that they are all at the same proficiency level. Previous research found the reliability of the test be almost 0.9 (Geranpayeh, 2003). Based on vocabulary and grammar knowledge, QPT tests learners' overall proficiency in English in two parts. The first part includes easier questions compared to the second part, so students are directed by the supervisors whether to continue the second part or not in the original test. However, since the participants' overall proficiency was of concern for the purposes of this study, all participants were asked to answer all 60 questions. Based on the results, it was observed that the proficiency level of the participants ranged from Upper-Intermediate (equivalent to at least 40 out of 60) to Advanced (equivalent to at least 48 out of 60) with a mean of 43 (*SD* =

5.6; Min.= 31, Max.= 56). Further, a one-way ANOVA was conducted on the QPT scores to explore if the groups differed in terms of the proficiency level. The results did not yield significant differences across the groups, F (3, 119) = 0.52, $p > .05$ showing that the groups were equal in terms of linguistic proficiency.

Before the exposure to the text, the participants' background knowledge related to the topic of "civilizations" was assessed through a scale prepared by the researcher (see Appendix A). A Shapiro-Wilk's test, visual inspection of their histograms, normal Q-Q plots and box plots exhibited that prior knowledge questionnaire scores were normally distributed for two groups (L-S and T-S groups) ($p> .05$) while the distribution was not normal for the remaining two groups (L-P and T-P groups) ($p< .05$). One-way analysis of variance across the experimental groups yielded no significant differences on the prior knowledge questionnaire, F (3, 119) = 1.57, $p > .05$, thus, the groups were considered equal in terms of topic-familiarity, as well.

Extra course credits were given to the students for their participation in the study. As such, the data came from 127 cases with upper-intermediate to advanced level of proficiency. After the analysis of data for normality, some outliers were deleted and normal distribution was ensured. Finally, the participant number in each experimental condition was 30. Each group was further divided into two halves as low WM vs. high WM capacity participants. Of the final 120 participants, 92 (77 %) were females and 28 (23 %) were males. Table 1 presents the distribution of participants across the groups and the number of participants in each cell.

Table 1. Number of Participants in Each Cell of Analysis

L-P *N* = 30		L-S *N* = 30		T-P *N* = 30		T-S *N* = 30	
Low WM *N* = 15	High WM *N* = 15	Low WM *N* = 15	High WM *N* = 15	Low WM *N* = 15	High WM *N* = 15	Low WM *N* = 15	High WM *N* = 15

3.3 Reading text

In order to simulate real reading experiences of this group of participants who constantly read lengthy passages in English as a requirement of their major, an article consisting of 2,580 words and nine paragraphs was selected from a book preparing students for proficiency exams (see Appendix B). As students of the Department of Foreign Language Teaching, the participants of this study have always been involved in reading texts related to social issues or humanities. Besides, building on the findings of recent instructional theories reporting the benefits of real-life tasks as a motivator for learning (Merrill, 2002; van Merrienboer & Kirschner, 2001), a social-sciences-related text is assumed to be motivating for this group of readers. Hence, both to replicate their usual reading environments and to motivate them to read, a lengthy article focusing on the reasons of societal collapse was chosen. Next, the text was converted into an electronic text by dividing the material into seven internally coherent sections for the users to pursue reading without getting distracted by scrolling down. The readers could navigate through the sections by clicking on "next" or "previous" buttons. They could also see which page they were on by looking at the navigation map which was placed on top of every section. The current page they were viewing was highlighted in bold characters on the navigation map (see Figure 2).

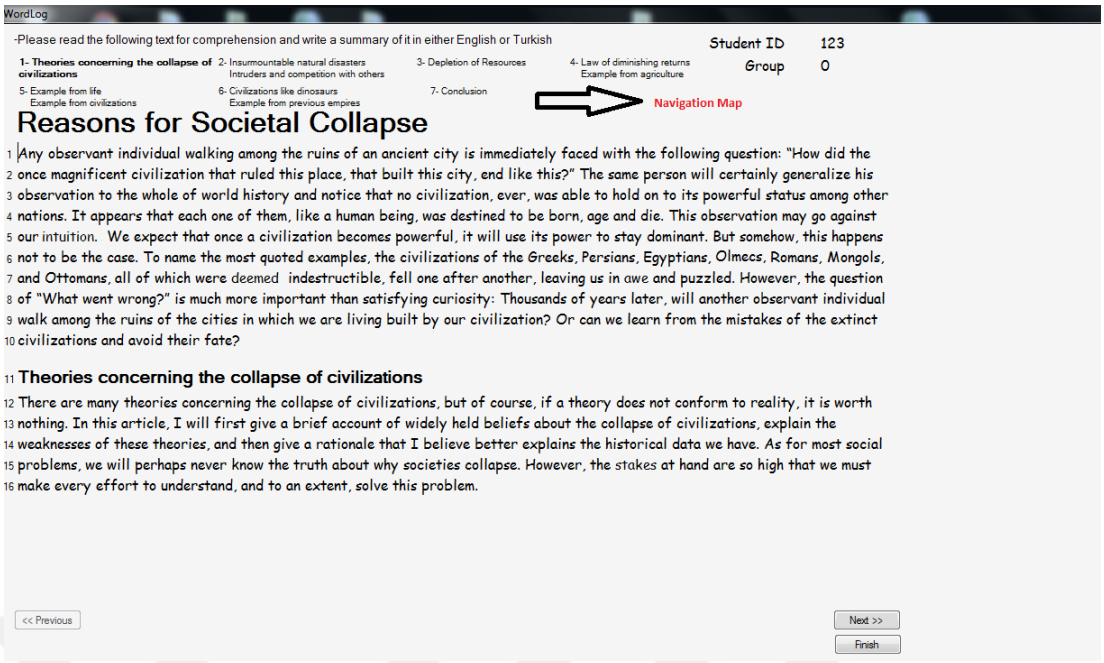


Figure 2. Navigation map in the hypertext

Word-level or topic-level annotations were incorporated into the text through hyperlinks. WordLog software using C# programming recorded the interactions of the readers in the form of frequency of access to annotations and total time spent on reading.

3.4 Piloting

The researcher conducted a piloting session so as to determine the words and the topics to be annotated. To this end, a group of twenty-nine students who did not take part in the actual study had been asked to read the pen-and-paper version of the text and underline any words that they did not know and circle around any topic on which they would need extra information to understand the passage. Upon reading the text, the comprehension test, and two vocabulary measures, namely vocabulary matching and vocabulary production tests were administered. The amount of time needed to complete the overall reading, and question-answering was calculated: approximately one and a half hours. For determining the words/ text parts to be annotated, the

frequencies were calculated. Accordingly, if a word was underlined by more than half of the participants, it was included in the list of the target words. As such, a total of twenty-eight words which met the criteria were chosen to be annotated in the actual study. Similarly, twelve topics which were circled around by at least half of the participants were given as topic-level glosses.

3.5 Data collection instruments

It has been known for long that the methods of testing and the use of specific tasks affect both readers' performance and the outcomes of the test (Alderson, 2000; Bernhardt, 1991, Shohamy, 2001). To investigate the issues in question within the current study multiple tools were used. In particular, the instruments used in the study were a topic interest questionnaire, a backward digit-span task, an electronic reading environment (which integrates the annotated text, recall task, comprehension test, vocabulary production and vocabulary recognition tests), two subjective cognitive load scales, and semi-structured interviews.

3.5.1 Independent variables

3.5.1.1 Gloss type

Words that contained either a definitional or a topic-level explanation were written in boldface.

3.5.1.1.1 Lexical annotations

L2 definitions which were taken from Oxford Monolingual Online Dictionary were provided for twenty-eight words within the text (see Appendix C). As already noted,

these target words were determined by the students taking part in the piloting phase. This condition is given in two different formats, namely as pop-up-window or as separate-window.

3.5.1.1.2 Topic-level annotations

Extra-textual information related to the previously-chosen twelve topics was given in English (see Appendix D). Information of this kind included short excerpts related to some concepts, important events, names of theorists, and major civilizations written in 30 to 70 words. The content to be included in the annotations was determined by the researcher according to the level of importance for later mental integration. Further, expert opinions were taken from the colleagues as to the content of annotations and necessary adjustments were made. Topic-level annotations were not directly related to any of the questions in the comprehension test. Rather, they were assumed to support global comprehension and inferencing indirectly by catering topic-related extra information. As with the lexical annotations, contextual annotations were given in two different conditions.

3.5.1.2 Gloss location

Lexical or topic-level annotations written in Tooltip plugin appeared in a pop-up box or in a separate window.

3.5.1.2.1 Pop-up-window

In this format readers without leaving the page could access the information by just moving the mouse on the highlighted word to access definitions of words (L-P condition) or to access extra information about the topic (T-P condition). These pop-

up-window annotations did not cover the space of the selected words so that the reader could see the part which the word takes place in and their explanations simultaneously in an integrated format. Annotation content disappeared as soon as the reader moved the mouse away from the annotated word (see Figures 3 and 4).

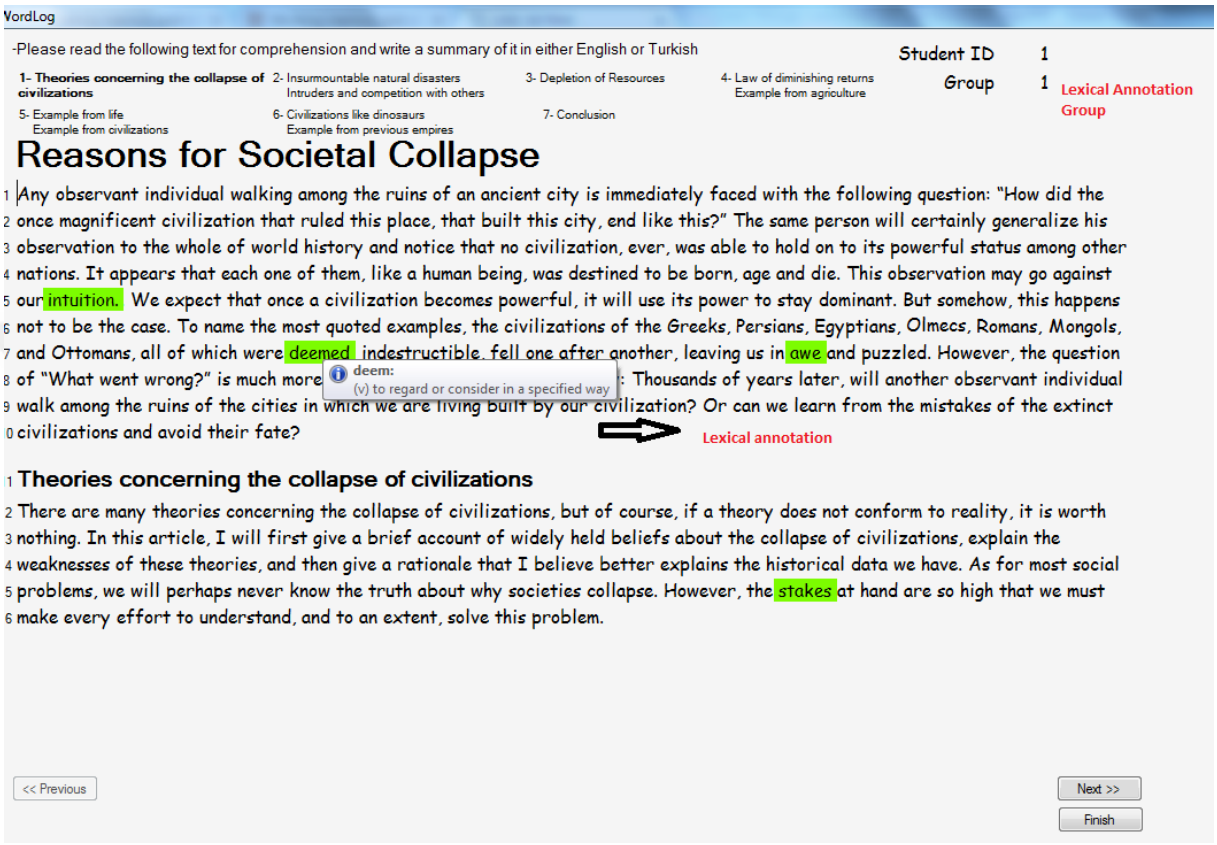


Figure 3. Lexical pop-up window annotation condition

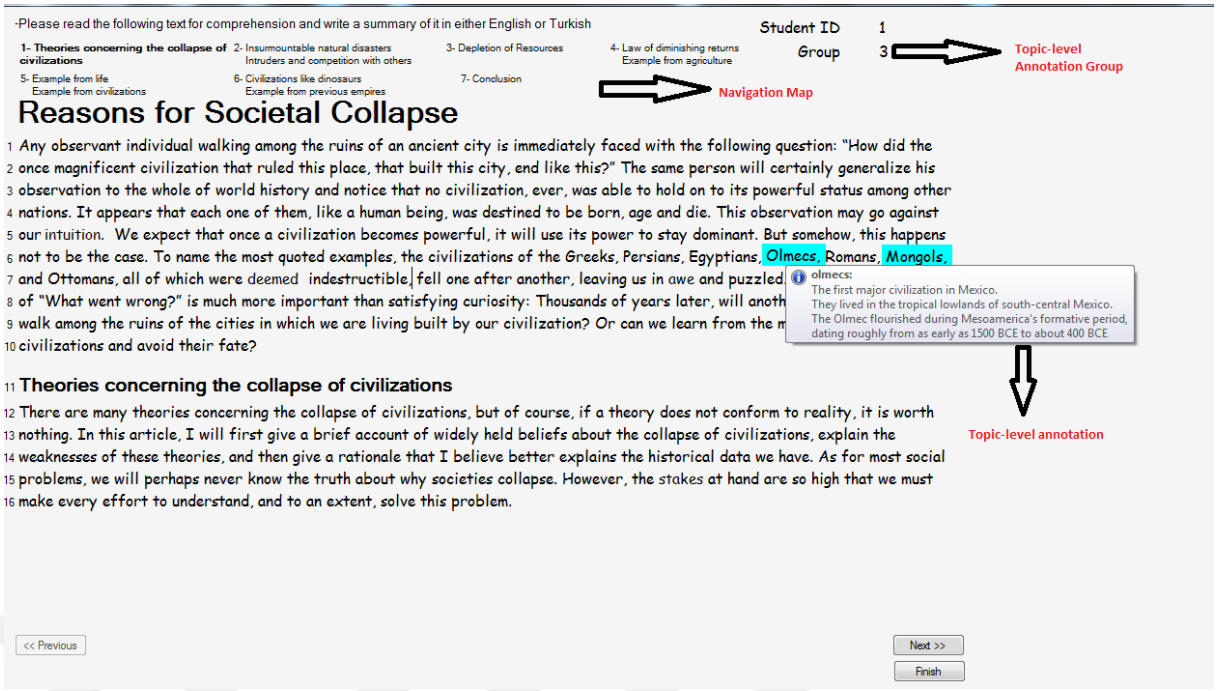


Figure 4. Topic-level pop-up window annotation condition

3.5.1.2.2 Separate-window

In this condition annotations were given in a box on white background which appeared on the desktop by closing the currently read page. Similar to the pop-up window condition, readers could access the annotation by just moving the mouse on the highlighted word in a separate window this time. When they read the annotation content, they had to click on a blank area on that separate page in order to return to the hypertext reading environment. Figure 5 provides a screenshot of separate window lexical annotations (the L-S condition) and Figure 6 shows the separate window contextual annotation (the T-S condition).

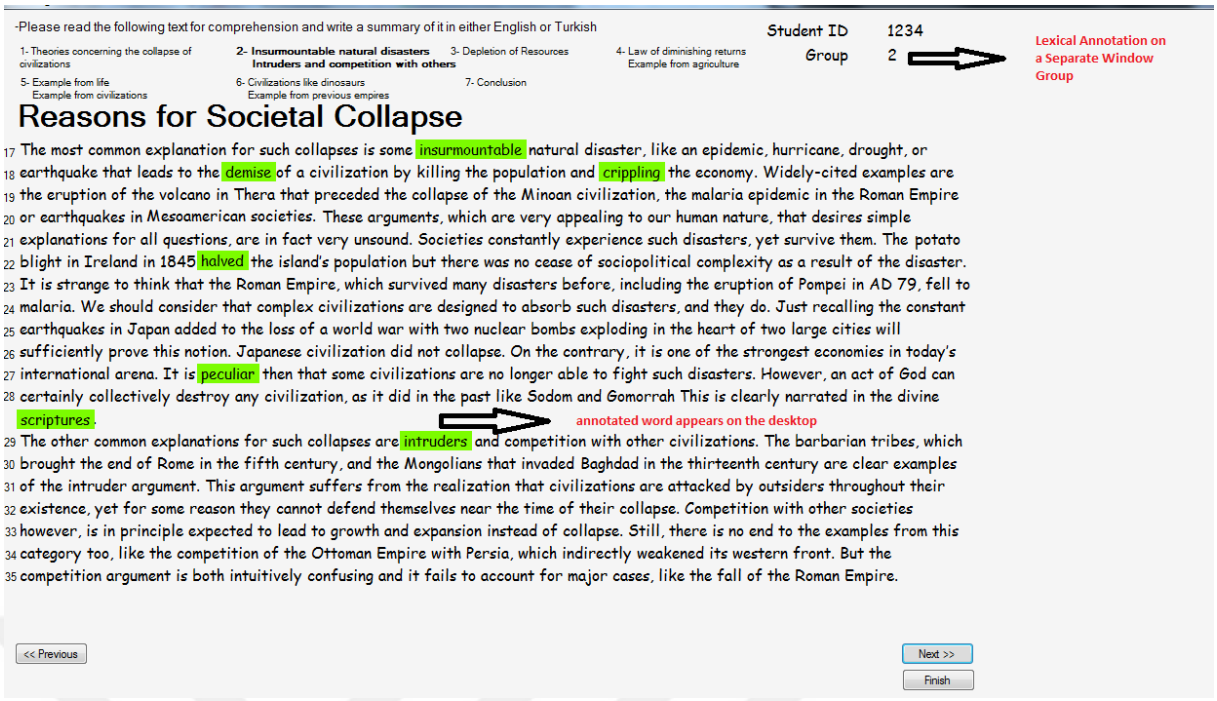


Figure 5. Lexical separate window annotation condition

When the reader chooses the word '*intruder'*, for example, the current page on which the reader is reading the text closes, and a new window containing the gloss information opens up on the screen as it is shown in Figure 5:

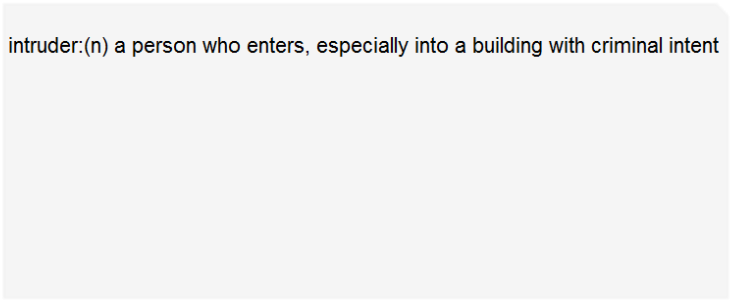


Figure 6. Lexical annotation box viewed on a separate window

As for topic-level annotations, if the reader wants to get additional information on the concept '*potato blight'*, a box containing the information on a separate-window like the one shown in Figure 8 could be viewed.

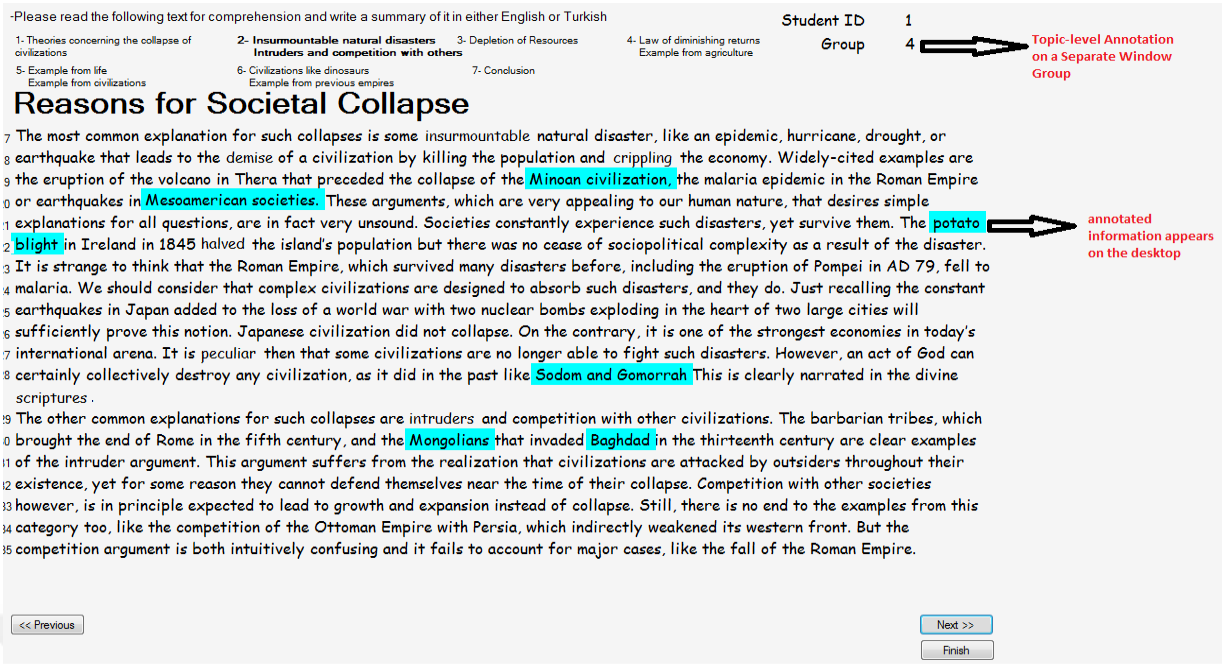


Figure 7. Topic-level separate window annotation condition

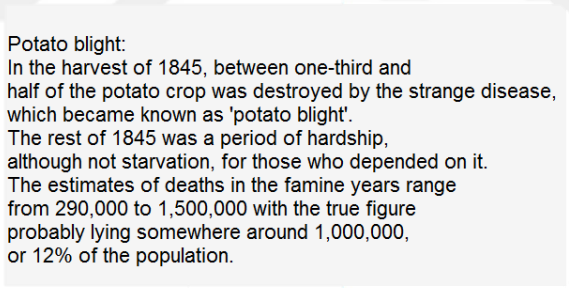


Figure 8. Topic-level annotation box viewed on a separate window

3.5.1.3 Backward digit span task

The Memory for Digit Spans from Wechsler's Intelligence Scale was used to determine WM capacity of the participants for the current study. Digits Forward is a simple span test which gauges the storage and maintenance components of working memory. Digits Backward is a more complex span task which entails the manipulation of verbal information while it is in temporary storage. Testees listen

and repeat sequences of digits read aloud by the examiner in Digits Forward task. In Digits Backward task, testees listen to and repeat the sequences of numbers with increasing number of digits in the reverse order. The task was to keep the numbers in memory while counting them backwards in the order they were first told. In comparison to the simple span tasks which measure the amount of units to be stored for a limited time, complex span tasks, such as backward digit span, require the individual to execute additional processing tasks while performing a short-term storage task at the same time. It is intended to probe central executive component of Baddeley's (1986) model of working memory (Schüler, Scheiter, & van Genuchten, 2011). Only Digits Backward task was employed in the present study because it is thought to load more on WM processing as a more complex measurement (Psychological Corporation, 1997). The assessment was administered in Turkish by the researcher one-on-one with each participant. The total maximum score was 14. During the analysis, it was turned into a categorical variable consisting of two levels: low and high. As such, scores below the median (7) were marked as low WM capacity and scores above the median were marked as high WM capacity.

