ABSTRACT

Title of Dissertation:	THE EFFECTS OF TRAINING FOR AUTOMATICITY ON MULTIWORD RECOGNITION
	Hyojin Jeong, Doctor of Philosophy, 2020
Dissertation directed by:	Dr. Robert M. DeKeyser, Second Language Acquisition

Automatic lexical processing, such as accessing and retrieving the correct meaning of a word quickly and automatically in reading, is fundamental for successful and fluent language use. However, the focus of the previous studies has been limited to the processing of single words. No study has investigated whether skilled recognition of newly learned multiword expressions, such as *cook the books*, can be fostered by training as predicted by skill acquisition theory. The present study, therefore, investigated the effects of multiword-recognition training on the development of automaticity in collocation-recognition processing, and also examined the roles of first language (L1) congruency (having word-for-word translation equivalents in L1), training condition (verbatim repetition vs. varied repetition), working memory capacity, and procedural learning ability in automatization of collocation recognition.

Seventy-two Korean learners of English were assigned to one of two experimental conditions (verbatim repetition and varied repetition conditions). All the participants engaged in three training sessions. The training consisted of a series of fill-in-the-blank exercises that only differed in terms of training materials. The target items for the training consisted of 16 congruent (corresponding to the participants' L1) and 16 incongruent verb-noun phrases in English. Two tests (phrasal decision task and sentence-and-word recall task) were used to assess the participants' learning gains from the training.

In line with the predictions of skill acquisition theory, the results showed that repeated practice through explicit vocabulary exercises led to significant improvement in recognition performance of collocations in terms of speed, processing stability and cognitive capacity usage. L1 congruency was found to play a facilitative role in collocation-recognition speed only in the early stages of learning. No reliable evidence for differential effects of the two types of training condition on developing automaticity in collocation recognition was found in the current study. Working memory played a facilitative role in collocation-recognition performance as measured by the sentence-and-word recall task but not in collocation-recognition speed or processing stability as measured by the phrasal decision task. On the other hand, no reliable evidence for a facilitative role of procedural learning ability in automatization of collocation recognition was found.

THE EFFECTS OF TRAINING FOR AUTOMATICITY ON MULTIWORD RECOGNITION

by

Hyojin Jeong

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2020

Advisory Committee:
Professor Robert M. DeKeyser, Chair
Professor Kira Gor
Professor Nan Jiang
Professor Steven J. Ross
Professor Yasmeen Faroqi-Shah

© Copyright by
Hyojin Jeong
2020

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my advisor, Dr. Robert DeKeyser, for his invaluable support and guidance, detailed comments and insightful suggestions throughout the completion of my studies. He is a great role model for me as a scholar and a researcher who is always responsive and reliable whenever students have questions or need advice and who has great passion and commitment to advancing the field of second language acquisition. I feel very fortunate to have worked with such a wonderful mentor who greatly inspires his students.

I am also truly grateful to Dr. Nan Jiang for his advice and support throughout the PhD process and especially for introducing me to the fascinating issues of lexical processing and development. I have become very interested in these topics and have continued to explore them since I joined the program. My sincere thanks also go to Dr. Kira Gor for her valuable and incisive comments and kind encouragements. I am also very grateful to Dr. Steven Ross for his valuable comments and advice on statistical analyses. I would like to express my thanks to Dr. Yasmeen Faroqi-Shah for being the Dean's representative on my dissertation committee.

I would also like to thank my classmates and friends in the SLA program who helped me complete this journey with joy and delight. My thanks especially go to Jiyong Lee, Kyoko Kobayashi Hillman, Yoonjee Hong, and Qian Zhou who provided great friendship and fruitful discussions. I would also like to thank Sun Young Ahn,

Jihye Moon, and Sunyoung Lee-Ellis, alumnae in the SLA program, for their advice and warm encouragements at different stages of my graduate school years.

I am also grateful to members and faculty of SLLC and the Korean program. My special thanks go to Yuk Fan Tai, Dr. Lauretta Clough, Josiland Chambers, and David Watson who always provided great help and assistance throughout my study at UMD. I am also truly grateful to Dr. Robert Ramsey and Prof. Younghi Ramsey for giving me invaluable opportunities to teach Korean language courses, which allowed me to realize the great joy of teaching Korean and inspired me to conduct a research on advanced learners of Korean. I owe special thanks to my M.A. advisor, Dr. Haemoon Lee, who introduced me to the exciting field of SLA and has always been a supportive and inspiring mentor from the very beginning of this long journey.

My deepest gratitude goes to my family. I would like to thank my parents and parents-in-law for their unwavering support and encouragement. I am especially indebted to my sister who helped me with participant recruitment and development of materials for my past projects as well as dissertation research. My sincere thanks also go to my brother who has always been encouraging and caring. And last but not least, my deepest thanks go to my husband, Daeyoon, for his understanding and belief in my abilities and endless support and encouragement throughout this journey. I would not have made through this journey without him.

Table of Contents

Acknowledgements ii
Table of Contents iv
List of Tables vi
List of Figures vii
Chapter 1: Introduction 1
Chapter 2: Review of the Literature and Theoretical Framework 5
Defining collocations 5
L2 collocation learning 7
Skill acquisition theory 11
Automaticity of lexical recognition in L2 reading 12
Important criteria for automatic lexical access and their measurement 14
Empirical studies of automatic word recognition in SLA 20
Types of repetition and collocation learning 28
The relationship between L1 congruency and collocation learning 32
The relationship between cognitive abilities and collocation learning 36
Chapter 3: Research Questions and Hypotheses 41
Chapter 4: Methods 46
Design of the study 46
Participants 48
Materials 49