3.5.2 Dependent variables

3.5.2.1 Text recall

As off-line performance measure of reading, participants were asked to produce free recall of the text. Free recall protocols are conceived as valid and integrative tools for checking comprehension subsequent to reading (Berndhart, 1991) unlike discrete point tests which portray a "fragmented, compartmentalized" (Riley & Lee, 1996, p. 174) view of comprehension. Berndhart (1983) features them especially because they

diminish the factor of guessing answers and they do not interfere with students' comprehension of the text via preset questions. By following the same logic, free recall protocols were used in this study as a measure of text recall. When the participants finished reading, they wrote what they could remember from the text either in English or in Turkish without any time constraints. Language restriction was not imposed so that the students' reproductive abilities would not impede their performance (Lee, 1986; Lee & Riley, 1990). As such, 52 % of participants preferred writing in Turkish while 48 % of them produced the written recall in English.

Alderson's (2000) scheme of propositional analysis based on main idea units was used for the quantification of summary data. To do so, macrostructures including the coherent and connected macro-propositions from the text were created to give a more global account of the text (Kintsch & van Dijk, 1978). Accordingly, a recall protocol for the text in the form of main idea units was constructed by the researcher and a PhD candidate who volunteered to read the text. Next, the total number of main idea units were obtained which yielded a total possible score of 26 for the recall task. Finally, the number of main idea units in each participant's recall was compared to this framework (see Appendix E). Each correct idea unit was given one point; unrelated sentences or micro propositions which did not comply with the pre-set scheme were not given any points.

A subset of recall protocols (12 %) was rated by an independent rater who was also a non-native EFL instructor at college level with more than 10 years of experience especially on teaching reading and writing. The inter-rater reliability coefficient was calculated to be 0.852, and the disputed propositions were resolved.

3.5.2.2 Reading comprehension

Free recall task, on its own, cannot be counted as a measure of overall comprehension as it would also include memory effects (Alderson, 2000). Furthermore, as assessing reading by using only one method is believed to be insufficient, Alderson (2000) recommends complementing subjective techniques with objective ones. Daneman and Hannon (2001) showed that although multiple-choice tests were criticized as a measure of reading comprehension for the possibility of success without even consulting the passages, they do actually tap comprehension if readers take active part in reading. Martinez (1999) notes that multiple choice items can tap complex cognitions, such as understanding, prediction, evaluation, and problem-solving. Hence, in order to erase memory effects and assess a deeper and analytical understanding of the text, a reading comprehension test was added.

After finishing the recall task, the readers took the comprehension test which consisted of 14 multiple choice items tapping of literal and inferential (higher levels of Bloom's taxonomy of learning) comprehension (see Appendix F). The test was constructed by the researcher with the continual feedback of the advisor and additional feedback obtained from the colleagues. Based on this feedback, five question stems were changed so that they could be easily answered by the test-takers. As a third step, five students who did not take part in the actual study had taken the pilot test. They were told to take notes about any ambiguous or confusing items while taking the test. Not many adjustments were required as a result of the piloting, yet some distractors were paraphrased in order to make them easily understandable. Lastly, Cronbach's alpha was computed for this 14-item test (α =.718), and the test was found to have a sufficient level of internal consistency. A screen shot of the page with the comprehension questions is displayed in Figure 9.

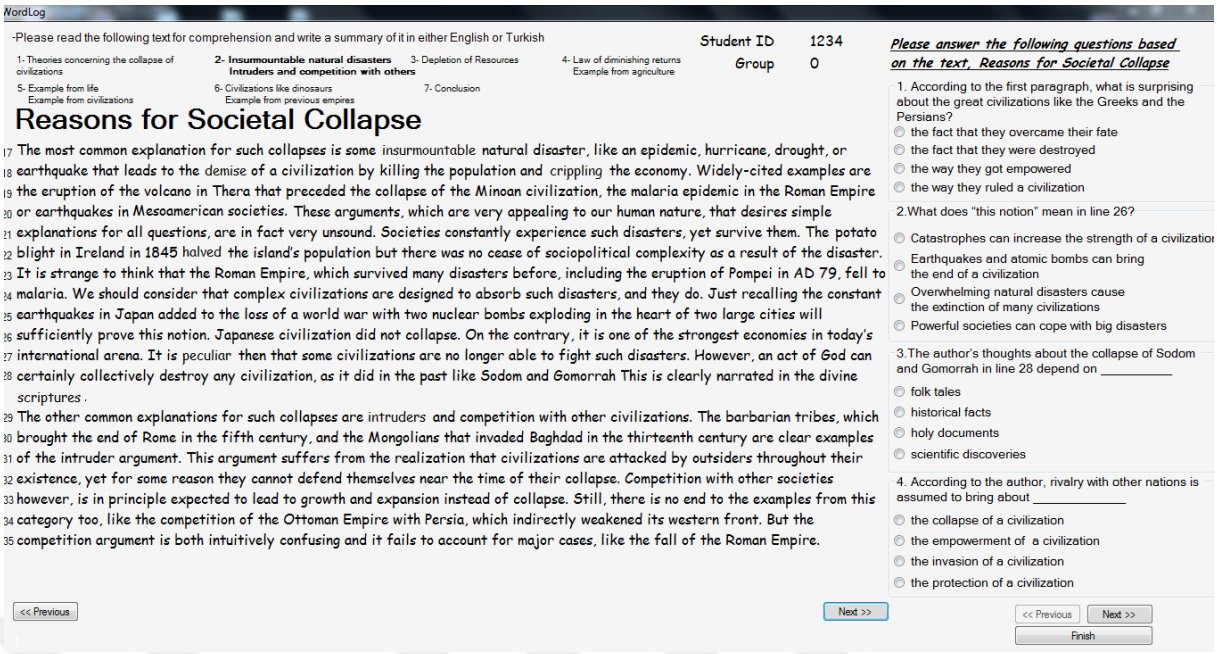


Figure 9. A screenshot from the comprehension test

3.5.2.3 Vocabulary production test

Based on Laufer and Goldstein's (2004) dichotomy of active versus passive knowledge of words, a vocabulary production task which required the participants to supply the definitions of all 28 target words either in English or in Turkish was prepared by the researcher (see Appendix G). The test represents passive recall from the hierarchy of form-meaning relationship which requires the test taker to show an understanding of the L2 word. Two detailed answer keys were prepared by the researcher for both the English and Turkish definitions of the words. The English definitions in the key were the same as the definitions given in the lexical glosses which were taken from the Oxford Monolingual Online Dictionary. L1 definitions in the key were constructed based on the Cambridge English-Turkish Dictionary. The answers were rated by the researcher, and one point was granted for the correct description of each word meaning. A subset of test papers (12 %) was assessed by an independent rater who was also a non-native EFL instructor teaching reading and

writing at tertiary level for more than 10 years. An inter-rater reliability of .923 was obtained, and the conflicted items were resolved in a conference. Three weeks later, the participants were given the pen-and-paper version of the same test to assess vocabulary retention. A moderate correlation was obtained between the immediate and the delayed vocabulary production tests (r= .54) but looking at the significance of the value (p = .000), the robustness of the tests is reassured.

3.5.2.4 Vocabulary recognition test

Receptive knowledge of readers regarding the words they encountered during reading was tested through a matching task. The test is representative of active recognition in Laufer and Goldstein's (2004) hierarchical framework since the test taker had to match the form with the most appropriate meaning. Out of 28 target words to which hyperlinks were given in two of the experimental conditions (i.e., L-P vs. L-S), 14 were included in this matching task. The items were chosen with the intention to make the list a representative sample of the whole vocabulary list; namely giving equal proportion of nouns, verbs, adjectives, and adverbs. This was not a simple recognition activity as all 28 words were given in the matching list to choose from, but there were only 14 definitions to match with them. The task was not simply to match the words with the appropriate definition, but also to eliminate the distractors. In this regard, it was not only a lexical intake but also a lexical gain task. The test was designed by the researcher with the help of the advisor. In the second draft, expert opinions were taken from three colleagues who were also graduate students, and the five students mentioned above took the test for piloting. Accordingly, definitions of four words were rewritten to include further explanations and examples to better illustrate the meaning (see Appendix H). The Cronbach's

alpha was .701 which indicates a sufficient level of internal consistency for the test. Additionally, by conducting Spearman's rho, a moderate correlation (r= .42) was observed between the immediate and delayed vocabulary recognition tests, but again the high significance level (p < .001) assures that this is not a chance result rather a consequence of the large sample size (N= 120).

3.5.2.5 Perceived cognitive load

After reading the text, the participants in all conditions were given rating scales to report the perceived CL evoked by various treatment conditions (see Appendix I). Subjective Cognitive Load Scale of Paas (1992) was employed as an indirect measure of cognitive load. Difficulty and mental effort scales were handed to the participants upon the completion of tasks. Originally, Paas estimated a reliability coefficient of .90 for this measurement, and its internal consistency had an alpha value of .79 in the current study. This unidimensional scale asked subjects to rate the mental effort invested in understanding the reading text on a 9-point symmetrical rating scale. 1 indicated very, very low mental effort, and 9 corresponded to very, very high mental load. In the difficulty scale, the participants had to choose from a 9-point scale, again, ranging from 1 (very, very easy) to 9 (extremely difficult).

3.5.2.6 Annotation use

Information regarding online performance variables such as total reading time, and the frequency of access to annotations came from a built-in tracking software within the electronic reading environment. They were considered as more direct indicators of CL assuming that participants were mentally involved with comprehending the text. Of particular interest was the relationship between annotation use (which would

increase as the amount of time spent on annotations and the frequency of access to annotations increase) and the reported CL.

3.6 Covariate

3.6.1 Topic interest

Prior to reading the text, readers filled out a topic interest scale. Schiefele and Krapp's (1996) topic interest questionnaire was used in this study with some adjustments in order to determine "feeling-related valences" and "value-related-valences" of readers. Specifically, footnotes were added at the end of the questionnaire to further elaborate what is meant by some concepts (i.e., *stimulated, involved,* and *engaged*) (see Appendix J). On the page of the questionnaire, the topic of the reading passage was explained in order to give brief information on the text. In the first part, "feeling-related-valences" ("bored", "engaged", "interested", "stimulated", "indifferent", "involved") are given with a four-point rating scale ranging from "completely true" to "not true at all". Upon reading the brief information about the text, participants expressed how they expected to feel while reading it. In the second part, another subscale was used to uncover how important the topic is for the participants personally. Three "value-related" adjectives ("meaningful", "useful", and "worthless") were given with a four-point rating scale ranging from "very" to "not true at all".

Schiefele (1990) reported the reliability coefficient of the feeling-related subscale as .91, and of the value-related subscale as .89. While scoring, feeling-related-valences and value-related-valences are added up to each other to reach the total score of the participants (the possible maximum score = 36). Two items (*bored,*

and *worthless*) are reversely coded during scoring to make them compatible with the rest of the scale.

Table 2 provides the descriptive statistics for the topic interest scores across the groups. Topic-related interest of these groups of readers was fairly high since their means were well above the average. The Shapiro-Wilk's test ($p > .05$) (Shapiro & Wilk, 1965) and a visual exhibition of their histograms, normal Q-Q plots and box plots revealed that the topic interest scores were approximately normally distributed across four treatment conditions. An analysis of variance showed that the effect of topic interest was nonsignificant across the four treatment conditions, F (3, 119) = .633, $p > .05$, so the groups were considered equal on this covariate.

Table 2. Descriptive Statistics for the Topic Interest Scale across the Groups

Groups	*M*	*SD*	*SE*	*Min.*	*Max.*
L-P	23	4.5	0.82	16	32
L-S	24	4.2	0.76	15	31
T-P	24	3.9	0.73	16	30
T-S	24	4.2	0.77	14	33

3.7 Semi-structured interviews

Within a week after the implementation of the computer-based reading, semi-structured individual, paired, and focus group interviews were conducted in Turkish. Interviews were held with a total of 18 volunteers. The distribution of interviewed participants across the groups were as follows: two students from the L-P condition, three students from the L-S condition, six participants from the T-P condition, and seven participants from the T-S condition. Semi-structured interview format was adopted mainly because it lets the interviewer to flexibly handle the responses, and attend to any recurring topics. Merriam (2009) puts forward that in semi-structured interviews either all questions are flexibly structured or the interview is conducted with the equal distribution of more and less structured questions. Accordingly, the

questions determining the focal point of attention are pre-constructed, but the other points to be taken are not so strictly worded nor ordered beforehand. Krueger (2015) contends that "focus groups work particularly well to explore perceptions, feelings, and thinking about issues, ideas, products, services, or opportunities" (p. 7). Aiming at probing the effects of a particular electronic reading environment, focus group interviews together with individual and paired formats were adopted for the current study. There were seven pre-determined questions for the interview within the current study regarding their experiences with print vs. electronic texts, their preferences, the strengths or the weaknesses of the hypertext reading environment designed for this study, and whether they experienced any difficulties (see Appendix K).

3.8 Procedures

Data collection started with the implementation of QPT exam and continued with one-on-one implementation of the Backward Digit Span task which took almost a month. Prior to reading, each group was given the prior knowledge and topic-interest questionnaires. According to the availability of the department lab, four groups took the treatment in different times within a period not surpassing a semester. The treatment was given at participants' self-pace. That is to say, there was no time restriction during reading or for the upcoming activities. As the aim was to elicit their free recall of the text, they were not allowed to take down notes or to look up words in the dictionary while reading. When they finished reading, a blank page came to screen on which they would type whatever they understood and remembered from the text in either English or Turkish. During the recall task, they had no access to the text. Following this, a multiple-choice comprehension test page came to the screen.

While answering these questions, they had access to the text at the same time. Upon the submission of the comprehension test, the participants took two consecutive vocabulary tests: a passive recall test and an active recognition test. The text was not available for the completion of these tests, either. Then, they completed Subjective Cognitive Load Scales. Three weeks after the treatment, all groups took the vocabulary tests again. Lastly, interviews were held with individual, pairs or groups of participants with the volunteers. Except for the semi-structured interviews, all the participants took all of the data collection instruments. During none of these data collection procedures, the participants were made explicit remarks in relation to the purposes of any of the tools so as to ensure the objectivity and neutrality of data collection.

3.9 Data analysis

The effects of gloss type (lexical versus topic-level), gloss position (pop-up versus separate window) and WM capacity (i.e., high versus low), on the dependent variables, namely, text recall, comprehension test, vocabulary measures, perceived CL, frequency of access to annotations and total reading time, were analyzed via quantitative methods of data analysis. First, descriptive statistics were obtained for all groups across all variables. Next, data were analyzed for normality of distribution and for outliers. Following the deletion of outliers, the sample size for each group was fixed to 30. Square root transformations were applied for the whole data set as a positively skewed distribution was observed in the raw scores. For the combined effects of WM capacity and the treatments, a three-way analysis of variance was conducted on the influence of three independent variables (WM capacity, gloss type, and gloss position) on the analyses of comprehension measures, namely free recall

and comprehension tests. WM capacity included two levels (high versus low), gloss type consisted of two levels (lexical versus topic-level), and gloss position contained two levels (pop-up window versus SW). As for the effects of WM capacity, treatments, and time, a 2 (WM capacity: high vs. low) X 2 (gloss type: lexical vs. topic-level) X 2 (gloss position: pop-up window vs. SW) X 2 (time: immediate vs. delayed) mixed ANOVA was run for both vocabulary recognition and vocabulary production measures. Independent samples and paired samples *t*-tests were employed to determine the level of significance for any statistically meaningful interaction. For determining the effects of different treatment conditions and the mediating role of WM capacity on the perceived CL, a Kruskal-Wallis *H* test and a Mann-Whitney *U* test were conducted on the difficulty and mental effort ratings for the reading process. Two separate three-way ANOVAs were computed to probe the effects WM capacity, gloss type, and gloss position on the frequency of access to glosses and the total reading time.

As regards the semi-structured interviews, audio-taped interviews were first transcribed and translated into English. The comparison of the Turkish and English versions of the transcribed interviews was shown to a colleague who is also a PhD candidate in the ELT department for expert opinions. Consensus was reached on the translated version by compromising on the disputed wordings. Next, they were qualitatively analyzed for any recurrent themes. The correspondence of emerging themes with the focuses of this study was sought for the categorization process. As a result, interview data were consulted to further explain any parts which needed elaboration, and also to shed light on readers' perceptions of electronic reading.

CHAPTER 4

RESULTS

This chapter will seek to present the findings obtained from both quantitative and qualitative analyses of data. Quantitative data which constitute the major part of the analyses come from the scores of text recall, text comprehension, vocabulary tests, topic interest and cognitive load scales, and log files. Qualitative data were obtained from the analyses of semi-structured interviews conducted with the volunteered participants. Descriptive statistics based on transformed scores were obtained for the quantitative measures. Inferential statistics were also conducted on the transformed scores to further interpret these results. After the elimination of outliers, sample size was fixed to 30 in each experimental condition; i.e., lexical pop-up window annotation group, lexical separate window annotation group, topic-level pop-up window annotation group, and topic-level separate window annotation group.

4.1 Effects of WM capacity, gloss type and gloss position on text recall and comprehension

4.1.1 Free recall

Descriptive statistics across the treatment conditions for the free recall task as measured through the recall protocols are provided in Table 3. It should be noted that the table presents the means and standard deviations based on the transformed scores.

Table 3. Descriptive Statistics for the Free Recall Task

Condition	*M*	*SD*	*SE*	*Min.*	*Max*
L-P	3.06	0.54	0.09	2.45	4.12
L-S	3.28	0.72	0.13	1.73	4.69
T-P	3.14	0.64	0.11	2.00	4.58
T-S	2.84	0.46	0.08	2.00	3.74

The participants who accessed lexical annotations in a separate window produced the best results on free recall while those who accessed topic-level annotations in a separate window had the lowest mean. A three-way between groups ANOVA with gloss type (lexical versus topic-level), gloss position (pop-up window versus SW), and WM capacity (high WM capacity versus low WM capacity) as between group factors was conducted on text recall. The assumption of equality of variances was accepted since Levene's test for equality of variances did not yield significant results for the present analysis (F = 1.2, p =.282). The results are summarized in Table 4.

Table 4. ANOVA Summary of Text Recall Scores

Source	*SS*	*df*	*MS*	*F*	*Sig.*	*Partial Eta²*
WM capacity	1.505	1	1.505	4.243	.042*	.036
Gloss Type (GT)	1.013	1	1.013	2.857	.094	
Gloss Position (GP)	.048	1	.048	.134	.715	
WM capacity X GT	.269	1	.269	.757	.386	
WM capacity X GP	.019	1	.019	.053	.818	
GT X GP	1.996	1	1.996	5.626	.019*	.048
WM capacity X GT X GP	7.477E-5	1	7.477E-5	.000	.988	
Error	39.726	112	.355			

*$p < .05$

The ANOVA results for text recall yielded a significant main effect only for WM capacity, displaying a statistically significant difference between high capacity (M = 3.19, SD = 0.64) and low capacity (M = 2.96, SD = 0.56) readers in their free recall scores regardless of the glossing condition. The only significant interaction was observed between gloss type and gloss position, which means that the effect of gloss type on readers' recall scores depended on the position of glosses as is shown in Figure 10. The other effects were not significant.

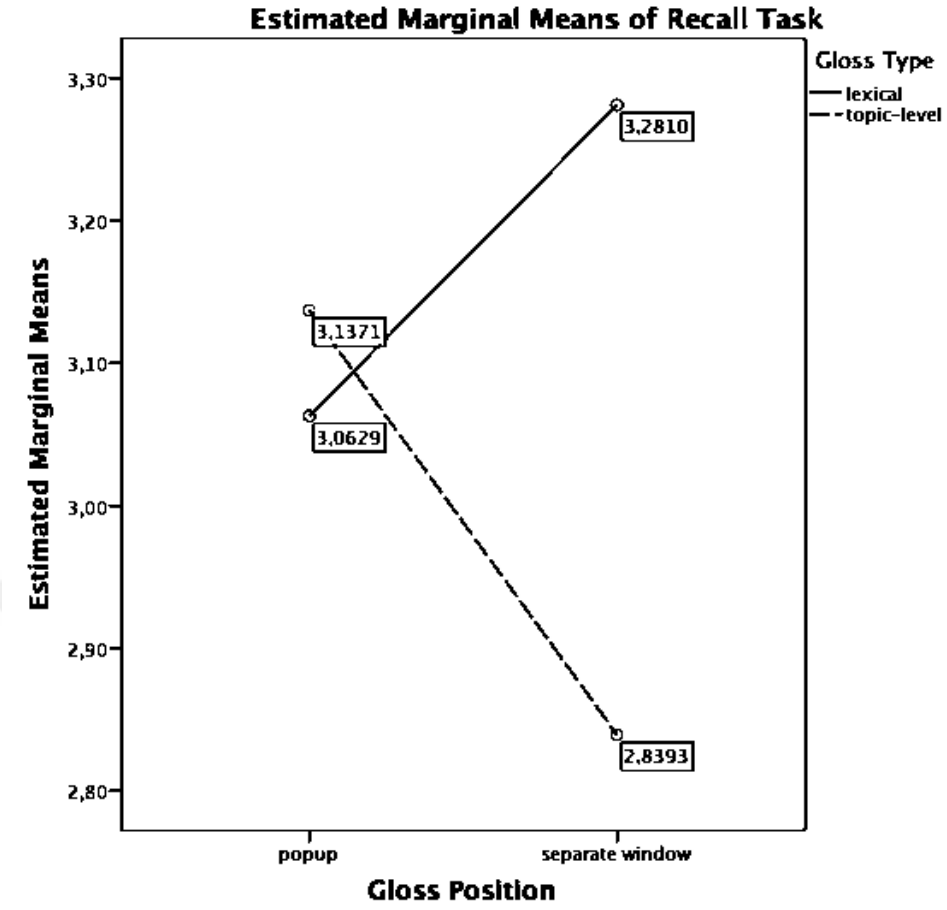


Figure 10. Interaction between gloss type and gloss position in terms of text recall

Figure 10 indicates that the pop-up window condition neutralizes the effect of gloss type. Specifically, independent samples *t*-tests revealed that the difference between the L-P and the T-P condition was not significant ($p > .05$). On the other hand, the participants in the L-S condition generated significantly more propositions on the recall task ($M = 3.28$, $SD = .72$) than those in the T-S condition ($M = 2.84$, $SD = .46$), $t(58) = 2.829$, $p = .007$, $d = 0.87$. Additionally, while the L-P and L-S conditions did not lead to significantly different number of propositions ($p > .05$), there existed a significant difference between the T-P ($M = 3.14$, $SD = .64$), and the T-S conditions ($M = 2.84$, $SD = .46$), $t(29) = 2.828$, $p = .042$, $d = 0.53$. Overall, these results suggest that topic-level annotations, when presented in a separate window, led to significantly fewer recall of ideas compared to the other three conditions.

4.1.2 Comprehension

Apart from recall, the participants' comprehension of the text was tested via a fourteen-item reading comprehension test. Descriptive statistics based on the transformed scores are provided in Table 5.

Table 5. Descriptive Statistics for the Comprehension Test

Condition	*M*	*SD*	*SE*	*Min.*	*Max*
L-P	2.88	0.35	0.06	2	3.46
L-S	2.79	0.3	0.05	2.24	3.32
T-P	3.01	0.47	0.08	2	3.74
T-S	2.96	0.26	0.04	2.24	3.46

Different from the recall task, topic-level annotations in a pop-up window condition yielded the best scores for the comprehension test. A three-way between groups ANOVA was conducted to determine if there existed an interaction between WM capacity, gloss type, and gloss position in terms of comprehension.

Table 6. ANOVA Summary of Comprehension Test Scores

Source	*SS*	*df*	*MS*	*F*	*Sig.*	*Partial Eta²*
WM capacity	21.67	1	21.67	5.70	.019*	.048
Gloss Type (GT)	25.21	1	25.21	6.62	.011*	.056
Gloss Position (GP)	8.00	1	8	2.11	.150	
WM capacity X GT	1.41	1	1.41	.37	.544	
WM capacity X GP	20.01	1	20.01	5.26	.024*	.045
GT X GP	.075	1	.075	.020	.889	
WM capacity X GT X GP	6.075	1	6.075	1.60	.209	
Error	426.13	112	3.81			

*$p < .05$

The homogeneity of variances was assumed for each combination of groups since the calculated p-value in Levene's test for homogeneity of variances was bigger than the significance level (p= .05). As with the text recall, WM capacity had a significant main effect on text comprehension. Accordingly, comprehension scores of high capacity readers (M = 2.99, SD = 0.36) were significantly higher than those of low capacity readers (M = 2.84, SD = 0.34). The main effect of gloss type tells us that if we ignore all the other factors, there is a significant difference in comprehension scores between lexical (M = 2.83, SD = 0.33) and topic-level (M = 2.99, SD = 0.37)

annotation groups, with the latter outperforming the former. The only interaction detected was between WM capacity and gloss position (Figure 11).

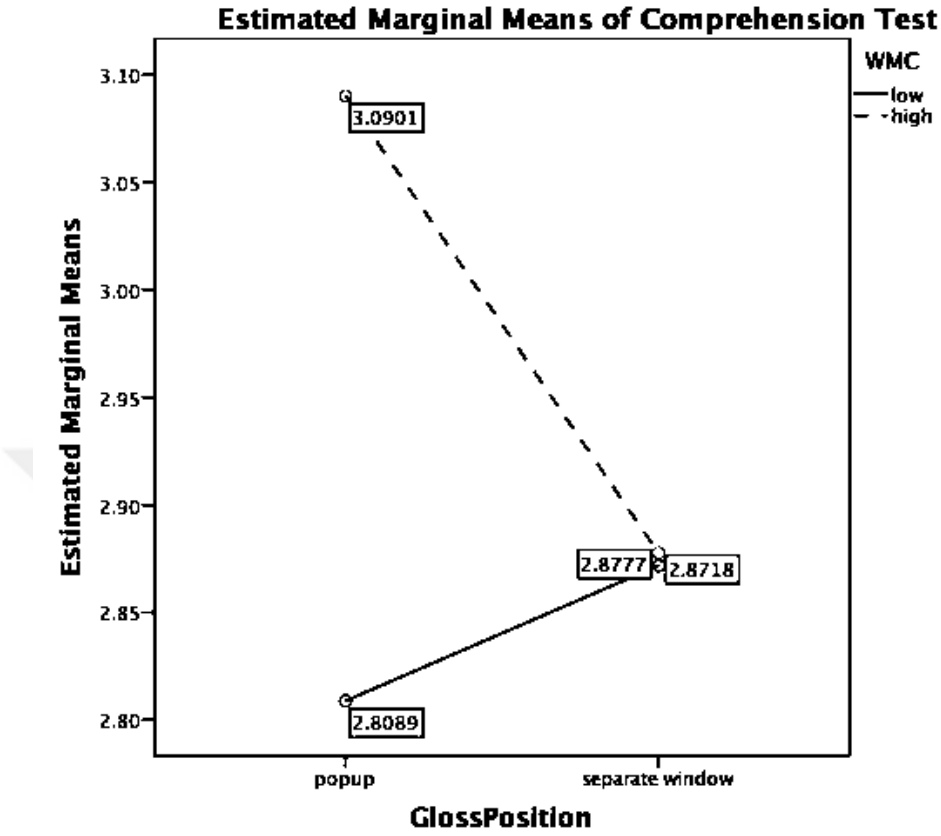


Figure 11. Interaction between WM capacity and gloss position in terms of comprehension

As shown in Figure 11, high capacity readers outperformed their low capacity counterparts in the pop-up condition. Whether this is a significant difference or not was further investigated via independent samples *t*-tests. The results depicted a statistically significant difference in the pop-window group in that high capacity readers (M = 3.09, SD = 0.39) scored significantly better than low capacity readers (M = 2.81, SD = 0.38) in the same annotation format, $t(58) = 2.79$, $p = .007$, $d = 0.72$. In addition, to show the variations from pop-up to separate window conditions among low WM and high WM participants, additional independent samples *t*-tests were conducted. While the difference among low WM participants was not significant, $t(58) = .707$, $p > .05$, the difference between the pop-up group (M = 3.09,

SD = 0.39) and the separate window group (*M* = 2.87 *SD* = 0.29) for high WM participants was significant, $t(58) = 2.35$, $p = .022$, $d = 0.64$. Overall, these results suggest that while the performance of the low capacity participants was similar between the pop-up and separate-window conditions, the high capacity participants' performance decreased significantly from the pop-up to the SW condition.

4.2 Effects of WM capacity, gloss type, gloss position and time on incidental vocabulary learning

The participants' vocabulary gains and retention were measured with the help of two measures (vocabulary matching and vocabulary production) tapping recognition and production performance respectively on two different occasions (immediate and delayed vocabulary post-tests).

4.2.1 Vocabulary recognition

A fourteen-item vocabulary-matching task was used to test vocabulary recognition. Descriptive statistics regarding this task are displayed in Table 7 which indicate that the group who accessed lexical annotations in a separate window was able to match most number of items on average.

Table 7. Descriptive Statistics for Post and Delayed Matching Test

	Condition	*M*	*SD*	*SE*	*Min.*	*Max*
Post-test	L-P	7.85	1.29	0.23	3.78	10
	L-S	8.27	0.92	0.17	5.98	9.64
	T-P	7.76	0.87	0.16	5.98	10
	T-S	7.02	1.62	0.3	3.78	9.26
Delayed Test	L-P	7.18	1.03	0.19	4.63	9.26
	L-S	7.15	1.23	0.23	4.63	9.26
	T-P	7.51	1.57	0.29	3.78	10
	T-S	6.65	1.73	0.32	3.78	9.64

A four-way mixed ANOVA with gloss type (lexical versus topic-level), gloss position (pop-up window versus separate window), and WM capacity (high WM

capacity versus low WM capacity), as between group factors and time (immediate versus delayed) as repeated measures factor was conducted on vocabulary recognition tests (see Table 8).

Table 8. ANOVA Summary of Vocabulary Recognition Scores

Source	*SS*	*df*	*MS*	*F*	*Partial Eta*^{2}
Between		119			
Gloss Type (GT)	8.61	1	8.61	3.49	
Gloss Position (GP)	5.46	1	5.46	2.22	
WM capacity	6.76	1	6.76	2.75	
GT X GP	14.76	1	14.76	5.99*	.051
GT X WM capacity	.020	1	.020	.008	
GP X WM capacity	.075	1	.075	.030	
GT X GP X WM capacity	4.64	1	4.64	1.88	
Error	275.92	112	2.46		
Within		120			
Time	21.94	1	21.94	20.87**	.157
Time X GT	5.15	1	5.15	4.90*	.042
Time X GP	1.24	1	1.24	1.18	
Time X WM capacity	.365	1	.365	.347	
Time X GT X GP	.392	1	.392	.373	
Time X GT X WM capacity	.206	1	.206	.196	
Time X GP X WM capacity	.095	1	.095	.091	
Time X GT X GP X WM capacity	.370	1	.370	.352	
Error (time)	117.77	112	1.051		

*$p < .05$. **$p < .01$.

The only significant main effect was that of time; that is the overall mean of the immediate matching task (M= 7.73, SD = 1.28) was significantly higher than that of the delayed matching task (M= 7.12, SD = 1.46). No significant four-way or three-way interactions were found. The interactions observed were two-way; one being between gloss type and gloss position, the other being between time and gloss type. The interaction between gloss type and gloss position is shown in Figure 12.

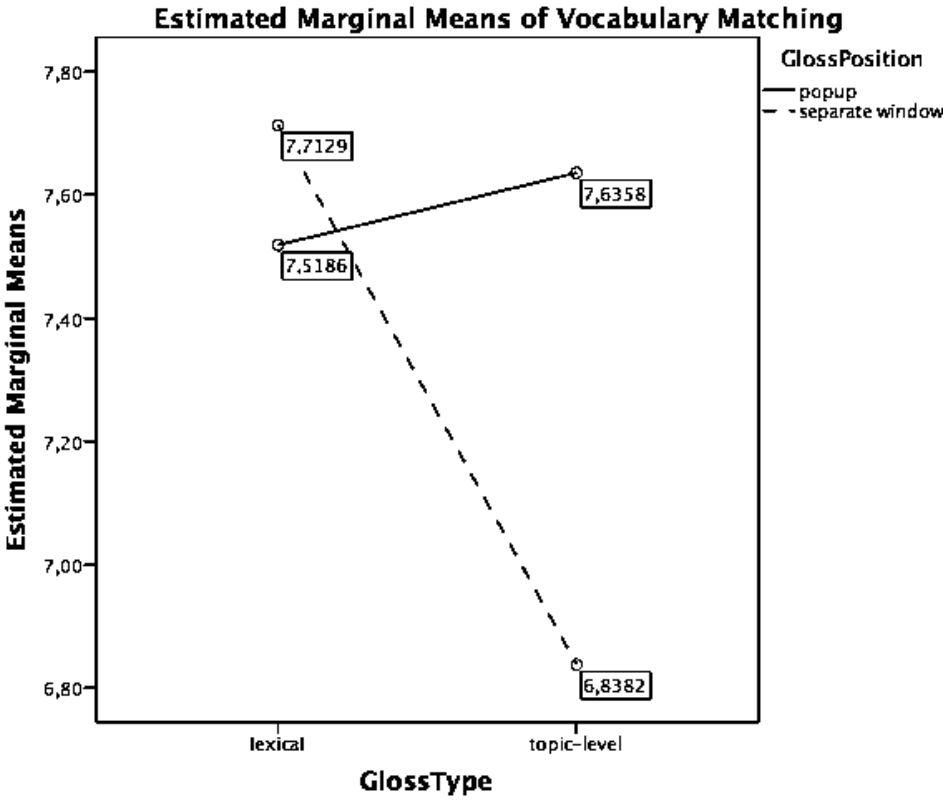


Figure 12. Interaction between gloss type and gloss position in terms of vocabulary recognition

Figure 12 demonstrates a distinct gap between the L-S and T-S conditions. An independent samples *t*-test was computed to understand whether the difference between these groups is significant, and it was observed that in the separate window conditions, lexical group performed significantly better than the topic-level group on vocabulary recognition tests (M = 7.71, SD = 0.83; M = 6.84, SD = 1.43 respectively), $t(58)=2.90$, $p=.006$, d = 0.75. Additional *t*-tests were also conducted to see whether the differences within the lexical groups and topic-level groups are significant. The results demonstrated that the performance of the participants who accessed lexical annotations did not differ depending on gloss position ($p > .05$). However, those who accessed topic-level annotations in a pop-up window outperformed those who saw the same annotations in a separate window, (M = 7.64, SD = 1.07; M = 6.84, SD = 1.43 respectively), $t(58)=2.45$, $p=.017$, d = 0.63.

The second two-way interaction which is between time and gloss type is illustrated in Figure 13.

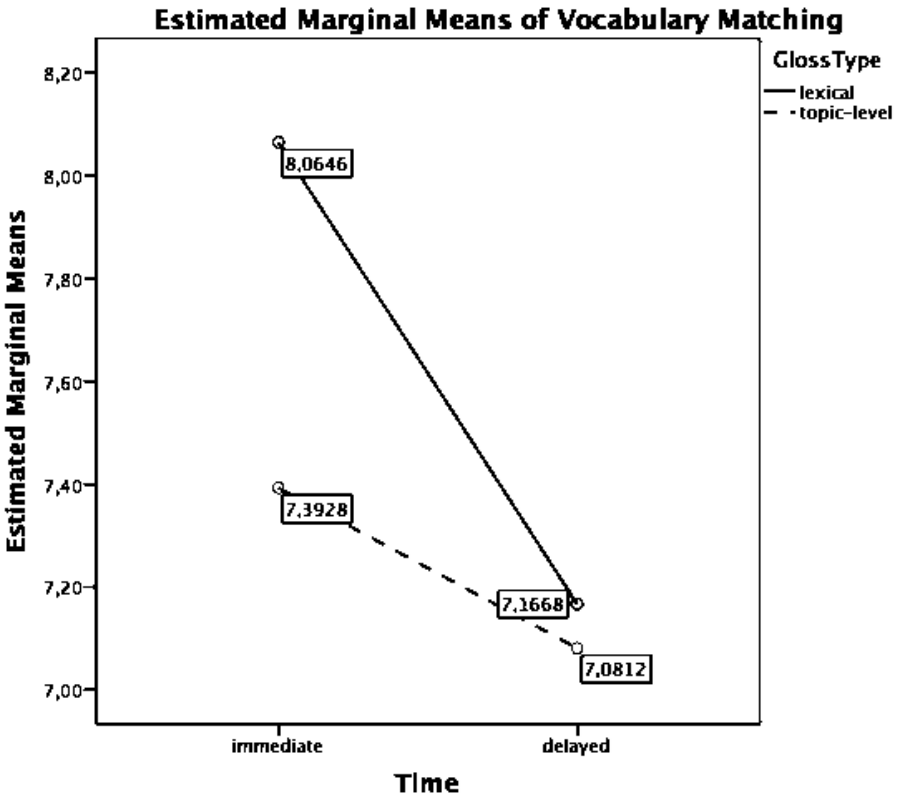


Figure 13. Interaction between gloss type and time in terms of vocabulary recognition

The figure shows that while there is not much variance between lexical and topic-level annotation groups in terms of the delayed post-test scores, there appears to be a vivid difference between these two groups on the immediate post-test. Independent samples *t*-tests indicated that the lexical annotation group had a significantly higher mean (M = 8.06; SD = 1.14) than the topic-level annotation group (M = 7.39; SD = 1.34) on the immediate vocabulary recognition test, $t(118) = 2.96$, $p <. 05$, $d = 0.53$. However, this difference cannot be observed on the delayed vocabulary recognition test ($p >.05$).

4.2.2 Vocabulary production

As for production scores, the descriptive statistics for the immediate and delayed post-test are presented in Table 9.

Table 9. Descriptive Statistics for Post-test and Delayed Production Test

	Condition	*M*	*SD*	*SE*	*Min.*	*Max*
Post-test	L-P	6.92	1.58	0.29	3.27	9.82
	L-S	7.54	1.56	0.28	5	10
	T-P	6.3	1.26	0.23	3.78	9.82
	T-S	5.7	1	0.18	3.27	7.07
Delayed Test	L-P	5.87	1.12	0.2	3.78	8.02
	L-S	5.47	1.46	0.27	3.27	9.26
	T-P	5.61	1.5	0.27	3.27	9.26
	T-S	5.52	1.2	0.22	3.27	8.02

Based on the results of the descriptive analyses, the L-S group seems to outperform the other groups in the post-test whereas the L-P group performed best in the delayed test. A 2 (high vs. low WM capacity) X 2 (lexical vs. topic-level annotation) X 2 (pop-up window vs. separate window annotation) X 2 (immediate vs. delayed post-test) factorial ANOVA yielded significant main effects of time, gloss type, and WM capacity (see Table 10).

Table 10. ANOVA Summary of Vocabulary Production Scores

Source	*SS*	*df*	*MS*	*F*	*Partial Eta*^{2}
Between		119			
Gloss Type (GT)	26.44	1	25.44	9.74*	.080
Gloss Position (GP)	.838	1	.838	.309	
WM capacity	11.29	1	11.29	4.16*	.036
GT X GP	3.06	1	3.06	1.13	
GT X WM capacity	1.10	1	1.10	.407	
GP X WM capacity	2.25	1	2.25	.830	
GT X GP X WM capacity	2.40	1	2.40	.883	
Error	303.96	112	2.71		
Within		120			
Time	59.6	1	59.6	68.31**	.379
Time X GT	18.72	1	18.72	21.34**	.160
Time X GP	.959	1	.959	1.09	
Time X WM capacity	.295	1	.295	.337	
Time X GT X GP	8.7	1	8.7	9.91*	.081
Time X GT X WM capacity	.641	1	.641	.731	
Time X GP X WM capacity	4.49	1	4.49	5.11*	.044
Time X GT X GP X WM capacity	.802	1	.802	.914	
Error (time)	98.244	112	.877		

*$p < .05$. **$p < .01$.

The main effect of gloss type suggests that the lexical annotation group (M = 7.23, SD = 1.59) had significantly higher scores than the topic-level annotation group (M = 6.00, SD = 1.16). The main effect of WM capacity indicates that high WM participants (M = 6.87, SD = 1.51) had higher productive vocabulary scores compared to low WM participants (M = 6.36, SD = 1.49). Lastly, the significant main effect of time displays that the mean scores of the immediate vocabulary production test were significantly higher (M = 6.62, SD = 1.52) than those of the delayed vocabulary production test (M = 5.62, SD = 1.33). Additionally, there is a lower-order interaction between time and gloss type (Figure 14).

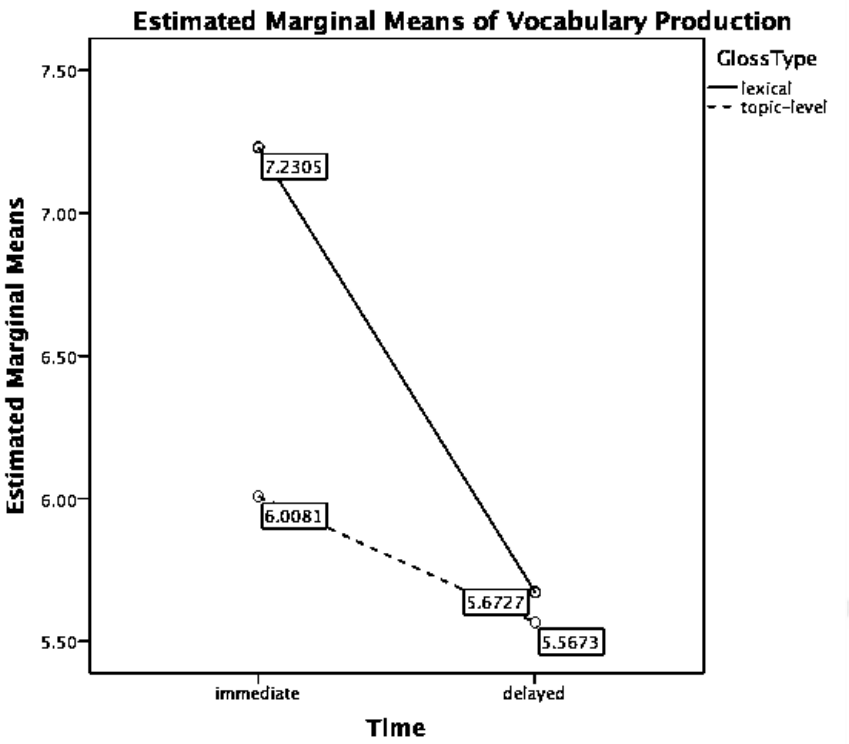


Figure 14. Interaction between time and gloss type in terms of vocabulary production

The figure demonstrates that there is a decline in the scores of both lexical and topic-level groups from the immediate post-test to the delayed post-test, but still there is a difference between these two groups. While the gap between the lexical and the topic-level groups was large for the post-test, it decreases in the delayed-test. To further investigate this significant interaction, an independent-samples *t*-test was computed. Results indicated that the scores of lexical groups were significantly higher (M= 7.23, SD = 1.59) than those of the topic-level groups (M= 6.00, SD = 1.16) on the immediate post-test, $t(118)= 4.8$, $p < .001$, $d = 0.88$ whereas there is not a significant difference between them on the delayed post-test ($p > .05$).

As Table 10 shows, there are two higher order interactions which were between time, gloss type, and gloss position, and between time, gloss position, and WM capacity. The first three-way interaction between gloss type, gloss position, and time is displayed in Figures 15 and 16.

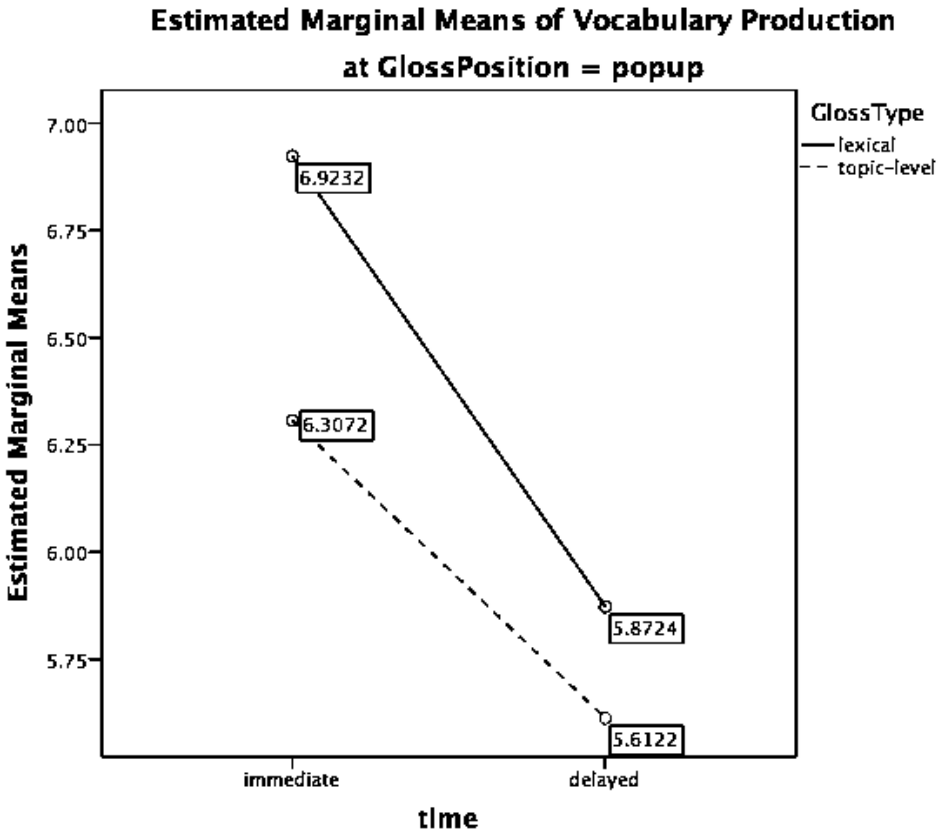


Figure 15. Interaction between time and gloss type for pop-up window condition

It is evident in Figure 15 that in the pop-up condition, the immediate post-test mean of both the lexical (M_{LP} = 6.92, SD = 1.5) and topic-level (M = 6.31, SD = 1.26) annotation groups were higher than their delayed test means ($M_{LP\text{-}delayed}$ = 5.87, $SD_{LP\text{-}delayed}$ = 1.13; $M_{TP\text{-}delayed}$ = 5.61, $SD_{TP\text{-}delayed}$ = 1.51). The paired samples t-test indicated that this was a significant decrease for both the lexical, $t(29) = 4.78$, $p < .001$, $d = 0.76$ and the topic-level annotation group, $t(29) = 3.42$, $p < .05$, $d = 0.57$. On the other hand, the independent samples t-tests yielded that the difference between L-P and T-P groups was not significant on either the immediate test or the delayed test ($p > .05$).

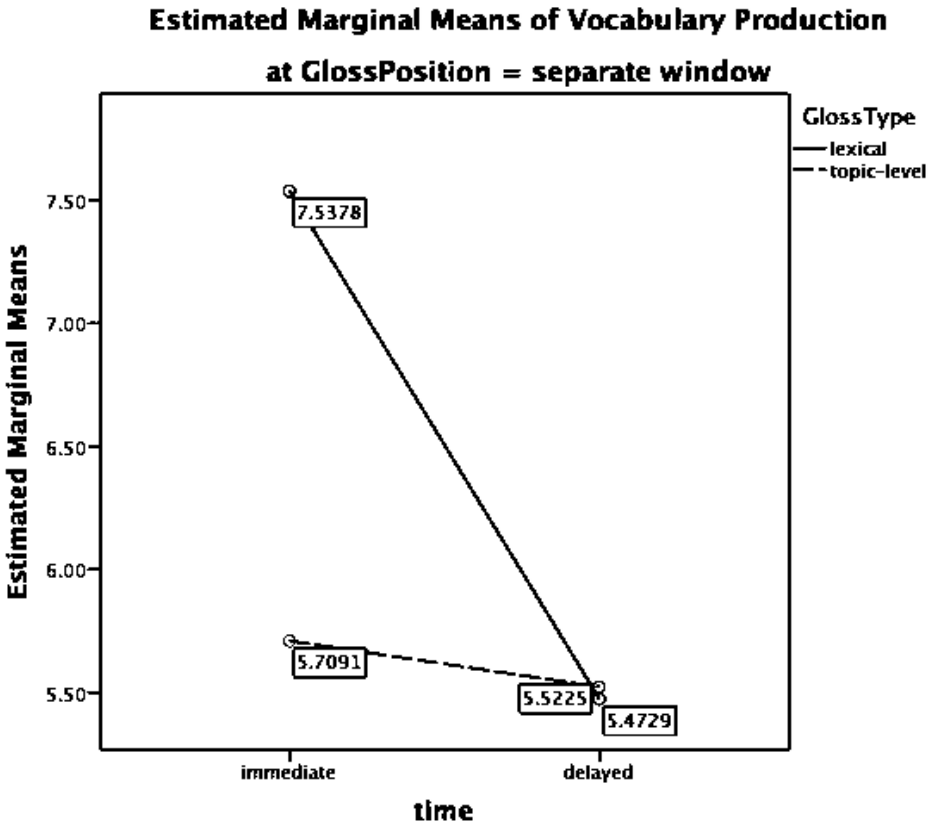


Figure 16. Interaction between time and gloss type for separate window condition

Figure 16 figure depicts that, in the separate window condition, the lexical annotation group (M = 7.54, SD = 1.56) outperformed the topic-level annotation group (M = 5.71, SD = 1.01) on the immediate test, t (58)=5.4, $p < .001$, d = 1.39 but not on the delayed test, t(58)= -.049, $p > .05$. In addition, the production scores of the former group decreased significantly from the immediate to the delayed test, $t(29)$ = 6.28, $p < .001$, d = 1.37 while that of the topic-level annotation group did not change, t (29) = .903 $p > .05$.

Another three-way interaction was observed between time, WM capacity, and gloss position (see Figures 17 and 18).

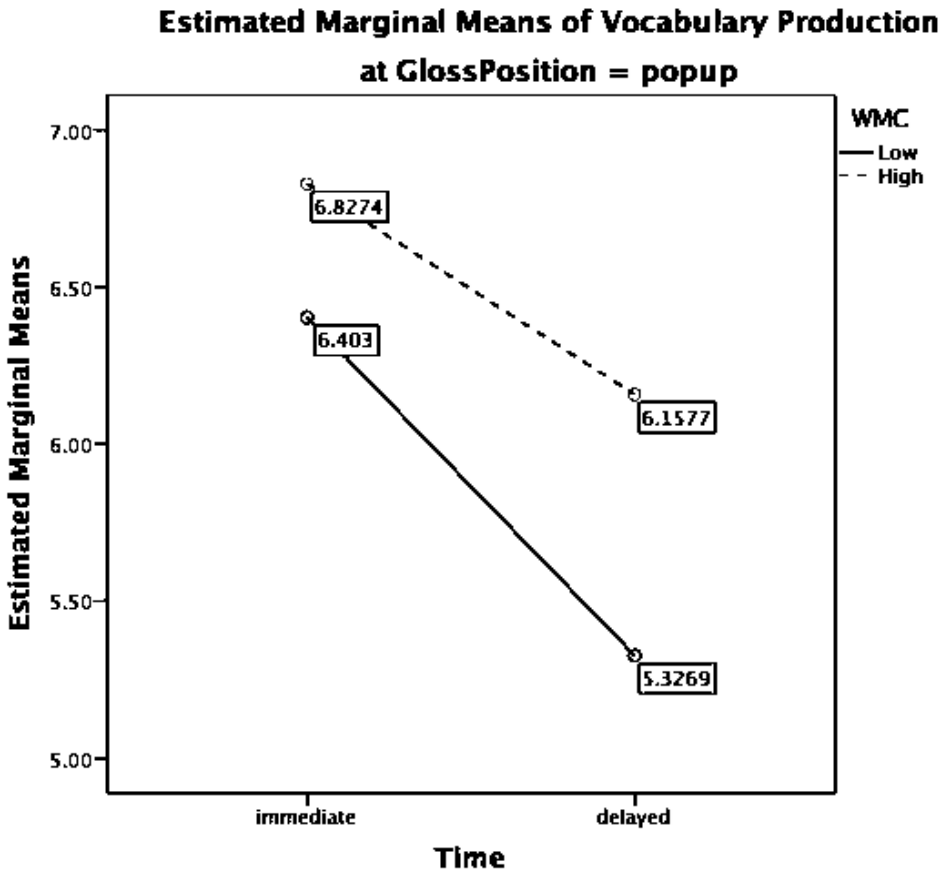


Figure 17. Interaction between time and WM capacity for pop-up window condition

Figure 17 depicts an overall difference between high capacity and low capacity readers in the pop-up window condition at both times of testing. A closer look with *t*-test analyses revealed a significant difference between low-WM (M = 5.32, SD = 1.14) and high-WM (M = 6.15, SD = 1.38) participants on the delayed test, $t(58) = 2.53$, $p = .014$, $d = 0.65$ but not on the immediate test ($Ml_{owWM} = 6.40$, $SD_{lowWM} = 1.51$; $M_{highWM} = 6.82$, $SD_{highWM} = 1.37$) for the pop-up condition. In addition, paired samples t-tests indicated that within the pop-up condition both low- and high-WM participants had significantly lower scores on the delayed test compared to the immediate post-test, $t(29) = 3.94$, $p < .001$, $d = 0.71$ for the low-WM group; $t(29) = 5.55$, $p < .001$, $d = 1.01$ for the high-WM group.

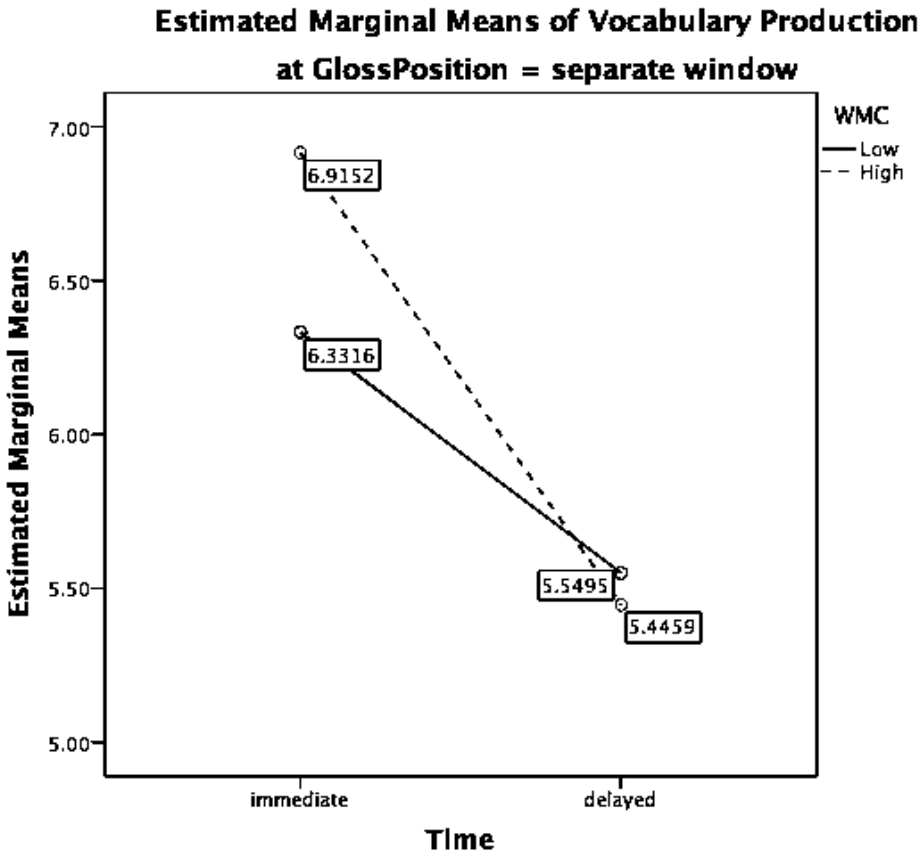


Figure 18. Interaction between time and WM capacity for separate window condition

For the separate window condition, Figure 18 demonstrates a distinct decline for both WM capacity groups from the immediate test to the delayed-test. Paired samples *t*-tests also revealed that the post-test sores of both low-WM (M = 6.33, SD = 1.49) and high-WM participants (M = 6.91, SD = 1.65) were significantly higher compared to the delayed scores (M_{lowWM} = 5.54, SD_{lowWM} = 1.46; M_{highWM} = 5.44, SD_{highWM} = 1.2), $t_{lowWM}(29) = 2.39$, $p = .024$, $d = 0.43$, and $t_{highWM}(29) = 4.73$, $p = .000$, $d = 0.86$. On the other hand, independent samples *t*-tests did not yield significant differences between the low- and high-WM participants either on the immediate post-test or on the delayed test ($p > .05$) in the separate window condition.

Altogether, these findings suggest that, in the pop-up condition, which supposedly induces lower CL, readers with high capacity were able to retain more

words from the text compared to the low capacity readers. On the other hand, in the separate window condition with higher associated CL, the groups performed similarly both on the immediate and delayed tests.

4.3 Effects of WM capacity, gloss type and gloss position on the perceived cognitive load and the recorded annotation use

The participants' ratings on the two sub-dimensions of the Subjective Cognitive Load Scales, namely the Difficulty and Mental Effort (ME), were analyzed. Means and standard deviations across the groups regarding the difficulty and ME ratings on the reading passage are provided in Table 11.

Table 11. Descriptive Statistics of the Perceived Cognitive Load

Tasks	Lexical				Topic-level			
	Pop-up		Separate		Pop-up		Separate	
	M	SD	M	SD	M	SD	M	SD
Difficulty	3.43	1.70	4.2	1.69	4.53	1.52	3.9	1.49
ME	5.33	1.77	5.3	1.68	5.2	1.74	5.76	1.25

It is evident in the table that the difficulty level reported by the T-P group is greater than the rest of the groups while the closest follower is the L-S group. Besides, it is apparent that there are no great differences across the groups regarding the ME ratings except for the highest rating of the T-S group. To further probe whether the differences across the groups are significant, two Kruskal-Wallis ANOVAs were conducted. No differences between groups were found for either difficulty ratings (Kruskall-Wallis $H = 6.83$, $df = 3$, $p = 0.07$) or mental effort ratings (Kruskall-Wallis $H = 2.62$, $df = 3$, $p = 0.45$). To probe the WM capacity effect, the Mann-Whitney U tests were computed; however, the difference between the low capacity and high capacity participants was not significant for either scales again (Mann-Whitney $U_{\text{difficulty}} = 1527$, $p > .05$; Mann-Whitney $U_{\text{mentalEffort}} = 1648$, $p > .05$). Interestingly, based on the self-reports of the participants, the perceived difficulty of reading the

text or the mental effort invested during reading did not change according to either WM capacity or the electronic reading conditions manipulated in this study.

The participants' use of glosses were investigated in terms of two aspects, namely the frequency of access to glosses and the total amount of time spent on reading (i.e., time-on-task) which were recorded through a built-in tracking software. Table 12 presents frequency of access in percentage and total reading time in seconds. Since the number of word-level (28 in total) and topic-level (12 in total) annotations were different, raw frequencies were converted into percentages by dividing the total number of times each participant accessed annotations by the total number of annotations available in a given condition and multiplying this ratio with 100. It should be noted that the average frequency could exceed 100 since the participants were allowed to access glosses more than once.

Table 12. Descriptive Statistics of the Access to Glosses and Reading Time

		Access Frequency			Time on Task		
WM	*Condition*	*M*	*SD*	*N*	*M*	*SD*	*N*
1	L-P	103.57	54.14	15	1566.93	465.16	15
	L-S	86.66	36.95	15	1833.93	744.14	15
	T-P	80.00	48.77	15	1616.20	384.56	15
	T-S	75.55	64.04	15	1326.53	584.50	15
	TOTAL	86.44	51.69	60	1585.90	576.16	60
2	L-P	98.09	57.94	15	1685.87	538.36	15
	L-S	59.28	42.48	15	1650.87	768.33	15
	T-P	98.88	56.68	15	1428.80	406.88	15
	T-S	63.88	48.04	15	1488.60	404.98	15
	TOTAL	80.03	53.70	60	1563.53	546.66	60
	L-P TOTAL	100.83	55.17	30	1626.40	498.03	30
	L-S TOTAL	72.97	41.53	30	1742.40	748.98	30
	T-P TOTAL	89.44	52.83	30	1522.50	400.50	30
	T-S TOTAL	69.72	55.94	30	1407.57	500.90	30

Two independent three way ANOVAs with WM capacity (low vs. high), gloss type (lexical vs. topic-level) and gloss position (pop-up vs. SW) as between group factors were conducted on the frequency of access (see Table 13) and total reading time (see Table 14).

Table 13. ANOVA Summary of Frequency of Access to Glosses

Source	*SS*	*df*	*MS*	*F*	*Sig.*	*Partial Eta²*
WM capacity	1232.15	1	1232.15	.459	0.499	
Gloss Type (GT)	1608.10	1	1608.10	.599	.440	
Gloss Position (GP)	16978.5	1	16978.5	6.328	.013*	.053
WM capacity X GT	3011.92	1	3011.92	1123	.292	
WM capacity X GP	5160.16	1	5160.16	1923	.168	
GT X GP	496.32	1	496.32	.185	.668	
WM capacity X GT X GP	140.31	1	140.31	.052	.820	
Error	300480	112	2682.86			

*$p < .05$

The table demonstrates that only the main effect of gloss position is significant which shows that the overall mean for the pop-up conditions (M = 95.13, SD = 6.68) is significantly higher than that of the separate window conditions (M = 71.34 SD = 6.68). The ANOVA results did not yield any significant interaction for this analysis, though. Taken altogether, these suggest that the participants in the integrated-format accessed the glossed words more frequently than the participants in the split-format.

Table 14. ANOVA Summary of Total Reading Time

Source	*SS*	*df*	*MS*	*F*	*Sig.*	*Partial Eta²*
WM capacity	15008	1	15008	.049	826	
Gloss Type (GT)	1.44E+6	1	1.44E+6	4.677	.033*	.04
Gloss Position (GP)	8.533	1	8.533	.000	.996	
WM capacity X GT	2822.7	1	2822.7	.009	.924	
WM capacity X GP	4224.53	1	4224.53	.014	.907	
GT X GP	399977	1	399977	1.296	.257	
WM capacity X GT X GP	759767	1	759767	2.578	.111	
Error	3.46E+7	112	308675			

*$p < .05$

ANOVA results for the total time spent on reading did not produce any significant interactions. There is only one significant main effect which is of the gloss type. Accordingly, the participants accessing lexical annotations (M= 1684.40, SD= 71.72) spent significantly more time for reading the text than the participants accessing topic-level annotations (M= 1465.03, SD= 71.72)

Pearson product-moment correlations were obtained to examine whether the frequency of access to annotations and total reading time were related to reading comprehension and vocabulary learning for each glossing condition. For the L-S condition, there is a modest negative correlation between access frequency and the immediate vocabulary recognition task (r= - 40, $p < .05$) which indicated that as the participants reached more split-format annotations, their vocabulary gain scores decreased. In addition, as the amount of time spent on reading increased, the recall scores of the participants increased (r = 47, $p < .05$). In the T-S condition, access frequency correlated significantly with the comprehension scores (r= 43, $p < .05$). Moreover, total reading time yielded a modest significant positive correlation with the comprehension test (r = 37, $p < .05$). No significant correlations were detected for the L-P and T-P conditions.

4.4 Qualitative results

The last research question concerned the participants' perceptions of the electronic text format as to the nature of reading, its effects on recall, comprehension, and vocabulary tasks, and to the cognitive demands associated. The interviews were analyzed in the light of the research questions, and emerging themes were added into the analysis. Unanimity of feedback and opinions was sought at first to provide the general picture followed by more idiosyncratic but related comments.

Firstly, regarding the effects of gloss type and position on their understanding of the text and vocabulary learning, the participants' opinions varied. Although most of them thought that the annotations were user-friendly (since they can immediately access them) others (especially in the topic-level groups) confessed that they did not use the annotations much except for the ones which aroused curiosity. Whereas the

lexical annotation groups reported consulting annotations frequently to access the meanings of unfamiliar words, the topic-level annotation groups indicated that they wanted to view the annotations only because they were informative. Yet, they thought although the content of topic-level annotations included interesting facts/ information, they only helped the reading process, not the upcoming vocabulary tasks. In a way, as ELT students and prospective teachers, they considered that those annotations were given as extra information and not essential for the completion of some tasks (i.e., vocabulary tasks).

> I liked the ease the pop-window vocabulary annotations provided because in some other cases where the glosses are given at the bottom of the page or at the end of the text, we still have to look up them, but this format was more practical and time-saving.
>
> (L-P group, paired interview)

> Vocabulary annotations were useful and essential especially in the short term for reading and completing the tasks. Since the new words were annotated, the text was not difficult for me to understand. But if we don't use those words, we will for sure forget them in the long run.
>
> (L-P group, paired interview)

> Vocabulary annotations were useful in terms of having quick access to word meanings while reading, and helped me especially complete the vocabulary tasks. Otherwise, in some cases though we recognize the word, we couldn't write down the definitions, so vocabulary annotations helped especially at this point (*during the productive vocabulary task).*
>
> (L-S group, paired interview)

> Annotations were easy to use and did not require extra effort. And it was easy to integrate them into the text content because it was easy to see the connection. Most of the content in pop-up windows was new information for us and helped us understand the text better without requiring extra effort. Since we took those topical annotations as examples of the content covered in the text, they facilitated our completion of the upcoming tasks and increased the text comprehension.
>
> (T-P group, focus-group interview)

> I didn't use those annotations much, actually. I rarely opened them. I opened especially the ones which appealed my attention. But I immediately forgot them afterwards. They did not help my understanding of the text or the completion of other tasks.
>
> (T-S group, individual interview)

Actually, it seems that lexical annotation groups mostly emphasized the facilitative effects of annotations for vocabulary learning whereas contextual annotation groups focused on the facilitative effects on comprehension. This finding, in part, corroborates some of the quantitative findings presented above. That topic-level annotations enabled better comprehension while lexical annotations facilitated incidental learning of words is in line with these ideas derived from the interviews. Furthermore, topic-level annotation groups seemed to get as much help from the navigation map as the annotations for constructing the mental model of the text:

> The subheadings were really useful. By remembering those subheadings while writing the summaries, I recalled what each part contained. Before hitting the "finish" button, I went over them again. Then, while writing the summaries, I remembered the content under each subheading.
>
> (T-P group, focus group interview)

> The navigation map helped us a lot. While reading, when I sometimes wanted to turn to the previous parts, I directly looked at those subsection headings. Or while reading, when I checked the subtitle, I understood which part I was currently reading. While writing the summaries, I directly built my summary based on those sections.
>
> (T-S group, focus group interview)

In terms of the perceived difficulty of the electronic text they read, most of them thought that the text was appropriate for their level except for some unknown vocabularies which were especially pronounced by participants in topic-level gloss conditions. Select comments from this group applicable to this kind are as follows:

> The word-choice in the text was heavy. Since the senior year in high school, I have not encountered so many new words. That's why I have forgotten a lot of words that I knew before. I recognized them, but I couldn't remember their meanings.
>
> (T-P group, focus group interview)

> We did not notice those new words while reading. But when we encountered them in the vocabulary tests, we noticed them. I was able to guess the meaning from the content while reading, but when I was asked to supply their definitions, it was hard. We used to have such comprehensive vocabulary knowledge four years ago in the last year of high school while we were getting prepared for the university entrance exam.
>
> (T-P group, focus group interview)

> The text was not difficult, but just some unknown words were hard. We paused reading to try to guess their meanings. There were some new words for me.
>
> (T-S group, paired interview)

> The text was not difficult, but since there were some new words, I don't think that I completely understood it.
>
> (T-P group, focus group interview)

Based on these responses, it can be asserted that the lack of access to word meanings turned this electronic reading experience into a more difficult process especially for the readers in the topic-level annotation groups. Although the perceived CL did not change significantly across the group, and the reading span for the topic-level groups was much shorter than that of the lexical groups, they reported having difficulties in relation to the number of unknown words. In a way this suggests that spending less time on a task does not always indicate lower levels of mental effort or higher mental efficiency.

Topic interest is taken constant in this study since the scale did not yield significant differences across the groups, and it produced fairly high scores. Yet, the interview findings revealed contradictory information in that they expressed that they did not normally enjoy reading on this topic (i.e., history of civilizations). They reported that in general they were not interested in history or civilizations, per se, but the text prompted them to go on reading. Based on the distinction between individual interest versus situational interest, this type of emerging arousal at the time of reading can be taken under situational interest. Hence, it can be argued that although

these students do not possess individual interest towards the topic of civilizations due to the inherent characteristics of the text, they developed situational interest. Overall, the interviewees were of the opinion that the text was intrinsically interesting and motivated them to read till the end. Excerpts from the interviews portraying this effect are as follows:

> I was not, actually, interested in this topic, but the text attracted my attention. It was appealing as a topic and in terms of content, it was interesting.
> (L-P Group, paired interview)

> I am normally not interested in this topic, but I did not get bored while reading the text. It somehow attracted my interest.
> (T-P group, individual interview)

> Although there were some new words, we did not experience much trouble guessing their meanings because the text was very coherent and interesting. But I think the topic interest changes from person to person, history does not arouse my interest… *(Another student interrupted at this point and said)* I also don't like texts about history, but I did not get bored while reading this passage; the text was not boring compared to other texts on history.
> (T-S group, focus group interview)

Based on these insights from the interviews, it can be said that they enjoyed this electronic reading experience thanks to immediate access to glosses and its easy to navigate layout. They think that topic-level annotations and the presence of section sub-headings helped them understand and reproduce the text while lexical annotations facilitated word-learning. Moreover, they found the topic interesting although they do not normally like reading texts about history or civilizations.

4.5 General findings

Following the inferential analyses, the research hypotheses and the obtained results pertaining to those particular hypotheses are provided in Table 15.

Table 15. Summary of Research Hypotheses, Main Findings, and Conclusions

HYPOTHESES		FINDINGS	CONCLUSIONS
H1	L = T in terms of recall	Nonsignificant main effect of gloss type	Confirmed
H2	T > L in terms of comprehension	Significant main effect of gloss type	Confirmed
H3	L > T in terms of vocab recognition & production	The facilitative effect of lexical glosses was observed both on recognition and production. But in terms of immediate gains but not delayed gains.	Partially confirmed
H4	T-S < T-P=L-S= L-P in terms of recall, comprehension, vocabulary	1) Significant interaction between gloss type *position; T-S exhibited the lowest recall mean. In addition, the difference between T-P and T-S conditions was significant while it was nonsignificant for the lexical groups. 2) Nonsignificant interaction between GT*GP in terms of comprehension 3) Significant interaction between GT*GP in terms of vocabulary recognition; TS exhibited the lowest vocabulary recognition. 4) Significant interaction between time*GP*GT in terms of vocab production; the groups performed similarly in the pop-up condition but in the SW condition, the lexical group had significantly higher immediate test score.	Partially confirmed
H5	T-S > T-P=L-S= L-P in terms of CL	Nonsignificant interaction in terms of the perceived CL ratings	Not confirmed
H6	High WM > Low WM in terms of recall, comprehension, vocabulary	Significant main effect of WM in terms of recall, comprehension, and vocabulary production (not recognition).	Confirmed
H7	High WM < Low WM in terms of CL	Nonsignificant main effect of WM	Not confirmed
H8	Lexical > Topic-level in terms of access to glosses	Nonsignificant main effect of gloss type.	Not confirmed
H9	Pop-up > SW in terms of access to glosses	Significant main effect of gloss position. Pop-up groups accessed annotations more frequently than SW groups.	Confirmed
H10	Lexical > Topic-level in terms of total reading time	Significant main effect of gloss type. Lexical groups spent more time on reading the text than topic-level groups.	Confirmed
H11	Pop-up < SW in terms of total reading time	Nonsignificant main effect of gloss position.	Not confirmed.

4.5.1 Additional findings

1) Significant interaction between WM*gloss position in terms of comprehension. In pop-up condition, high WM scored significantly higher than low WM. In separate window condition, WM did not have an effect. In addition, low WM participants performed similarly in both condition but high WM participants' performance decreased significantly from the pop-up to the SW condition. This suggests that low CL condition boosts the performance of high WM participants (The rich get richer effect).
2) Significant interaction between time*gloss type in terms of both vocabulary recognition and vocabulary production. Readers who took lexical annotations performed significantly better than readers who took topic-level annotations on the immediate tests.
3) Significant main effect of time in terms of both vocabulary recognition and production. On both tasks, immediate test scores were significantly higher than delayed test scores.
4) Significant interaction between time*GP*WM. In the pop-up condition with lower associated CL, readers with high WM were able to retain more words from the text compared to the low WM readers. On the other hand, in the separate window condition with higher associated CL, the groups performed similarly both on the immediate and delayed tests.

CHAPTER 5

DISCUSSION

Theoretically built on Kintsch's (1988) construction-integration (CI) model and Sweller's (1994) cognitive load theory (CLT), this study mainly attempted to test the intermediary role of the content of annotations on reading comprehension and incidental vocabulary learning, and also on the cognitive load (CL) induced by reading an electronic text. Another aim of the current study was to investigate the various effects of the location of annotations by probing the split-attention effect on the aforementioned dependent variables. Lexical annotations consisting of L2 definitions of 28 words and topic-level annotations containing factual or contextual information regarding 12 text parts were used so as to explore their differing effects. In a factorial design, annotation combinations were generated to examine their relative influences. Accordingly, the study consisted of four experimental conditions: lexical pop-up window annotations (L-P), lexical separate window annotations (L-S), topic-level pop-up window annotations (T-P), and topic-level separate window annotations (T-S). A built-in tracking software was integrated into the reading environment to record the interaction of the readers with the text. Working memory capacity (WM capacity) was also taken into account as a reader-related mediator of comprehension and vocabulary learning process, whereas topic interest was incorporated into the study as a reader-related covariate factor. As such, a number of research hypotheses were constructed, and the findings were interpreted within the framework of CI model of text-comprehension, the split-attention effect of cognitive load theory, and WM research.

5.1 Annotation use and reading

At the onset, it was assumed that lexical annotations integrated into the text would enhance the construction of text base model as the knowledge of individual words would facilitate automatic word recognition which, in turn, would improve formation of the propositional base of the text. Access to topic-level annotations, instead of definitions of words, on the other hand, should not hamper the construction of text base because these high proficiency learners were expected to use other complementary strategies to figure out the meanings of unknown words from the context to form a mental model of the text (Ariew & Erçetin, 2004). The lack of a significant main effect of gloss type confirms the first hypothesis that word-level and topic-level annotations would result in similar levels of performance in terms of text recall which relies on the formation of a text base. The interaction detected between gloss type and gloss location on free recall portrayed that readers accessing lexical annotations in a separate window (L-S) recalled a significantly higher number of propositions than readers taking topic-level annotations in a separate window (T-S). This finding was also supported by the log data since text recall scores of the L-S group participants increased as the total reading time increased. This might suggest that as the participants got more engaged with the reading task, ergo made effective use of the annotations, their performance on the text recall also augmented. Besides, an increase in total reading time might signal an increase in CL, as well, though this high mental load does not necessarily lower performance (Paas et al., 2003). Hence, though the participants spent longer times on reading the text, this did not affect text recall negatively although it did so for the short-term vocabulary learning (see below). In the pop-up condition, the difference was not significant.

On the other hand, within the same gloss position, participants who accessed topic level annotations generated significantly fewer ideas. This finding confirms the expectation that split attention effect should be observed with larger amounts of information or element interactivity (Hypothesis 4). In view of the increased cognitive demands associated with annotations in a separate window, extraneous load must have taken place triggered by the high element interactivity of topic-level glosses. Instructional materials that contain more than one source of information run the potential risk of creating split-attention effect (Ayres & Sweller, 2005). Element interactivity, "the number of elements that must be simultaneously processed in working memory in order to understand the information" (p. 141-142) is of concern when mentally integrating multiple sources of information. For one, instructional material which is high in element interactivity, namely complex and heavily demanding designs, induces intrinsic CL. For another, if this heavy processing load is doubled with the mental load of integrating disparate sources of information, the split-attention effect takes place which is a form of extraneous CL (Chandler & Sweller, 1996). Reversely, having fewer elements to simultaneously process to be able to make sense of the instruction (Kalyuga et al., 2003) compared to topic-level glosses, lexical annotations did not put undue processing strain on working memory. In consequence, the processes of recalling information from the text and forming a proper text base were not hindered for the L-S readers even though they might have experienced extraneous load resulting from mentally integrating disparate information sources (hence spent more time for reading the text). That is to say, they may not have suffered from splitting their attention between the physically disparate sources of information since intrinsic load imposed by high element interactivity was not the case for lexical annotations. A comparison within the gloss types also implied

a split-attention effect for the topic-level annotation groups as their scores differed in the recall task from pop-up to separate conditions. Consequently, they performed better when they took topic-level annotations in a pop-up window. As such, the gloss type associated with high CL led to better results when integrated into the main material. To the contrary, the location of annotations did not significantly affect the performance of lexical groups suggesting that their performance did not change a lot based on the location of the glosses. In the end, it can be maintained that both gloss types contributed to text recall in different interactions with the gloss position in varying degrees. This was also corroborated with the interviews as both lexical and topic-level groups reported that they found annotations useful for the completion of tasks.

Apart from recall protocols, a comprehension test was used within the current design to gauge readers' comprehension upon reading the electronic text. The test contained a total of fourteen questions which were a combination of referential and inferential questions. Related to the performance on the multiple-choice comprehension test, it was predicted topic-level annotations would lead to better results than lexical annotations. It was hypothesized that answering comprehension questions requires the ability to bring various text parts together as well as connecting them with the prior knowledge. Being a form of prior knowledge in content, topic-level annotations were assumed to enhance situation model building (i.e., combining textual information with already existing schemata to form inferences). In fact, the results confirmed this assumption (Hypothesis 2) as there was a statistically significant superiority of topic-level annotation groups in the comprehension test over the lexical annotation groups. Additionally, within the T-S condition, comprehension test scores increased as the frequency of access to glosses

and the total reading time increased reiterating the effect of topic-level glosses for comprehension even in the non-conducive format. Despite the overall less frequent use of glosses by the T-S group resulting in decreased performance across tasks, access to annotations assisted them only on the task which required prior knowledge the most. The reason for the observed dominance of topic-level glosses over word-level glosses only in the comprehension test but not in the free recall task can be explained in terms of the type of information each of the tasks taps and the level of comprehension associated with them. First, heavily relying on memory, free recall represents mostly the main ideas or explicit information from the text (Alderson, 2000). Comprehension test, on the other hand, as an amalgamation of lower-order and higher-order questions, taps literal understanding as well as inferential comprehension. This type of processing obviously requires more than definitions of words as comprehension supplements. Thereby, these advanced readers equipped with a range of reading strategies, such as guessing the meaning from the context (Ariew & Erçetin, 2004) to cope with unknown words, must have excelled in the presence of contextual support in the MC test entailing deeper processing as opposed to those who took lexical annotations. This also suggests that as the number of interacting elements augments, the number of information to be kept and processed in WM also increases. Thus, participants receiving higher-order annotations (i.e., topic-level glosses) and managing to process these lengthy descriptions and to integrate them into their mental model were also advantaged in terms of succeeding in the more intriguing task of multiple choice comprehension test. A similar effect was also documented in Türk and Erçetin (2014) whereby readers viewing verbal and visual information simultaneously in a single gloss excelled in the more compelling and cognitively demanding tasks, such as problem-solving in comparison

to readers given the option to view either verbal or visual content alone. Few studies integrated topic-level glosses to explore their comparative effects next to word-level glosses (Erçetin, 2003; Erçetin, 2010). Unlike the present study, an inverse relationship was reported in Erçetin (2003) where intermediate learners accessed both textual and contextual annotations more frequently than advanced learners whereas this high frequency of access to annotations yielded deleterious effects on reading comprehension.

Overall, WM capacity was found to be a consistent predictor of reading comprehension on both measures, namely on text recall and comprehension test. Actually, for almost all the dependent variables, except for vocabulary recognition, WM capacity was found to be the common denominator of success. As suggested by Kintsch (1998), comprehension operates at several levels including micro-level and macro-level processes. The coordination of disparate knowledge sources such as linguistic information derived from the surface text code, schematic knowledge and knowledge of genres stored in long-term memory entails the active involvement of WM during text comprehension. Given the significant role of WM capacity in processing and storing the upcoming information from the text, integrating it with already existing knowledge and retrieving information from LTM, it is expected to determine passage comprehension which was also compromised by the findings of this study. This finding also corroborates Harrington and Sawyer's (1992) study in which WM capacity, operationalized as a *trade-off* between processing and storage, was found to be an index of reading skill as measured by text recall.

Specifically, in the current investigation, high-WM capacity readers who took pop-up window annotations performed better in the comprehension test compared to low-WM capacity readers who took pop-up window annotations. Pop-up window

annotations were already expected to reduce the extraneous load by physically integrating disparate sources of information and both low WM and high WM capacity readers were expected to benefit from the presence of them. Partially confirming research Hypothesis 6 which took into account the strengthening effects of WM capacity, this finding is also consistent with Leeser's (2007) study which also displayed an effect of WM capacity on text recall only in combination with topic familiarity. According to the *rich-get-richer model* (Stanovich, 1986) mentioned in that study, WM capacity enhances the effects of prior knowledge. As such, it was noticed that high WM capacity subjects got more help from domain knowledge than low WM capacity subjects. As with the effect of domain knowledge in the rich-get-richer model, in this study the integration of annotations within the text made the role of the WM capacity more predominant. Likewise, high capacity readers getting pop-up window annotations which mediated the adverse effects of extraneous load excelled in the comprehension test. Reversely, in the separate window annotation conditions where the extraneous load is high, the annotation use equally affected/ overloaded high- and low-WM capacity readers. However, since the combined effects in the focus of that study (i.e., WM capacity and domain knowledge) differed from the present investigation, a direct comparison would not be appropriate. Rather, based on the rich-get-richer model, it is quite understandable that the combined effects of positioning annotations in the immediate environment and high WM capacity actually complemented comprehension.

Morrison (2004), with a similar annotation format, namely rollover annotations which resemble tooltips used for pop-up annotations in the current study, also advocated the use of such careful techniques to present glosses so as not to exceed the cognitive capacities of readers. A similar observation was attained in

Chen and Yen's (2013) study which demonstrated the contributions of annotation use, pop-up window annotations in particular, to reading comprehension which was attributed to the "physically immediate availability and nonintrusive quality" (p. 421) of glosses in pop-up windows. Likewise, in the present study pop-up window annotations proved to be more effective than separate window annotations when taken into analyses in the company of different variables for measuring situation-model building. AbuSeileek (2008), in a similar way, reported an effect of hypermedia annotations rather than traditional end-of-the text glossaries on reading comprehension. That the bottom-of-screen glosses yielded lower scores for both reading and vocabulary measures than the marginal glosses did was related to the split-attention which might have occurred in managing information coming from different locations in that study, too. However, unlike that study no main effect of gloss position was detected in the current study except for the increased frequency of access to pop-up window glosses. In other words, positioning the annotations in pop-up or separate windows did not trigger any individual effect within the current design neither on recall nor on comprehension tasks. Plus, in that study, the pop-up format produced the lowest results like it did in Chen's (2016) study. Both of the researchers attributed this finding to the inability of low proficiency level learners participating in their study to take the advantage of the pop-up format. Nevertheless, in the current study high proficiency level learners benefited from it as the integrated format boosted up the look-up behavior of the participants although it did not cause any other significant performance change on its own. Instead, the location of annotations mattered when the interaction with a particular annotation type or with WM capacity was at work.

5.2 Annotation use and vocabulary learning

The effect of annotation use was especially observed for both vocabulary recognition and vocabulary production. Evidently, not only does vocabulary knowledge aid building the mental-model of the text by enabling attending to surface code easier, but reading process also assists enriching the mental lexicon. The relationship between reading and vocabulary learning is reciprocal wherein one supplements the other. When readers encounter the words within a text, contextual knowledge of words which is defined by Stahl (2003) as the recognition of how words act across different contexts develops at the same time along with definitional knowledge. Through consistent and multiple exposures to words in context, gain and retention of vocabularies are likely to take place. Hence, the supremacy of lexical annotations on vocabulary gain is an unsurprising result of the present study which also validates Hypothesis 3 in part (i.e., only for the immediate effects). In this respect, this finding also corroborates Akbulut's (2007a, 2007b) and Cheng and Good's (2009) studies which found a strong relationship between vocabulary learning and gloss use. More specifically, in this study, it was seen that readers who took lexical annotations outperformed readers who took topic-level annotations on the unannounced immediate vocabulary matching and on the immediate definition production tasks. This suggests that access to word-level information during reading offers more benefits for incidental vocabulary learning in the short run. Looking at the total time they spent on reading which is significantly higher than that of the topic-level groups, the finding that lexical groups learned more words from the text than the topic-level groups is accentuated. Aligned with this intensive use, the interviewees accessing lexical annotations reported that they found the annotations useful for the completion of the upcoming vocabulary tasks and learning new words in general.

Taken all together, these corroborate the overall supremacy of lexical gloss type over topic-level gloss type in terms of vocabulary learning as well as the advantage topic-level glosses provide readers in terms of deeper-level comprehension. This particular finding corroborates Gettys et al.'s (2001) study in which it was observed that while word-level glosses yielded better vocabulary learning which was contended to result from a deeper level processing, the participants found reading with sentence-level glosses more enjoyable and satisfactory.

The short-term effect of glossing is also in tune with the extant research on incidental word learning as Hulstijn (2001) argues that evidence from the experiments at hand suggests that one-time exposure to words rarely turns into permanent learning which would be adequate for future retrieval of them from memory. Though this is a short-term effect, the fact that lexical annotations per se, not topic-level ones, increased vocabulary learning – even when they took the vocabulary tests unannounced – underlies the influence of noticing in incidental learning of words at the same time. As Schmidt (1990) argues, even in incidental learning, there is a degree of consciousness involved where learners notice new input. In the same vein, highlighting word-level annotations for lexical gloss groups accompanied with their definitions must have nurtured readers' attendance to form-meaning mapping which in turn increased their lexical gains. In line with this finding, interviewees from the lexical annotation groups reported positive feelings toward annotation use and found the annotations useful in their task accomplishment. Conflicting with these findings, the correlations obtained from the log data showed that the receptive vocabulary gains in the L-S group decreased as their frequency of access to annotations increased, though (see below). Since no such specific attention was allocated to new words in the topic-level annotation groups, the finding that no

difference was observed in their vocabulary scores from immediate to delayed testing is not much of a surprise.

This short-term benefit of lexical glosses could not be maintained in the long run since lexical and topic-level annotation groups performed alike in the unannounced delayed test contradicting the research Hypothesis 3. While the incidental vocabulary recognition gains from lexical annotations underwent a dramatic decrease over time, topic-level annotations remained constant suggesting no effect on vocabulary learning whatsoever which is quite reasonable by considering they spent less time on reading the text which implies that they may not have used annotations very effectively. Nevertheless, the location of the glosses mediated the processing of unfamiliar words during reading for the topic-level annotations. Accordingly, the readers in the T-P group recognized more words on the immediate vocabulary matching test than the readers in the T-S group. The integration of topic-level annotations might have offered learners a context with which the word(s) can be linked. Further, the overall better comprehension topic-level glosses catered might have also facilitated word learning. Still, this is a finding that must be handled cautiously since a comparison within the gloss types showed that the mean of the lexical group was significantly higher than that of topic-level group, and the effect of the latter was only present with the interaction of glossing position. One explanation could be that topic-level annotations which granted lengthier explanations and deeper levels of knowledge to process simultaneously must have triggered intrinsic CL as explained above. This high intrinsic load combined with the extraneous load imposed by information coming from a separate window must have overloaded WM by creating the split-attention effect in the T-S condition. This finding, in part, corroborated Hypothesis 4 favoring pop-up window annotations over separate

window annotations thanks to their facilitative effects for the mental integration of corresponding representations (Kalyuga et al. 2003). From another aspect, this also implies that readers who managed to accomplish the intrinsically demanding task of integrating longer and indirectly-related information into their meaning-making process must have also been competent enough to employ wise strategies to match unknown words with their definitions which was a requirement of the vocabulary recognition task. As skilled L2 readers they might have employed other cognitive strategies to retrieve the meanings of unknown words from the context while reading the text in the absence of word-level information (e.g., guessing the meaning from context) just as the advanced readers in Ariew and Erçetin's (2004) study did. Without the interference emanating from the separate positioning of topic-level glosses, these expert readers in the T-P format were able to allocate resources to process the newly encountered words while getting contextual help at the same time. It should also be noted that during the interviews, the participants from the T-P condition stated that annotations were easy to use and helpful for text comprehension. Possibly, this increased comprehension and the use of strategies to cope with unknown words during mental modelling might have provoked the incidental acquisition of them as vocabulary development through incidental learning is documented to improve when reading is supported by high quality contextual clues (Webb, 2008). At the same time, one can also assume that this resulted from the more frequent access to glosses by the pop-up groups as revealed by the log data. Since the participants in the split-formats did not use the annotations as frequently as the participants in the integrated-formats, they may not have benefited from the presence of annotations that much.

For the receptive and productive vocabulary gains, the interaction between gloss type and gloss position demonstrated that there was a statistically significant difference between the L-S and T-S groups whereby the former outperformed the latter. Although, at the first glance, this finding could be attributed to the physical separation of lexical glosses from the text, a closer look at the log data showed that actually more frequent use of glosses lowered the gains on vocabulary recognition in the L-S condition. Accordingly, the participants performed worse on the immediate vocabulary matching test as they accessed lexical glosses in a separate window more frequently. While the L-S group's receptive vocabulary scores were the best, they obtained this success not through the use of glosses, but through the disuse of glosses. This is an interesting finding (although the correlation is modest) which runs counter to the overall dominance of lexical glosses for word-learning and the increased reading time of the lexical annotation groups. This suggests that these highly proficient readers opted for not accessing glosses when they were presented in an isolated format, rather they got help from their own resources (e.g., reading strategies) when the target words were highlighted. From a split-attention perspective, this may mean that the readers were actually overwhelmed by the physical separation of word-level glossaries, but still managed to improve their vocabulary knowledge when encountered the highlighted words within context by applying other strategies to guess their meanings.

With regard to the productive vocabulary gains of incidental learning, lexical gloss over topic-level gloss for the immediate effects and high WM capacity over low WM capacity for the extended effects seem to underlie success. The effects of providing lexical glosses and WM capacity became more visible in the productive vocabulary task which is intuitive as it necessitated retrieving word meanings from

memory which would be facilitated more with the presence of word-level glosses and higher WM capacity. The major finding of this measure was that readers performed better on the immediate vocabulary production task upon viewing lexical rather than topic-level annotation in a separate window. One more time, this result together with the significant main effect of gloss type confirms the research hypothesis that lexical annotations would lead to better learning in all vocabulary measures. Given the abundance of evidence on learning new words incidentally when the aim is to comprehend a text (Chun & Plass, 1996; Fang, 2009; Al-Seghayer, 2001; Watanabe, 1997; Xu, 2010; Yoshii, 2006), the significant relationship found in this study between access to lexical glosses and lexical intake across both vocabulary measures makes perfect sense. In fact, the increased time on the reading task for the lexical group suggests that the readers were more cognitively involved in schema construction and/or organization in this condition. In the end, this high engagement with the text must have lent itself to better short-term vocabulary learning.

Next, a significant interaction was detected between time, gloss type and gloss location. As such, it was seen that lexical and topic-level groups performed alike in the pop-up window condition in both times while in the separate window condition lexical group outperformed topic-level group on the immediate post-test. The emergent interaction between time, gloss type, and gloss position can be interpreted in the light of the limited capacity working memory assumption. As elaborated above, the split-attention effect takes place only when the instructional materials were of high element interactivity. Hence, it comes as no surprise that lexical annotations, naturally fostering vocabulary learning coupled with the benefits of low element interactivity, led to greater differences between word-level and topic-

level annotations in the separate window condition than they did in the pop-up window condition. In a way, the effect of gloss type becomes more predominant in the non-conducive gloss position. This result confirms the fourth research hypothesis predicting that the T-S group would get the lowest scores across tasks among all the groups. In terms of the performance of the other groups, no difference was expected across the remaining conditions (L-P, LS, and T-P). Thus, the lack of a significant difference across the pop-up groups in a way confirms this prediction. That is, as gloss position associated with lower CL, the integrated format was already expected to facilitate learning. Although topic-level annotations were not assumed to lead to vocabulary learning, they still mediated this process when given within the text as they did for the vocabulary recognition gains. Still, the single effect of gloss type was already high suggesting the dominance of lexical gloss type over topic-level glosses in terms of productive vocabulary gains though the performance of the topic-level group approached to that of the lexical one when combined with the gloss position effect.

A second major finding for productive vocabulary gains concerns the distinction between high WM capacity and low WM capacity readers as to the location of annotations they received and the time of testing. Particularly, it was observed that in the low CL condition (i.e., pop-up window), high capacity readers were able to supply more definitions for the targeted words than low capacity readers on the delayed test whereas the difference between high WM and low WM groups was not significant in the high CL condition (i.e., separate window). Even though this is an interaction which cannot be attributed to a single variable, it is obvious that the individual characteristics of readers supported by the instructional design were influential in determining the success for lexical retention. It seems that the effect of

WM capacity interacted with the effect of time, and a significant difference was observed in the retention of productive vocabulary between high and low capacity readers in the pop-up window conditions. Although this finding confirms the assumption on the determining effect of WM across the tasks for one condition, it conflicts with the same assumption for the other. Actually, the influence of WM should have been greater in the condition associated with higher CL since it was assumed to mediate the processing and retrieval of information. Given the limited nature of WM, the distinction between the high capacity and low capacity readers was predicted to widen when the instructional material imposes extraneous CL evoked by the format of the task. High capacity readers were expected to cope with this increased processing demands induced by the separate location of glosses significantly better. Yet, they did not outscore despite this advantage in the more demanding condition while they excelled in the presence of the more conducive conditions. It follows that high capacity readers got richer combined with the effect of the integrated annotation format in terms of the long-term benefits of incidental vocabulary learning through reading. The proximity of annotations without interrupting the flow of reading must have better enabled the processing and storage of new lexical items in the long-term memory. Thus, long-term vocabulary learning benefited more from the integrated than the separate presentation of annotations when it is combined with the effects of higher WM capacity creating another instance of the rich-gets-richer effect for the present study.

Conflicting findings arouse from the discussion above. On the immediate test no difference was observed in the pop-up condition between lexical and topic-level annotations in terms of productive vocabulary gains. The difference was notable in the separate-window condition, though. On the delayed test, however, no difference

in the separate-window condition was detected, while in the pop-up condition the difference between high WM and low WM was remarkable. Note that the lack of word-level glosses did not prevent readers in the T-P group from noticing new words and using other cognitive strategies to compensate for this lack of word-level annotations. T-S group, on the other hand, performed well below the L-S group implying the effects of split-attention in terms of short-term vocabulary gains. The interaction between time, gloss position and WM went in the opposite direction for long-term vocabulary learning. This time, the difference between high WM and low WM readers became more visible in the pop-up window condition not in the separate-window condition. The comparison of these two results are not quite straightforward though, because of the difference in the members of interaction. In the former, gloss type changed the results interplaying with the effects of time and gloss position, whereas in the latter WM capacity affected the results in collaboration with time and gloss position effects. We must also keep in mind the overall single effect of lexical gloss type on immediate vocabulary gains while its effect disappeared on the delayed testing of vocabulary. This suggests that when the benefits coming with the use of word-level glosses were not available, the effect of gloss position changed direction in favor of the low CL (pop-up) condition reinforced with high WM capacity. In general, all groups performed better when tested immediately after reading than when tested after a certain time period except for this interaction detected for the delayed post-test. The facilitative effect of WM capacity for long-term vocabulary learning can be attributed to that fact that readers make more use of their cognitive capacities in more favorable conditions rather than pushing their limits in the less favorable contexts.

On the whole, if we are to summarize the findings related to the effects of gloss position on reading and vocabulary learning, inconsistent results emerge. For instance, while word-level glosses in a separate window led to better text recall, topic-level annotations integrated within the passage improved recall more. Plus, while the integrated format of the topic-level glosses benefited active recognition, the separated format of lexical glosses benefited both active recognition and passive recall in the short run. In the end, it can be said that when readers' cognitive resources were not overloaded, the presentation mode of the annotations, be it pop-up or separate, did not have significant effects on text comprehension or vocabulary learning on their own.

5.3 Annotation use and cognitive load

Concerning the cognitive load perceived by the readers across the treatment conditions, it was assumed that topic-level annotations which were high in element interactivity would generate the highest perceived CL when given in a separate window associated with high extraneous load. Readers may suffer from the disparate positioning of these lengthier annotations and perceive increased load for all the task types. However, no significant effects emerged from the results in terms of the participants' self-ratings which is consistent with Yao's (2006) findings depicting no effects of glossing format on the CL. Additionally, high-WM readers were expected to report lower levels of perceived CL than low WM readers, but no influence of WM capacity was observed on the perceived CL either. That the electronic reading environment was self-paced might have erased any effects related to WM capacity since they may have read the text flexibly without constraining their capacities, as a result of which they did not perceive an increased mental load. Nevertheless, if we

are to take tracking data as relatively direct measures of CL, that the lexical annotations caused longer reading time can give us some clues in interpreting the self-reports. The longer reading time along with the higher levels of CL have been considered as signs of cognitive overload not necessarily leading to lowered performance (Tuovinen & Paas, 2004; van Merriënboer et al, 2002). In fact, it was assumed that due to the lower intrinsic CL associated with lexical annotations, they would trigger more frequent access which would in turn extend the total time spent on reading in this self-paced electronic reading environment. Yet, apparently the reading time did not extend because of a more frequent look-up behavior as there was not a significant difference between lexical and topic-level groups' frequency of access to glosses. The readers in the topic-level groups accessed the glosses as frequently as the ones in the lexical groups, yet this frequent access did not extend the total reading time though the content of topic-level annotations were lengthier than that of the lexical ones. It may be the case that they reached those annotations just out of curiosity just as the readers in Erçetin's (2003) study and did not elaborate so much on the content of annotations which would have otherwise increased the time spent on reading. On the other hand, the reason for the longer reading time in the lexical groups can be the longer time needed for the integration of new words into the mental lexicon. Obviously, access to word meanings is required not only for form-meaning mapping but also for the comprehension of the propositional meaning. These processes of schema construction and/ or organization might have put more excessive demands on the reader compared to topic-level annotations which might have been perceived as inessential for comprehension. Despite this heightened mental load, lexical conditions performed better across vocabulary tasks and in the recall task combined with the effect of the gloss position. This corroborates van

Merriënboer et al.'s (2002) study where the learners who spent more time on the task and invested more effort performed better on transfer and retention tasks. Instead of affecting performance adversely, such co-occurrence of high performance and high mental effort might be an indicator of high mental efficiency (Paas et al., 2003). Also, despite spending longer time on reading, the interviewees from lexical groups reported that they found the text appropriate for their level while those from topic-level groups found it difficult due to the number of unfamiliar words. As the perceived difficulty of the text for the lexical group was not notably different from that of the topic-level group, we might claim that the provision of word-level annotations though extended the reading time did not lead to task inefficiency.

In terms of the frequency of access to glosses, pop-up window annotations were used more frequently than separate window annotations which confirmed the hypothesis that easy-to-access nature of these annotations would lead to an increase in the access. The integration of annotations within the text must have triggered the look-up behavior without intervening the flow of reading although it did not shorten the reading time contrary to what was expected. Initially, it was predicted that pop-up window annotations would lessen the total time spent on reading because they were already integrated into the text and access to them would not require as much time as separate window annotations would do (Hypothesis 11). Yet, as the separate window groups did not access the annotations as frequently as the pop-up groups, this might suggest that they may not have even attempted to read the content of every annotation in the first place. As a result, no significant differences emerged between these two groups in terms of the total reading time. To conclude, across all the measures the facilitative effects of pop-up window glosses were noticed, especially in the case of topic-level annotations and for high capacity readers. Moreover, this

frequent use did not cause any cognitive overload as they did not increase reading time or were not reported to have caused higher perceived CL.

CHAPTER 6

CONCLUSION

6.1 Main findings

Initially motivated by the desire to explore the various factors influential on comprehension, word learning, and the perceived cognitive load such as reader-related factors (i.e., working memory capacity) and text-related factors (i.e., the use of glosses), the present investigation has shed light on several crucial issues related to electronic reading in L2. To articulate the general conclusions to be drawn from this study, firstly the effect of gloss type was found to be dynamic; changing according to the task type used to gauge comprehension. Primarily, lexical annotations given in a separated format were more useful for remembering text parts to reproduce its hierarchical structure whereas the integration of topic-level annotations benefited learning more. Log data also revealed that in the split format, text recall of the participants accessing lexical annotations and comprehension scores of the participants accessing topic-level annotations improved as the time spent on reading increased. This moderately complies with the frame of reference this investigation adheres to in the sense that topic-level glosses enhanced situation model building via enabling necessary prior information to build a situation model of the text (Kintsch, 1998). For vocabulary acquisition, access to lexical annotations was the consistent predictor of vocabulary gain; however, it did not lead to the retention of those newly learned words. In other words, any effects on the vocabulary gains related to the presence of glosses are short-lived. This finding is quite consistent with the incidental vocabulary learning research since it is contended to entail multiple exposure to words to turn them into measurable gains (Nation &

Wang, 1999; Waring & Nation, 2004; Waring & Takaki, 2003). The analyses also unraveled a dramatic decrease in the vocabulary learning of the groups receiving lexical annotations from immediate to delayed post-test, while the vocabulary gains of the groups receiving topic-level annotations remained constant across both times of testing. All in all, the second conclusion regarding the effects of gloss type was that glossing unknown words improves short-term vocabulary gains though this effect could not be maintained in the long run. For the general conclusion to be drawn, it can be claimed that gloss use benefited vocabulary gain better whereas no clear-cut findings were obtained in terms of reading comprehension.

The general findings as to the effects of the location of annotations in reading comprehension and vocabulary learning displayed no meaningful main effects for this variable; however, various interactions were observed at different levels of analyses for reading and vocabulary measures. To start with, as expected, topic-level annotations provided in a separate window produced the lowest results across reading and vocabulary measures. However, their comprehension scores improved as they accessed more glosses and spent more time on reading highlighting the influence of topic-level glosses for the comprehension test even in the non-conducive glossing format. Next, lexical annotations presented in isolated format benefited recall and vocabulary gains (except for vocabulary recognition which was negatively affected by the split-format) more than topic-level annotations given in separate windows. However, topic-level annotations led to better recall and better recognition of words when they were presented in a pop-up format. In fact, the integration of annotations within the text was assumed to yield better learning by preventing the effects of split-attention. The finding that both lexical and topic-level groups exhibited similar performance in the pop-up condition, whereas that the difference

between them became more visible in the separate window condition on the immediate testing of vocabulary and text recall complies with this expectation. Not only did the difference between the two gloss types increase in the split format, but the one associated with low CL also led to better results in recall and vocabulary measures depicting that they were not influenced by the effects of extraneous load as much as the high ICL conditions (i.e., topic-level groups). On the other hand, pop-up window annotations resulted in more durable learning for high capacity readers in terms of the retention of productive vocabulary and text comprehension. Taken all together as the third main finding, these results suggest that the effect of gloss position depended on the CL associated with the gloss type and the characteristics of readers (i.e., WM capacity).

Fourth, working memory capacity was found to play a major role in all processes of comprehension (i.e., construction and integration), and in some aspects of vocabulary acquisition, (i.e., in the retention of productive vocabulary). Among all the other factors, WM capacity was the consistent determinant of reading performance signaling its crucial role to predict success in L2 reading. For both constructing the propositional text base and forming a situation model, high WM capacity readers were at an advantage and excelled even more in the presence of conducive gloss locations or times of testing. Nevertheless, WM capacity was not effective on the perceived cognitive load triggered by the treatment conditions. When the task demands do not exceed the available resources of readers in an instructional design which allows the allocation of memory to processing new information, WM capacity is expected to determine success between high versus low capacity readers (Paas et al., 2004). In other words, a difference between high capacity and low capacity readers is predicted in the levels of perceived CL associated with learning

tasks. Yet, contrary to this assumption, WM capacity did not change the levels of mental effort invested or the difficulty perceived during reading across the treatment groups.

Fifth, mixed results were obtained in regard to readers' actual use of the electronic text as revealed by the log data. First, pop-up glosses were accessed more frequently than the separate window glosses though this frequent access did not cause significant changes in terms of the overall reading time. Next, as expected lexical glosses extended the total reading time while the frequency of access to them was not significantly different from that of the topic-level glosses running counter to the expectations.

Lastly, the participants' attitudes toward screen reading and annotations were diverse and changed according to the gloss type. That is, while participants receiving lexical annotations found them beneficial, participants in the topic-level annotation groups found them inessential particularly for the upcoming vocabulary tasks. Besides, according to the participants from the lexical groups, the text was not difficult whereas topic-level groups reported it to be difficult due to the presence of unfamiliar words.

To portray the general picture emerging from these conclusions:

i. For the effects of gloss type, lexical annotations catered to overall vocabulary learning while topic-level annotations benefited comprehension.

ii. Within lexical annotations, the split-format made the effects of gloss use more prominent especially for recall and vocabulary learning. In terms of topic-level glosses, the integrated format highlighted the effects of gloss use for comprehension combined with the effects of

WM capacity. For also vocabulary retention, the integrated format accentuated the effects of WM capacity.

iii. In terms of the role of WM capacity, it was observed that while the effects of other factors (i.e., gloss type and gloss position) changed depending on the task type, the role of WM capacity was more fixed particularly for reading. However, it did not affect the perceived CL during reading.

iv. For the reported CL, no effects of WM, gloss type, or gloss location were detected. However, the log data that tracked the participants' use of the electronic text revealed that participants in the lexical gloss conditions spent longer time for reading though it did not negatively affect their performance nor did it increase the perceived CL.

v. With regard to readers' use of electronic reading, the integration of annotation content into electronic text triggered more access to these annotations than the separated formats.

vi. As for the readers' attitudes toward electronic reading, lexical glosses were considered to improve vocabulary learning and to facilitate the reading process, while topic-level glosses were conceived to enhance comprehension.

6.2 Pedagogical implications

A number of implications regarding the use of electronic texts in EFL reading for academic purposes can be drawn from this between groups design. Electronic reading by nature is suitable for enabling access to authentic materials which will be rendered easier to read thanks to individualized help coming in the form of glosses or

hyperlinks. These kinds of flexible and self-regulated reading environments could pave the way to learner autonomy (Erçetin, 2003) since it is up to the reader to select a particular annotation and to pursue reading at their own pace. For instructional materials developers, this means rather than the bare presentation of reading texts, presenting them with the additions of appropriate aids such as glossaries, or navigation maps would improve reading ability and vocabulary learning. Yet, in so doing, care must be taken for the risks of overloading cognitive capacities of readers because in poorly designed materials cognitive overload may set in. Thereby, the ultimate conclusion would be that as the amount of information provided in the annotation gets longer and denser, pop-up conditions produce better solutions. In terms of short amount of information, though, such as word-level annotations, whether they are given in a pop-up or separate window does not really matter much. Given the overriding evidence on the facilitative effects of glossing on vocabulary learning (Fang, 2009; Hulstijn, 1997; Hulstijn, Hollander & Greidanus, 1996; Jacobs et al., 1994; Watanabe, 1997), and further support from the present study, it is obvious that extensive reading materials which are expected to cater to incidental vocabulary learning should be supported with word-level glosses to promote autonomous vocabulary learning and independent reading in L2. Note that topic-level annotations can also be embedded to support reading as well as vocabulary acquisition which was documented to develop as a byproduct of comprehension processes facilitated through the use of contextual clues. By looking at the insights derived from the interviews, it can be maintained that L2 readers also favor instructional support of this kind while reading authentic texts on the screen.

Besides, among the individual factors affecting reading comprehension and vocabulary learning, working memory was found to be a significant predictor of

success. The vigorous relationship between working memory and learning implies that instructional materials should be designed in such a way that would scaffold information processing and information retrieval or storage for low capacity learners through environments which would minimize the amount of excessive load on the limited resources of working memory. To exemplify, separating annotations from the text might exceed the capacities of even high WM readers especially when the number of elements to be processed simultaneously is high, thus would increase the processing constraints on WM. So as to spare resources for the meaningful integration of text parts into LTM for both low WM and high WM readers, electronic text designers can construct materials that would improve the performance of these readers by optimizing the CL according to their expertise.

From an assessment perspective, it is possible to claim that free recall and MC comprehension tests tap different constructs. While the former is more suitable for assessing the comprehension of main ideas, the latter incorporates a deeper level analysis and synthesis of the text (Kintsch, 1998). Furthermore, the requirement to reproduce the text based on memory may exert greater CL as opposed to answering MC questions and possibly prevents learners from displaying elaborate comprehension. Even though MC tests were criticized for representing a fragmented view of comprehension, the integration of textually as well as scriptally implicit questions (Pearson & Johnson, 1978; as cited in Alptekin & Erçetin, 2009) into the test entails the transfer of acquired knowledge in novel contexts (which is a form of higher-order processing). As a result, it would be fair to recommend the use of both approaches rather than the adoption of one or the other while assessing reading comprehension to be able to extract a reader's true performance.

6.3 Limitations and future research directions

While drawing generalizations or implications from this study, there are certain limitations to take into account which could have inadvertently affected the results. First of all, this is a quasi-experimental design in which groups of students were randomly given any one of the treatment conditions, but individual students could not be randomly assigned to those groups, rather intact classes had to be used for grouping because of the limited availability of the department lab and the inflexible course hours of students. As a result, a true randomization was not the case for this design, so the results could have been influenced by some other confounding variables such as motivation or affect even though participants proved to be homogenous in terms of language proficiency, topic familiarity, and topic interest and were registered in the same department of the same university. As has already been noted, the proficiency level of learners was fairly high in this study, thus further research with a purely experimental design and groups of students with different educational background and proficiency levels could yield different findings.

Next, data were collected in a short duration of time and limited to the performance on only one text. Integrating other texts with varying lengths and topics might produce comparable data which would portray a more accurate picture of reading ability and vocabulary learning through reading. Moreover, extending the time period in which the treatments were given with a variety of reading texts might provide more reliable measures of performance refined from other variables (e.g., motivation, mood, text difficulty, or text type). Fourth, there was not a vocabulary pre-test in this study assessing their familiarity with the targeted words. Although those words had been selected by students with a similar profile as unknown words to be targeted, vocabulary gains measured only through post-test results cannot be

safely linked to the treatment effects. Future studies can eliminate this limitation by adding a vocabulary pre-test into the design. Fifth, although the group size for the four experimental conditions were 30, each group was further divided into two halves to create low WM vs. high WM groups. As a result, in each cell of analysis the number of participants was 15 which can decrease the power. That's why Bonferroni adjustments could not be run during the analyses which might have increased the risk of committing Type 1 error. Lastly, we do not really know how much content of the annotations the participants actually read as log data only provided data on the frequency of access to annotations and the time they spent on reading. This prevented us from making safe claims if the shorter reading time of the topic-level groups did indeed stem from their not reading the whole annotation content or not (as we would expect an extended reading time if they really read the whole content). Implementing think-aloud protocols in the future studies can give more straightforward information as to their real-time involvement with the text and the annotations.

In the future, to better understand the role of WM capacity on recall and comprehension-based tasks, a stricter compartmentalization of questions in the comprehension test as to referential and inferential questions and whether they tap storage or processing capacities of working memory should be better analyzed. Furthermore, giving the topic-interest scale prior to reading did not produce significant differences across the groups in this study, so it was taken as a covariate. However, during the interviews the participants in the topic-level groups indicated they reached annotations just to satisfy their curiosity. In the future, then, the topic-interest scale can be given after reading the text to see how the presence or absence of topic-level annotations affects their overall interest towards the topic.

APPENDIX A

PRIOR KNOWLEDGE QUESTIONNAIRE

Name-Surname: Student ID:

You are going to read a text about the collapse of civilizations. The researcher would like to learn what your prior knowledge about the topic is. For each of the following statements, circle T if you know the statement is True, circle F if you know the statement is False.

Civilizations Questionnaire

1	The old Egyptian civilization emerged in the Nile valley.	T	F
2	The Sumerians were the first humans to form a civilization.	T	F
3	The Babylonians are often credited with inventing the wheel.	T	F
4	Hammurabi ruled the Roman Empire from 1792 to 1850 B.C.E.	T	F
5	Roman coins were used to publicize the emperor, his achievements, and his family.	T	F
6	Assyrians were the predecessors of the Persian Empire.	T	F
7	Hippocrates, who is known as 'father of medicine' belonged to old Indus valley civilization.	T	F
8	The Mongols invaded Baghdad which was under The Abbasid Caliphate in the thirteenth century.	T	F
9	Persian was the official language of the Ottoman Empire.	T	F
10	The longest consistent civilization in the human history is that of China.	T	F

APPENDIX B

THE READING TEXT

---Please read the following text for comprehension and write a summary of it in either English or Turkish

Reasons for Societal Collapse

Any observant individual walking among the ruins of an ancient city is immediately faced with the following question: "How did the once magnificent civilization that ruled this place, that built this city, end like this?" The same person will certainly generalize his observation to the whole of world history and notice that no civilization, ever, was able to hold on to its powerful status among other nations. It appears that each one of them, like a human being, was destined to be born, age and die. This observation may go against our intuition. We expect that once a civilization becomes powerful, it will use its power to stay dominant. But somehow, this happens not to be the case. To name the most quoted examples, the civilizations of the Greeks, Persians, Egyptians, Olmecs, Romans, Mongols, and Ottomans, all of which were deemed indestructible, fell one after another, leaving us in awe and puzzled. However, the question of "What went wrong?" is much more important than satisfying curiosity: Thousands of years later, will another observant individual walk among the ruins of the cities in which we are living built by our civilization? Or can we learn from the mistakes of the extinct civilizations and avoid their fate?

Theories concerning the collapse of civilizations

There are many theories concerning the collapse of civilizations, but of course, if a theory does not conform to reality, it is worth nothing. In this article, I will first give a brief account of widely held beliefs about the collapse of civilizations, explain the weaknesses of these theories, and then give a rationale that I believe better explains the historical data we have. As for most social problems, we will perhaps never know the truth about why societies collapse. However, the stakes at hand are so high that we must make every effort to understand, and to an extent, solve this problem.

Insurmountable natural disasters, intruders, and competition with others

The most common explanation for such collapses is some insurmountable natural disaster, like an epidemic, hurricane, drought, or earthquake that leads to the demise of a civilization by killing the population and crippling the economy. Widely-cited examples are the eruption of the volcano in Thera that preceded the collapse of the Minoan civilization, the malaria epidemic in the Roman Empire or earthquakes in Mesoamerican societies. These arguments, which are very appealing to our human nature, that desires simple explanations for all questions, are in fact very unsound. Societies constantly experience such disasters, yet survive them. The potato blight in Ireland in 1845 halved the island's population but there was no cease of sociopolitical complexity as a result of the disaster. It is strange to think that the Roman Empire, which survived many disasters before, including the eruption of Pompei in AD 79, fell to malaria. We should consider that complex civilizations are designed to absorb such disasters, and they do. Just recalling the constant

earthquakes in Japan added to the loss of a world war with two nuclear bombs exploding in the heart of two large cities will sufficiently prove this notion. Japanese civilization did not collapse. On the contrary, it is one of the strongest economies in today's international arena. It is peculiar then that some civilizations are no longer able to fight such disasters. However, an act of God can certainly collectively destroy any civilization, as it did in the past like Sodom and Gomorrah. This is clearly narrated in the divine scriptures.

The other common explanations for such collapses are intruders and competition with other civilizations. The barbarian tribes, which brought the end of Rome in the fifth century, and the Mongolians that invaded Baghdad in the thirteenth century are clear examples of the intruder argument. This argument suffers from the realization that civilizations are attacked by outsiders throughout their existence, yet for some reason they cannot defend themselves near the time of their collapse. Competition with other societies however, is in principle expected to lead to growth and expansion instead of collapse. Still, there is no end to the examples from this category too, like the competition of the Ottoman Empire with Persia, which indirectly weakened its western front. But the competition argument is both intuitively confusing and it fails to account for major cases, like the fall of the Roman Empire.

Depletion of Resources

Another widely held belief about such collapses is that at a certain point in the life of a civilization a resource is depleted and the civilization that depends on this resource is prone to collapse. The Romans and the Ottomans both depended on military expansion for their economy, and when the relatively weak nations around them were engulfed or when they were barred from further expansion by geographical limitations, such as seas or large mountains, they were no longer able to use this resource. There seems to be some truth and lessons in this argument. To the uninformed, it is a very curious fact that the cradle of civilizations was Mesopotamia, where modern day Iraq is. How is it possible that the superpowers of that era, the Sumerians (~3000 BC) and the Babylonians (~1000 BC) chose to live in these deserts? How is it possible that they irrigated the land, raised armies, and built world wonders in these sand dunes? These questions actually are easily answered when we realize that Mesopotamia was not a dessert in that era after all. It is now a generally accepted theory that this place had a fragile ecosystem, which was destroyed after thousands of years of environmental pressure. The potential for these lands to accommodate great civilizations was lost after this fragile ecosystem was slowly destroyed by its inhabitants.

However, the argument of resource depletion inherently asserts that the elite of a civilization facing resource depletion passively waits for the predictable demise. I will argue below that this case, although strange, is true. Another difficulty of the resource depletion argument is that in some instances of collapse resources were never depleted. The fertile lands of Mesopotamia still remained green until later than 1000 AD, while many civilizations experienced collapses. The Romans, who used irrigation as a resource, kept farming till the very end. Finally, one may wonder why societies aim at possessing a higher amount of resources all the time. Population increase is only a partial answer to this question. We can easily imagine a society whose population stays the same; it is not a far-fetched hypothesis that this society

will naturally also try to increase its resources to fend off a variety of calamities it may experience, such as intruders and catastrophes. I believe herein lies an interesting rationale that brings together the mentioned theories that are flawed. To understand this, we first have to appreciate a law in economics, called the "law of diminishing returns."

Law of diminishing returns (Example from agriculture)

It is very rare in economics and in general social sciences that some series of observations can be identified as a "law." However, the "law of diminishing returns," first put forward in 1965 by Ester Boserup, is so comprehensive in its nature and explains such a variety of trends that it is now universally accepted. It goes: The return for an investment in a particular activity is great at the beginning, and then it gradually decreases. At a point, further investment brings no further benefits. At this point, the facility (a person, a group, a society, a factory) can no longer increase its returns, however much they would invest in that activity.

A simple example will clarify the law. Suppose we have a piece of land that we want to use for irrigation. In the beginning, we would just disperse seeds and wait for the crops to grow. Notice that our investment is minimal (say 1 unit of investment), and we get some food for our investment (again define this to be 1 unit of return). Then, if we want to increase the amount of crops we have, we may dig some canals for watering. It is straightforward to recognize that the canal digging is a lot harder than just dispersing seeds (say 5 times harder). However, it is again straightforward to recognize that although now we make 6 times more investment, we probably will not get 6 times the crop. Nevertheless, we want to maximize our return, so we still dig the canals. The next step would be to use motorized vehicles, which is maybe a 10 times increase in investment, but everybody will surely accept that it is not possible to get a crop that is 16 times greater than our original from the same plot of land. (Readers who may object that once the investments of canals and vehicles are made they will provide constant returns are reminded of the maintenance costs of these investments.) A further increase in returns may require genetically engineered crops that will require years of expensive research (more investment). The return per investment will always decrease for a certain type of activity, in this case irrigation.

Example from life and example from civilizations

This law is everywhere in life: If one week of studying suffices a result of 80 on one exam, in order to get 90, you need to study two more weeks. Most healthy people can run 100m in 20 seconds; to run it in 10 seconds you need years of exercising. Depending on one's abilities (which determine an individual's possible investment) these may even be impossible for many people. A vivid example is the heating problem in England during the nineteenth century. Heating, which was primarily carried out by burning wood from forests, with the increase in population had to be switched over to the burning of coal. The mining and distribution of coal, which is much more difficult than simply getting some wood from a nearby forest, was made even more difficult when the easily mined surface coal was rapidly depleted and deeper tunnels with lighting and airing problems had to be developed. It is intuitive why this law is in effect: Obviously, always, the easier solutions are adapted first, then the harder ones. Mining coal when you have easily available and plentiful wood is not reasonable. Consequently, we have a decline for our returns per investment.

The resources that civilizations use are no exception. A civilization that uses irrigation as a resource is bound to be limited by a certain level of return. Resource does not have to be depleted; it just cannot produce a return more than at a certain level. Another civilization that is dependent on taxation, mercenary or military expansion can achieve no more return after a certain level, no matter what adjustments it makes to its existing policy. Having said this, we can understand why a civilization that depends on a certain type of energy or resource cannot expand its influence beyond a certain level. Moreover, when energy becomes scarce, the civilization can become less agile in terms of trying new resources and new ways to produce returns, since agility and innovation mostly depend on using some of the surplus resources on strategies that will most probably yield no returns. Hence the rise of large architectural structures and many inefficient military operations are carried out during the ascent of a young civilization. These activities, which are easily buffered by the large returns that come from initial investment on the main resource of a civilization become impossibly costly later when the returns from the same investment is declining.

Civilizations like dinosaurs (Example from previous empires)
One last piece of the puzzle completes the rationale as to why civilizations collapse, and this piece is an easily accepted assumption: A civilization is like a dinosaur. It is large and strong, but it is adapted to the conditions into which it was born. The conditions change, however, the dinosaur cannot change its behavior. It helplessly tries to maximize the returns for the type of resource that it is adapted to use, and after a point, it simply cannot, thanks to the universal and unforgiving law of diminishing returns.

At this point, another civilization, that primarily uses another superior resource, will have larger returns, build a larger army to invade the former civilization, build larger ships to cut off the trade routes, and produce goods to cripple its economy… This is just a matter of time, and it is unavoidable. The strength of the civilization in its golden age is now its weakness. In such a weakness, since there is no extra resource to fight new problems-all resource is either used up by the population, or goes toward defense costs-even a natural disaster can bring an end to a civilization that once seemed to be indestructible.

The Ottoman Empire's strength in its rise was its perfect hierarchical organization which led to the accumulation of all power under the Sultan. Its main resource was military expansion and taxation of trade. These adaptations, which were ideal for the time between the thirteenth and fifteenth centuries, led to one of the most powerful empires that have ever reigned. However, by the sixteenth century, these strategies had become burdens: Due to the strong hierarchy, an intelligentsia that supported science and art as in the West could never develop. Military expansion had to stop. Taxation could no longer work since the Mediterranean Sea was no longer used for trade. The strategies were not abandoned though, instead, more investments were made in order to increase the returns, which as we saw above is a nonviable alternative. Eventually, other civilizations that used better resources brought about the end of the Empire. A similar order of events can be observed for other civilizations that collapsed. The great Arab historian Ibn Khaldun of the fourteenth century likens the lives of civilizations to the natural lifespan of individuals. They are

born, they grow old, and they die. In my view, a civilization does not die because it gets old; it dies because it cannot compete with a stronger civilization.

Conclusion

The natural question to ask is if our current civilization will collapse. From the analysis above, we can conclude that there are two reasons for the collapse of a civilization: 1) Dependence on a certain type of investment and failing to adapt to the new conditions. 2) The invention by another competitor civilization of a new type of investment with higher returns. In today's world, both of these reasons are in some ways different than those that existed in the past. First, with the advancement of science, the current civilizations are flexible in the resources they utilize, the options are constantly evaluated, the heating in United Kingdom does not collapse when wood is depleted; instead, coal, then gas, then nuclear power is used. The return for the investment made for some utility can be explained like this: whenever the return for a type of investment declines, we can shift to the next resource. Second, by the immense advancement in information processing and communication, the whole world is aware of the types of investments other societies are using, and a leading civilization that follows the developments in other countries is very unlikely to be threatened by a sudden development in a rival civilization. Third, because of progress in international trade, the old sense that any other civilization is an enemy has lost its significance.

Global warming and the depletion of petrol reserves were only two of the many alarming cues that we may have turned to declining returns for our investment curve. It is imperative to remember again that the stakes are very high. The next civilization to fall may bring about the fall of the human species.

YTU- Proficiency Exam Booklet (2007)

APPENDIX C

LIST OF ANNOTATED WORDS

1. intuition: (n) the ability to understand or know something because of a feeling rather than by considering the facts
2. deem: (v) to regard or consider in a specified way
3. awe: (n) a feeling of great respect and liking for someone or something
4. stakes: (n) [PLURAL] risk; the things that you can gain or lose by taking a risk, for example in business or politics
5. insurmountable: (adj.) too great to be overcome
6. demise: (n) a person's death
7. cripple: (v) to cause a severe problem for
8. halve: (v) to divide into two parts of equal or roughly equal size
9. peculiar: (adj.) different to what is normal or expected; strange
10. scriptures: (n) the sacred writings of Christianity contained in the Bible
11. intruder: (n) a person who enters, especially into a building with criminal intent
12. deplete: (v) to decrease seriously
13. engulf: (v) to completely surround or cover something
14. barred from: (v) to prevent or prohibit (someone) from doing something or from going somewhere
15. cradle: (n) a small bed for a baby, especially one that moves gently from side to side
16. dune: (n) a hill of sand formed by the wind, especially on the sea coast or in a desert
17. depletion: (n) reduction in the number or quantity of something
18. fend off: (v) to defend oneself from a blow, attack, or attacker
19. calamity: (n) an event causing great and often sudden damage; a disaster
20. herein: (adv.) in this document or book
21. disperse: (v) to distribute or spread over a wide area
22. switch over: (v) to change over, or change around
23. agile: (adj.) able to move quickly and easily
24. surplus: (n) an amount of something left over when requirements have been met
25. buffer: (v) to cushion, shield, or protect
26. reign: (v) to rule a nation or group of nations their king, queen, or emperor
27. intelligentsia: (n) intellectuals or highly educated people as a group, especially when regarded as possessing culture and political influence
28. nonviable: (adj.) not practical

APPENDIX D

TOPIC-LEVEL ANNOTATIONS

1. Olmecs: the first major civilization in Mexico. They lived in the tropical lowlands of south-central Mexico. The Olmec flourished during Mesoamerica's formative period, dating roughly from as early as 1500 BCE to about 400 BCE.

Retrieved from: https://www.boundless.com/world-history/textbooks/boundless-world-history-textbook/civilizations-in-the-americas-11/early-civilizations-of-mexico-and-mesoamerica-51/the-olmec-191-13286/

2. Mongols: an East-Central Asian ethnic group native to Mongolia and China's Inner Mongolia Autonomous Region. They also live as minorities in other regions of China (e.g., Xinjiang), as well as in Russia. The Mongol Empire which existed during the 13th and 14th centuries, was the largest contiguous land empire in history.

Retrieved from: https://en.wikipedia.org/wiki/Mongols

3. Minoan civilization: an Aegean Bronze Age civilization that arose on the island of Crete and other Aegean islands such as Santorini and flourished from approximately 2600 to 1400 BC.

Retrieved from: http://www.aoi.com.au/Cameos/CM601/

4. Mesoamerican societies: 'Meso' means 'middle', and these Mesoamerican cultures are the early advanced civilizations of Mexico and Central America. There were many unique groups inhabiting this region over time, including the Olmec, the Maya, and the Aztecs.

Retrieved from:
https://en.wikibooks.org/wiki/The_History_of_the_Native_Peoples_of_the_Americas/Mesoamerican_Cultures

5. Potato blight: In the harvest of 1845, between one-third and half of the potato crop was destroyed by the strange disease, which became known as 'potato blight'. The rest of 1845 was a period of hardship, although not starvation, for those who depended on it. The estimates of deaths in the famine years range from 290,000 to 1,500,000 with the true figure probably lying somewhere around 1,000,000, or 12% of the population.

Retrieved from: http://www.wesleyjohnston.com/users/ireland/past/famine/blight.html

6. Sodom and Gomorrah: cities mentioned in the Book of Genesis and throughout the Hebrew Bible, the New Testament as well as in the Qur'an and hadith. Sodom and

Gomorrah have become synonymous with impenitent sin, and their fall with a proverbial manifestation of divine punishment. Sodom and Gomorrah have been used as metaphors for vice and homosexuality viewed as a deviation.

Retrieved from: https://en.wikipedia.org/wiki/Sodom_and_Gomorrah

7. Baghdad: the capital of the Republic of Iraq. Located along the Tigris River, the city was founded in the 8th century and became the capital of the Abbasid Caliphate. Within a short time of its inception, Baghdad evolved into a significant cultural, commercial, and intellectual center for the Islamic world.

Retrieved from: https://en.wikipedia.org/wiki/Baghdad

8. Sumerians: Sumer was one of the ancient civilizations and historical regions in southern Mesopotamia, modern-day southern Iraq, during the Chalcolithic and the Early Bronze ages. Modern historians have suggested that Sumer was first permanently settled between 5500 and 4000 BC by non-Semitic people who spoke the Sumerian language

Retrieved from: https://en.wikipedia.org/wiki/Sumer

9. Babylonians: Babylonia was an ancient Akkadian-speaking Semitic state and cultural region based in central-southern Mesopotamia (present-day Iraq). Babylon greatly expanded during the reign of Hammurabi in the first half of the 18th century BC, becoming a major capital city. The earlier Akkadian and Sumerian traditions played a major role in Babylonian (and Assyrian) culture.

Retrieved from: https://en.wikipedia.org/wiki/Babylonia

10. Ester Boserup: (May 18, 1910 – September 24, 1999), a Danish economist. She studied economic and agricultural development, worked at the United Nations as well as other international organizations, and she wrote several books. Her works challenge the assumption dating back to Malthus's time (and still held in many quarters) that agricultural methods determine population (via food supply). Instead, Boserup argued that population determines agricultural methods. A major point of her book is that "necessity is the mother of invention".

Retrieved from: https://alchetron.com/Ester-Boserup-1362045-W

11. Mercenary: a person who takes part in an armed conflict who is not a national or a party to the conflict and is "motivated to take part in the hostilities by the desire for private gain." In other words, a mercenary is a person who fights for personal gains of money or other recompense instead of fighting for the ideological interests of a country, whether they be for or against the existing government.

Retrieved from: https://en.wikipedia.org/wiki/Mercenary

12. Ibn Khaldun: an Arab Muslim historiographer and historian, regarded to be the founding fathers of modern sociology, historiography, demography, and economics. His books influenced 17th-century Ottoman historians like Ḥajjī Khalīfa and Mustafa Naima who used the theories in the book to analyze the growth and decline of the Ottoman Empire.19th-century European scholars also acknowledged the significance of the book and considered Ibn Khaldun as one of the greatest philosophers of the Middle Ages.

Retrieved from: Cram101 Textbook Reviews Experience Sociology: Study Guide

APPENDIX E

SCORING PROCEDURE FOR RECALL PROTOCOLS

Reasons for Societal Collapse

1- The question of how the civilizations collapsed bewilders any observant individual walking among the ruins of once powerful civilizations.
2- Like human beings, civilizations are born, live, and die.
3- There are a number of theories concerning the collapse of civilizations.
4- First theory: Insurmountable natural disasters (earthquakes, draughts, plagues, etc.)
5- The eruption of the volcano in Thera preceded the collapse of the Minoan civilization.
6- The malaria epidemic in the Roman Empire or earthquakes in Mesoamerican societies caused the collapse of those societies.
7- The potato blight in Ireland in 1845 halved the island's population.
8- The Roman Empire fell to malaria after surviving many disasters including the eruption of Pompei in AD 79.
9- An act of God can collectively destroy any civilization, as it did in the past like Sodom and Gomorrah.
10- We should consider that complex civilizations are designed to absorb such disasters, and they do.
11- Second theory: Intruders and competition with others can cause the end of a civilization.
12- The barbarian tribes brought the end of Rome.
13- Mongolians invaded Baghdad.
14- The competition of the Ottoman Empire with Persia which indirectly weakened its western front.
15- Third theory: Depletion of Resources
16- The Romans and the Ottomans both depended on military expansion for their economy, but then came a moment when they could not use this resource anymore.
17- Mesopotamia (modern-day Iraq) was home to many civilizations (Sumerians, Babylonians, etc.) and very fertile, but it was also depleted.
18- A new theory: Law of diminishing returns: The return for an investment in a particular activity is great at the beginning, and then it gradually decreases.
19- Example from agriculture: when we want to grow something, we would disperse 1 seed, and get 1 unit of crop, but even if we increase our investments 5 times, the return you'll get from the land won't change.
20- Example from life: If one week of studying suffices a result of 80 on one exam, in order to get 90, you need to study two more weeks.
21- The heating problem in England; first they cut trees, then they mined for coal, last they turned to nuclear energy; but with every step the returns did not increase as much as the investments did.
22- Example from civilizations: A civilization that uses irrigation as a resource is bound to be limited by a certain level of return.

23- A civilization is like a dinosaur. It is large and strong, but it is adapted to the conditions into which it was born.
24- Example from previous empires: The Ottoman Empire's strength in its rise was its perfect hierarchical organization; the conditions changed, technology developed, but the civilization could not adapt itself to this change.
25- Will our civilization also collapse? The conditions are different now. thanks to technology we are aware of the developments in rival societies and threats that will come from them.
26- Conclusion: Global warming and the depletion of petrol reserves were only two of the many alarming cues.

APPENDIX F

COMPREHENSION TEST-KEY

---Please answer the following questions based on the text, *Reasons for Societal Collapse.*

1. According to the first paragraph, what is surprising about the great civilizations like the Greeks and the Persians?
a) the fact that they overcame their fate
b) the fact that they were destroyed
c) the way they got empowered
d) the way they ruled a civilization

2. What does ***"this notion"*** mean in line 37?
a) Catastrophes can increase the strength of a civilization
b) Earthquakes and atomic bombs can bring the end of a civilization
c) Overwhelming natural disasters cause the extinction of many civilizations
d) Powerful societies can cope with big disasters

3. The author's thoughts about the collapse of ***Sodom and Gomorrah*** in line 37 depend on _____
a) folk tales
b) historical facts
c) holy documents
d) scientific discoveries

4. According to the author, rivalry with other nations is assumed to bring about _____
a) the collapse of a civilization
b) the empowerment of a civilization
c) the invasion of a civilization
d) the protection of a civilization

5. Which of the below is **NOT** given as one of the reasons why the Ottoman Empire was the most powerful civilization between 1200s and 1400s?
a) hierarchy in the state
b) military power
c) support for science and art
d) taxes coming from trade

6. The author thinks that the strength of the Ottoman Empire at its rise became a burden later on because _____

a) it prevented the state's adaptation to the developments.

b) it required the use of new strategies.

c) the Empire consumed all of its resources.

d) the Empire reached the peak of its power.

7. While investments increase 15 times, returns may not provide that much gain simply because _____

a) expensive research will be necessary

b) initial gains will decrease more rapidly than the following ones

c) more investments have to be made for their maintenance

d) there is no need to maximize the investments

8. What does the author mean by ***"A civilization is like a dinosaur"*** in line 136?

a) A civilization continues its existence under tough conditions.

b) A civilization is famous for its flexibility.

c) It is difficult to accept that our civilization is getting old.

d) It is impossible to change the system of a civilization.

9. How does the heating problem in England prove "the law of diminishing returns" to be true?

a) Coal on the surface would balance the investment return cycle.

b) Feasible solutions would always be found with minimum effort.

c) Gains would decrease in time no matter what investments were made.

d) Reliance on forests would create maintenance problems in the future.

10. According to the text, the argument that best explains the reason for societal collapse is _____

a) competition with other civilizations

b) decline in returns

c) disruption by enemies

d) poor management of resources

11. What kind of a relationship exists between the amount of investment and the amount of return?

a) As one increases, the other will decrease.

b) A weak relationship exists.

c) No relationship exists

d) They increase together.

12. Which of the following is **FALSE** for current civilizations according to the text?
a) Advances in information technologies make it possible to be informed about the developments in rival countries.
b) Communication technologies slow down sudden developments in rival societies
c) Developments in science facilitate finding new resources in the case of consumption of old resources.
d) The other civilizations are not seen as enemies because of the spread of international trade.

13. The next civilization to collapse can be ours unless _____
a) we find better resources.
b) we increase our investments for returns.
c) we reevaluate the use of nuclear power.
d) we study the reasons of societal collapse in depth.

14. The following conclusion is taken from a NASA study:

"Technological change can raise the efficiency of resource use, but it also tends to raise both resource consumption and the extent of resource extraction. Therefore, the increases in consumption often compensate for the increased efficiency of resource use."

How can the "law of diminishing returns" explain these arguments of the NASA study?
a) Increasing the use of technology will create defects in resource policies
b) Technological devices will decrease resource consumption
c) There won't be much change in consumption related to technology
d) Trying to increase efficiency will inevitably increase the consumption

APPENDIX G

VOCABULARY PRODUCTION TASK

NAME-SURNAME: **STUDENT ID:** **DATE:**

Please write the meaning of each word either in English or in Turkish.

1. fend off: _____
2. surplus: _____
3. deplete: _____
4. awe: _____
5. demise: _____
6. halve: _____
7. intruder: _____
8. intuition: _____
9. engulf: _____
10. insurmountable: _____
11. calamity: _____
12. barred from: _____
13. cradle: _____
14. disperse: _____
15. dune: _____
16. deem: _____
17. depletion: _____
18. peculiar: _____
19. herein: _____
20. nonviable: _____
21. switch over: _____
22. cripple: _____
23. agile: _____
24. buffer: _____
25. stakes: _____

26. intelligentsia: _____

27. scriptures: _____

28. reign: _____

APPENDIX H

VOCABULARY RECOGNITION TASK - KEY

Please match the words given in Column B with the definitions given in Column A. Write the number you choose in the column provided. Some words are EXTRA in Column B.

Example:	*33*	*complaint*

	Column A
19	A. Able to move quickly and easily
9	B. A small bed for a baby, especially one that moves gently from side to side
12	C. A feeling of great respect and liking for someone or something
4	D. The sacred writings of Christianity contained in the Bible
16	E. Not practical
24	F. An event causing great and often sudden damage; a disaster
2	G. In this document or book
13	H. To prevent or prohibit (someone) from doing something or from going somewhere
22	I. To decrease seriously
27	J. To change over, or change around
15	K. The ability to understand or know something because of a feeling rather than by considering the facts
20	L. An amount of something left over when requirements have been met
8	M. To cause a severe problem for
25	N. To rule a nation or group of nations their king, queen, or emperor

Column B
1. fend off
2. herein
3. halve
4.scriptures
5. dune
6. mercenary
7. intruder
8. cripple
9. cradle
10. disperse
11. buffer
12. awe
13. barred from
14. engulf
15. intuition
16. nonviable
17. stakes
18. insurmountable
19. agile
20. surplus
21. peculiar
22. deplete
23. intelligentsia
24. calamity
25. reign
26. deem
27. switch over
28. demise

APPENDIX I

SUBJECTIVE COGNITIVE LOAD SCALES

Kullanıcı Adı (*User name):*

Mental Effort Scales

Aşağıda yaptığınız okuma aktivitesinin zorluğu ile ilgili bir derecelendirme ölçeği verilmiştir.
Below is given a rating scale on the difficulty of the reading activity you have just completed.

1. Parçayı okurken ne kadar zihinsel çaba harcadınız?
How much mental effort did you invest while reading the text?

Çok çok az zihinsel çaba harcadım/ Very very low mental effort								Çok çok fazla zihinsel çaba harcadım/ Very very high mental effort
1	2	3	4	5	6	7	8	9

Difficulty Scales

Aşağıda yaptığınız okuma aktivitesinin zorluğu ile ilgili bir derecelendirme ölçeği verilmiştir.
Below is given a rating scale on the difficulty of the listening activity you have just completed.

1. Okuduğunuz metin sizin için kolay mıydı?
Was the reading text easy for you?

Çok çok kolaydı/ Very very easy								Aşırı zordu/ Extremely difficult
1	2	3	4	5	6	7	8	9

APPENDIX J

TOPIC INTEREST QUESTIONNAIRE

Name-Surname: Student ID:

Measurement of (Text-Related) Topic Interest

The text you are going to read is about the collapse of civilizations. There are a number of theories behind the fall of a civilization. The text will discuss the reasons for the collapse of civilizations through some examples from previous civilizations.

1. While reading the text entitled "Societal Collapse" I expect to feel

	Not true at all	some-what	quite	completely true
Bored*	1	2	3	4
Stimulated[1]	1	2	3	4
Interested	1	2	3	4
Indifferent	1	2	3	4
Involved[2]	1	2	3	4
Engaged[3]	1	2	3	4

2. To me personally, the topic "Societal Collapse" is

	Not true at all	some-what	quite	very
Meaningful	1	2	3	4
Useful	1	2	3	4
Worthless*	1	2	3	4

[1]: filled with enthusiasm
[2]: If you are *involved* in a situation or activity, you are actively taking part in it.
[3]: intensely involved or occupied with an activity
*Reversed coding

APPENDIX K

INTERVIEW QUESTIONS

1) Do you prefer reading on screen or print reading?
2) Did you experience any difficulties while reading the text? Was it easy or difficult for you to understand? Why/ Why not?
3) Are you interested in this topic? Was the text interesting for you?
4) Was the electronic reading environment easy to use, navigate, or read?
5) Did you use the annotations given in the text? Were they helpful?
6) Were the upcoming tasks easy or difficult for you?
7) In which tasks did you experience particular difficulty?

REFERENCES

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.

AbuSeileek, A.F. (2008). Hypermedia annotation presentation: Learners' preferences and effect on EFL reading comprehension and vocabulary acquisition. *CALICO Journal, 25*(2), 260-275.

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,* 1281-1291.

Ainley, M. D. (1998). Interest in learning in the disposition of curiosity in secondary students: Investigating process and context. In L. Hoffman, A. Krapp, K. Renninger, & J. Baumert (Eds.), *Interest and learning: Proceedings of the Seeon conference on interest and gender* (pp. 257–266). Kiel, Germany: Institut für die Pädagogik der Naturwissenschaften.

Ainley, M. D., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94,* 545-561.

Akbulut, Y. (2007a). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science, 35,* 499–517.

Akbulut Y. (2007b). Variables predicting foreign language reading comprehension and vocabulary acquisition in a linear hypermedia environment. *TOJET, 6*(1), 53-59.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J., & Urquhart, A. (1988). This test is unfair: I'm not an economist. In P. Carrell, J. Define, & D. Eskey (Eds), *Interactive approaches to second language reading* (pp. 168-182). Cambridge, England: Cambridge University Press.

Alptekin, C., & Erçetin, G. (2009). Assessing the relationship of working memory to L2 reading: Does the nature of comprehension process and reading span task make a difference? *System, 37*(4), 627-639.

Alptekin, C., & Erçetin, G. (2010). The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading. *Journal of Research in Reading, 33*(2), 206-219.

Alptekin, C., & Erçetin, G. (2011). The effects of working memory capacity and content familiarity on literal and inferential comprehension in L2 reading. *TESOL Quarterly, 45*(2), 235-266.

Alptekin, C., Erçetin, G., & Özemir, O. (2014). Effects of variations in reading span task design on the relationship between working memory capacity and second language reading. *The Modern Language Journal, 98,* 536-552.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning and Technology, 5*(1), 202–232.

Al Shehri, S., & Gitsaki, C. (2010). Online reading: The impact of integrated and split-attention formats on L2 students' cognitive load. *ReCALL, 22* (3), 356-375.

Ariew, R., & Erçetin, G. (2004). Exploring the potential of hypermedia annotations for second language reading. *Computer-Assisted Language Learning, 17,* 237-259. http://dx.doi.org/10.1080/0958822042000334253

Atkins, P. W. B., & Baddeley, A. D. (1998). Working memory and distributed vocabulary learning. *Applied* Psycholinguistics, 19, 537–552

Ayres, P., & Sweller, J. (2005). The split attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135-146). New York: Cambridge University Press.

Baddeley, A.D. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences, 4,* 417–423.

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders, 36(3),*189–208.

Barnes, S. (1994). Hypertext literacy. *Interpersonal Computing and Technology, 2(4),* 24-36.

Bell, F.L. (2005). Comprehension aids, Internet technologies, and the reading of authentic materials by adult second language learners. *Dissertation Abstracts International, 66* (07), 2517. (UMI No. 3183039).

Bell, F. L. & LeBlanc, L. B. (2000). The language of glosses in L2 reading on computers: Learner preferences. *Hispania, 83*(2), 274-285.

Bellamy, C. (1999). The web, hypertext and history: A critical introduction. *The Electronic Journal of Australian and New Zealand History.*

Ben Salem, E. (2006). *The influence of electronic glosses on word retention and reading comprehension with Spanish language learners.* (Doctoral dissertation, University of Kansas). Retrieved from http://search.proquest.com/docview/304845561

Bernhardt, E. B. (1983). Testing foreign language reading comprehension: The immediate recall protocol. *Die Unterrichtspraxis, 16,* 27-33.

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

Bernhardt, E. B. & Kamil, M. L. (1995). Interpreting relationships between first language and second language reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics 16*, 1, 15-34.

Bray, G. B., & Barron, S. (2004). Assessing reading comprehension: The effects of text-based interest, gender, and ability. *Instructional Assessment, 9,* 107–128.

Brünken, R., Plass, J., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38,* 53–61.

Budd, D., Whitney, P., & Turley, K. J. (1995). Individual differences in working memory strategies for reading expository text. *Memory and Cognition, 23*(6), 735–748.

Bügel, K., & Buunk, B. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The Modern Language Journal*, *80,* 15–31.

Carrell, P. (1987). Content and formal schemata in ESL reading. *TESOL Quarterly, 21,* 461-481.

Carrell, P., & Wise, T. (1998). The relationship between prior knowledge and topic interest in second language reading. *Studies in Second Language Acquisition*, *20*, 285–309.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction,* 8, 293–332.

Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 151-170.

Chen, I.J. (2016). Hypertext glosses for foreign language reading comprehension and vocabulary acquisition: Effects of assessment methods. *Computer Assisted Language Learning*, *29*(2), 413-426.

Chen, I. J., & Yen, J.C. (2013). Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary acquisition in foreign languages. *Computers & Education, 63,* 416 423.

Cheng, Y.H. & Good, R. L. (2009). L1 glosses: Effects on EFL learners' reading comprehension and vocabulary retention. *Reading in a Foreign Language, 21*(2), 119-142.

Chun, D.M. (2006). CALL technologies for L2 reading. In L. Ducate & N. Arnold (Eds.), *Calling on CALL: From theory and research to new directions in foreign language teaching* (pp. 69–98). San Marcos, Texas: CALICO.

Chun, D. M. & Payne, J. S. (2004). What makes students click: working memory and look-up behavior. *System 32,* 481–504.

Chun, D. M. & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*(2), 183-198.

Chun, D. M., & Plass, J. (1997). Research on text comprehension in multimedia environment. *Language Learning & Technology, 1*(1), 60–81.

Clark, R. C., & Mayer, R. E. (2008). *E-learning and the science of instruction.* San Francisco: Pfeiffer.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3,* 149–210.

Coady, J. (1997). L2 vocabulary acquisition: A synthesis of the research. In J. Coady & T. Huckin (eds.), *Second language vocabulary* acquisition (pp. 273-290). Cambridge: Cambridge University Press.

Cohen, A. (1994). English for academic purposes in Brazil: The use of summary tasks. In C. Hill & K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 174-204). London: Longman.

Cohen, A. D. & Upton, T. A. (2006). Strategies in responding to the new TOEFL reading tasks. *Monograph 33,* Princeton, NJ: Educational Testing Service.

Conway, A. R. A., Kane, M. J., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review,* 12 (5), 769-786.

Cowan, N. (2005). *Working memory capacity.* Hove, UK: Psychology Press.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. *Educational Psychology Review, 26,* 197-223.

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

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior, 19,* 450-466.

Daneman, M. & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology, 130*(2), 208-223.

de Ridder, I. (2002). Visible or invisible links: Does the highlighting of hyperlinks affect incidental vocabulary learning, text comprehension, and the reading process? *Language Learning & Technology, 6*(1), 123-146.

de Ridder, I. (2003). *Reading from the screen in a second language: Empirical studies on the effect of marked hyperlinks on incidental vocabulary learning, text comprehension, and the reading process.* Antwerp: Garant.

DeStefano, D. & LeFevre, J.A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior, 23,*1616-1641.

de Vries, E. (2001). Hypermedia for Physics learning: The case of the energy concept. In J.F. Rouet, J. J. Levonen, & A. Biardeu (Eds), *Multimedia learning: Cognitive and instructional issues* (pp. 141-153). London: Pergamon.

Ellis, N. C. & Sinclair, S. G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology, 49*(1), 234-250.

Engle, R. W. (1989). *Working memory capacity: An individual differences approach.* Columbia, SC: AFOSR.

Erçetin, G. (2003). Exploring ESL learners' use of hypermedia reading glosses. *CALICO Journal, 20*(2), 261-283.

Erçetin, G. (2010). Effects of topic interest and prior knowledge on text recall and annotation use in reading a hypermedia text in the L2. *RECALL, 22*(2), 228–246.

Erçetin, G., & Alptekin, C. (2012). The explicit/implicit knowledge distinction and working memory: Implications for L2 reading comprehension. *Applied Psycholinguistics, 34,*727-753.

Fang, S. (2009). *Chinese gloss or English gloss: Which is more effective for incidental vocabulary acquisition through reading?* Retrieved from diva-portal.org

Flowerday, T., Schraw, G., & Stevens , J. (2004). The Role of Choice and Interest in Reader Engagement. *The Journal of Experimental Education, 72*(2), 93-114.

Foltz, P. W. (1993). Reading comprehension and strategies in linear text and hypertext. *ICS Technical Report, 93* (1), 7-84.

Foltz, P.W. (1996) Comprehension, Coherence and Strategies in Hypertext and Linear text. In Rouet, J.-F., Levonen, J.J., Dillon, A.P. & Spiro, R.J. (Eds.) *Hypertext and Cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Friedman, N.P., & Miyake, A. (2004) The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology, 133*(1), 101–135.

Garrett-Rucks, P., Howles, L., & Lake, W. (2015). Enhancing L2 reading comprehension with hypermedia texts: Student perceptions. *CALICO Journal, 32*(1), 26-51.

Geranpayeh, A. (2003). A quick review of the English Quick Placement Test. *Research Notes, 12,* 8-10.

Gettys, S., Imhof, L. A., & Kautz, J. O. (2001). Computer-assisted reading: The effect of glossing format on comprehension and vocabulary retention. *Foreign Language Annals, 34,* 91-106.

Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior, 23*(4), 467–486.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40,* 97–118.

Goodman, K. S. (1967). Reading: A Psycholinguistic Guessing Game. *Journal of the Reading Specialist 6,* 126-35.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly, 25*(3), 375-406.

Grabe, W. (2004). Research on teaching reading. *Annual Review of Applied Linguistics, 24,* 44–69.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York: Cambridge University Press.

Grabe, W. & Stoller, F. L. (2001). Reading for Academic Purposes: Guidelines for the ESL/EFL Teacher. In M. Celce-Murcia (ed.), *Teaching English as a Second or Foreign Language* (3rd Ed., pp. 187-203). Boston: Heinle & Heinle.

Grabe, W. & F. L. Stoller (2002). *Teaching and researching reading.* Harlow: Longman.

Guthrie, T. Wigfield, A., Humenick, N. M., Perencevich, K. C.,Taboada, A. & Barbosa, P. (2006). Influences of stimulating tasks on reading motivation and comprehension. *The Journal of Educational Research, 99*(4), 232-245.

Harrington, M., Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in Second Language Acquisition 14,* 25–38.

Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M., Juffs, A. & Wilson, L. (2010). Personalization of Reading Passages Improves Vocabulary Acquisition. *International Journal of Artificial Intelligence in Education 20,* 73-98

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research, 60,* 549–571.

Hidi, S. (1995). A re-examination of the role of attention in learning from text. *Educ. Psychol. Rev.* 7, 323–350.

Hidi, S.E. & McLaren, J. A. (1990). Motivational factors and writing: the role of topic interestingness. *European Journal of Psychology of Education, 6*(2), 187–197.

Hidi, S., and Harackiewicz, J. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 151–179.

Horiba, Y. & Fukaya, K. (2015). Reading and learning from L2 text: Effects of reading goal, topic familiarity, and language proficiency. *Reading in a Foreign Language, 27*(1), 22-46.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11,* 207-223.

Hudson, T. (1982). The effects of induced schemata on the "short circuit" in L2 reading: Non-decoding factors in L2 reading performance. *Language Learning, 32*, 3-31.

Hudson, T. (2007). *Teaching second language reading*. New York: Oxford University Press.

Hulleman, C. S., Durik, A. M., Schweigert, S. A., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology, 100,* 398–416.

Hulstijn, J.H. (1997). Mnemonic methods in foreign-language vocabulary learning: Theoretical considerations and pedagogical implications. In: J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 203-224). Cambridge, UK: Cambridge University Press.

Hulstijn, J.H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258-286). Cambridge: Cambridge University Press.

Hulstijn, J. H. & Laufer, B. (2001). Some empirical evidence for the Involvement Load Hypothesis in vocabulary acquisition. *Language Learning, 51* (3), 539-558.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal, 80,* 327–339.

Hyperlink. (2016). In *Merriam-Webster Online Dictionary.* Retrieved from https://www.merriam-webster.com/dictionary/hyperlink

Jacobs, G. M., Dufon, P., & Hong, F. C. (1994). L1 and L2 vocabulary glosses in L2 reading passages: Their effectiveness for increasing comprehension and vocabulary knowledge. *Journal of Research in Reading, 17*(1), 19–28.

Jenkins, J. R., & Dixon, R. (1983). Vocabulary learning. *Contemporary Educational Psychology, 8*(3), 237–260.

Johnson, P. (1982). Effects on reading comprehension of language complexity and cultural background of text. *TESOL Quarterly*, *16*, 169–181.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133.

Jonassen, D. H., & Wang, S. (1993). Acquiring structural knowledge from semantically structured hypertext. *Journal of Computer-Based Instruction, 20(1),* 1–8.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329-354.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122-149.

Kalyuga, S. (2008). *Managing Cognitive Load in Adaptive Multimedia Learning.* New York: Information Science Reference.

Kalyuga S. (2011). Cognitive load theory: How many types of load does it really need? *Educ. Psychol. Rev. 23*, 1–19. 10.1007/s10648-010-9150-7

Kalyuga, S., Ayres, P., Chandler, P. & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–32.

Khalifa, H. & Weir, C. (2009). *Examining Reading: Research and Practice in Assessing Second Language Reading*. Cambridge: Cambridge University Press.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95,* 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review,* 85, 363–394.

Kintsch, W., & Welsch, D. M. (1991). The construction– integration model: A framework for studying memory for text. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennett B. Murdock* (pp. 367–385). Hillsdale, NJ.: Erlbaum.

Kirschner, P. A. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learning and Instruction*, *12*(1), 1-10.

Kirschner, P., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory: The good, bad and the ugly. *Computers in Human Behavior, 27,* 99–105.

Knight, S. (1994). Dictionary use while reading: The effects on comprehension and vocabulary acquisition for students of different verbal abilities. *The Modern Language Journal, 78,* 285-299.

Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach.* Cambridge: Cambridge University Press.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. In K. Koda (Ed.), *Reading and language learning*, (pp.1-44). Special issue of *Language Learning Supplement, 57.*

Krapp, A. (1999). Interest, motivation and learning: An educational– psychological perspective. *European Journal of Psychology of Education, 14,* 23–40.

Krapp, A. (2002). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction, 12,* 383–409.

Krashen, S. (1995) *Principles and practice in second language acquisition.* New York: Phoenix.

Krueger, R., A. (2015). *Focus Groups: A Practical Guide for Applied Research.* Thousand Oaks, CA: SAGE.

Landow, G. P. (1992). *Hypertext: The convergence of contemporary critical theory and technology.* Baltimore and London: Johns Hopkins University Press.

Laufer, B. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge: Cambridge University Press.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

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

Laufer, B. & Kalovski- Ravenhorst, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22* (1), 15-30.

Lee, J. F. (1986). Background knowledge and L2 reading. *Modern Language Journal, 70,* 350-354.

Lee, J.F. & Riley, G.L. (1990). The effect of pre-reading, rhetorically-oriented frameworks on the recall of two structurally different expository texts. *Studies in Second Language Acquisition, 12,* 25-41.

Leeser, M. J. (2007). Learner-based factors in L2 reading comprehension and processing grammatical form: Topic familiarity and working memory. *Language Learning, 57,* 229–270.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2013). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bull Rev., 21,* 861-883.

Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *Journal of Documentation, 61*(6), 700-712. doi: 10.1108/00220410510632040

Lomicka, L. (1998). To gloss or not to gloss: An investigation of reading comprehension online. *Language Learning & Technology, 1* (2), 41-50. Retrieved from http://llt.msu.edu/vol1num2/article2/

Magliano, J. P, Durik, A. M., & Holt, J. K. (2011). Assessing the impact of topic interest on comprehension processes. *Law School Admission Council Research Report 11*-02.

Martinez, M. M. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*, 1–19.

Mayer, R. E. (2001). *Multimedia learning.* Cambridge: Cambridge University Press.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psycholo*gist, *63*(8), 760-769.

Mayer, R. E. & Moreno, R. (1998). Split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90* (2), 312-320.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1-43.

Merriam, S., B. (2009). *Qualitative Research: A Guide to Design and Implementation*. San Francisco: Jossey Bass.

Merrill, M.D. (2002). First principles of instruction. *Educational Technology, Research and Development, 50,* 43-59.

Miall, D. S., & Dobson, T. (2001). Reading hypertext and the experience of literature. *Journal of Digital Information, 2(*1), 1–20.

Miller. G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81-97.

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development, 53,* 35–45.

Morrison, J. (2004). *Reducing the cognitive load presented by definition presentation in electronic learning environments through the use of hypermedia rollovers* (Unpublished PhD thesis). University of Central Florida, Orlando, FL.

Nagy, W. & Anderson, R. (1984). How many words are there in school printed English? *Reading Research Quarterly, 19*(3), 304-330.

Nagy, W. E., Herman, P. A., & Anderson, R.C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233-253.

Nassaji, H. (2007). Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives. *Language Learning, 57,* 79-113.

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

Nation, P. and Wang, K. (1999) Graded readers and vocabulary. *Reading in a Foreign Language 12*(2), 355-380.

Nuttall, C. (2005). *Teaching reading skills in a foreign language.* Oxford: Macmillan Education.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal, 44,* 161–196.

Ozgungor, S., & Guthrie, J. T. (2004). Interactions among elaborative interrogation, knowledge, and interest in the process of constructing knowledge from text. *Journal of Educational Psychology, 96,* 437–443.

Paas, F. G. W. C. (1992) Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *Journal of Educational Psychology, 84,* 429-434.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory instructional implications of the interaction between information structures and cognitive architecture. *Instructional Science, 32,* 1–8.

Paas, F., Tuovinen, J. E., Tabbers, H. & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63-71.

Paas, F. & Van Merriënboer, J. J. G. (1994) Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach, *Journal of Educational Psychology 86,* 122–133.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Computers in Human Behavior, 19*, 221-243.

Perkins, K., Brutten, S. R., & Pohlmann, J. T. (1989). First and second language reading comprehension. *RELC Journal*, *10*(2), 1–9.

Potelle, H., & Rouet, J. F. (2003). Effects of content representation and readers' prior knowledge on the comprehension of hypertext. *International Journal of Human-Computer Studies, 58*(3), 327–345.

Psychological Corporation. (1997). *WAISIII WMS-III technical manual*. San Antonio, TX: Harcourt Brace.

Pulido, D. (2000). *The impact of topic familiarity, L2 reading proficiency, and L2 passage sight vocabulary on incidental vocabulary gain through reading for adult learners of Spanish as a foreign language.* (Unpublished PhD thesis). University of Illinois, Urbana-Champaign.

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning, 53,* 233–284.

Pulido, D. (2004). The effect of cultural familiarity on incidental vocabulary acquisition through reading. *The Reading Matrix: An Online International Journal, 4,* 20–53.

Pulido, D. C. (2007). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning, 54,* 155-199.

Reader, W. & Hammond, N. (1994). Computer-based tools to support learning from hypertext: Concept mapping tools and beyond. *Computers & Education, 22(*1-2), 99-106. doi: 10.1016/0360-1315(94)90078-7

Riley, G.L., & Lee, J.F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing, 13*(2), 173-190.

Roby, W. (1999). "What Is in A Gloss?" *Language Learning & Technology, 2*(2), 94-101.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition, 21,* 589–619.

Rouet, J.F. & Levonen, J. J. (1996). Studying and learning with hypertext: Empirical studies and their implications. In J.F. Rouet, J. J. Levonen, A. Dillon, & R.J.

Spiro (Eds), *Hypertext and Cognition* (pp. 9-23). Manwah, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E. (2004). Toward an Interactive Model of Reading. In R. B. Rumelhart & N. J. Unrau (Eds.). *Theoretical Models and Processes of Reading* (5th Ed., pp. 1149-1179). Newark, Delaware: International Reading Association.

Şakar, A., & Ercetin, G. (2004). Effectiveness of hypermedia annotations for foreign language reading. *Journal of Computer Assisted Learning, 21,* 28-38.

Salden, R. J. C. M., Paas, F., Broers, N. J. & van Merrienboer, J. J. G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science, 32,* 153-172.

Sawaki, Y. (2003). *A comparison of summary and free-recall as reading comprehension tasks in a web-based placement test of Japanese as a foreign language.* (Unpublished PhD thesis). University of California, Los Angeles.

Schiefele, U. (1990). The influence of topic interest, prior knowledge, and cognitive capabilities on text comprehension. In J. M. Pieters, K. Breuer & P. R. J. Simons (Eds). *Learning Environments*. Berlin: Springer.

Schiefele, U. (1996). Topic interest, text representation, and quality of experience. *Contemporary Educational Psychology, 21,* 3–18.

Schiefele, U. (2009). Situational and individual interest. In K. R. Wentzel & A. Wigfield (Eds), *Handbook of motivation at schools* (pp. 197-222). New York: Routledge.

Schiefele, U. (2012). Interest and learning. In N. M. Seel (Ed.), *Encyclopedia of Sciences of Learning* (pp.1623-1626). New York: Springer.

Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences, 8,* 141–160.

Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). A closer look at split visual attention in system- and self-paced instruction in multimedia learning. *Learning and Instruction, 20,* 100-110.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 129- 158.

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

Schüler, A., Scheiter, K., & van Genuchen, E. (2011). The role of working memory in multimedia instruction: Is working memory working during learning from text and pictures? *Educational Psychology Review, 23,* 389-411.

Seigneuric, A., Ehrlich, M. F., Oakhill, J. V., & Yuill, N. M. (2000). Working memory resources and children's reading comprehension. *Reading and Writing, 13*(1–2), 81–103.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika, 52* (3/4), 591-611.

Shapiro, A.M., & Niederhauser, D. (2004). Learning from hypertext: Research issues and findings. In D.H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2nd ed.) (pp. 605-620). Mahwah, NJ: Erlbaum.

Shohamy, E. G. (2001). *The power of tests: A critical perspective on the uses of language tests.* New York: Longman.

Sinatra, G. M., K. J. Brown & R. E. Reynolds (2002). Implications of cognitive resource allocation for comprehension strategies instruction. In C. C. Block and M. Pressley (eds.), *Comprehension instruction,* (pp. 62-76). New York: The Guilford Press.

Singer, L. M. & Alexander, P. A. (2017). Reading across mediums: Effects of reading digital and print texts on comprehension and calibration. *The Journal of Experimental Education, 85*(1), 155–172.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Smith, J. & Weiss, S. (1988). Hypertext. *Communications of the ACM, 31 (7),* 816–819.

Snow, C. (2002). *Reading for understanding: Toward an R&D Program in Reading*. CA: RAND.

Spiro, R.J., Feltovich, P.J., Jacobson, M.J., & Coulson, R.L. (1992). Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. *Educational Technology*, 31(5), 24-33.

Stahl, S. A. (2003). Vocabulary and readability: How knowing word meanings affects comprehension. *Topics in Language Disorder, 23* (3), 241–247.

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

Stanovich, K. E. (1980). Toward an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly 16*, 32-71.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly,21*(4), 360- 407.

Stromso, H. I. & Braten, I. (2002). Norwegian law students' use of multiple resources while reading expository texts. *Reading Research Quarterly, 37*(2), 208-27.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295-312.

Sweller, J. (2004). Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. *Instructional Science, 32,* 9-31.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 19 –30). New York: Cambridge University Press.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22,* 123- 138. doi:10.1007/s10648-010-9128-5

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction, 8,* 351-362.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12,* 185-233.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive Load Theory*. New York: Springer.

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review,* 10(3), 251–296.

Taylor, A. (2006). The effects of CALL versus traditional L1 glosses on L2 reading comprehension. *CALICO Journal, 23*(2), 309-318.

Tuovinen, J., & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science, 32,* 133– 152.

Turner, M. L. & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28*, 127-154.

Türk, E. & Erçetin, G. (2014). Effects of interactive versus simultaneous display of multimedia glosses on L2 reading comprehension and incidental vocabulary learning. *Computer Assisted Language Learning, 27* (1), 1-25, doi: 10.1080/09588221.2012.692384

Urquhart, S. & Weir, C. (1998). *Reading in a second language: Process, product, and practice*. New York: Longman.

van Dijk, T. A. & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

van Merrienboer, J.J.G., & Kirschner, P.A. (2001). Three worlds of instructional design: State of the art and future directions. *Instructional Science, 29*, 429-441.

van Merrienboer J.J.G., Schuurman, J. G., de Crook, M. B. M. & Paas, F. G. W. C. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction, 12,* 11-37.

van Merrienboer, J.J.G., Kirschner, P.A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational Psychologist, 38* (1), 5-13.

Wallen, E. S. (2002). *The effect of different types of hypertext annotations on vocabulary recall, text comprehension, and knowledge transfer in learning from scientific texts* (Unpublished doctoral dissertation). The University of New Mexico.

Wallen, E., Plass, J. L., & Brünken, R. (2005). The function of annotations in the comprehension of scientific texts: Cognitive load effects and the impact of verbal ability. *Educational Technology Research and Development, 53*(3), 59-71.

Waniek, J. (2012). How information organisation affects users' representation of hypertext structure and content. *Behavior and Information Technology, 31*(2), 143-154.

Waniek, J.,Brunstein, A.,&Naumann.A. (2003).Why should we use Eye Tracking for Hypertext Design? *MMI-Interaktiv, 6,* 35-40.

Waniek, J., Brunstein, A., Naumann, A. & Krems, J. (2003). Interaction between text structure representation and situation model in hypertext reading. *Swiss Journal of Psychology. 62* (2). 103-111.

Waring, R., & Nation, I. (2004). Second language reading and incidental vocabulary learning. *Angles on the English Speaking World, 4,* 97-110.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader *Reading in a Foreign language, 15*(2), 130-163.

Watanabe, Y. (1997). Input, intake and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19,* 287–307.

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20*(2), 232-245.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Xu, X. (2010). The effects of glosses on incidental vocabulary acquisition in reading. *Journal of Language Teaching and Research, 1*(2), 117-120.

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology, 13*, 48–67.

Yao, Y. (2006). *The effect of different presentation formats of hypertext annotations on cognitive load learning and learner control.* (Unpublished PhD thesis). University of Central Florida.

Yeh, Y., & Wang, C. (2003). Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *CALICO Journal 21*(1), 131-144.

Yeung, A. S. (1999). Cognitive load and learner expertise: Split-attention and redundancy effects in reading comprehension tasks with vocabulary definitions. *Journal of Experimental Education, 67*(3), 197-217. http://doi.org/cdtw89

Yeung, A. S., Jin, P., & Sweller, J. (1997). Cognitive load and learner expertise: Split-attention and redundancy effects in reading with explanatory notes. *Contemporary Educational Psychology, 23*(1), 1-21.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning and Technology, 10*(3), 85-101.

Yoshii, M. (2013). Effects of Gloss Types on Vocabulary Learning through Reading: Comparison of Single Translation and Multiple-choice Gloss Types. *CALICO Journal, 30*, 203-229.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning, 24*(1), 39-58.

Zumbach, J. (2006). Cognitive overhead in hypertext learning re-examined: overcoming the myths. *Journal of Educational Multimedia and Hypermedia*, 15(4), 411–432.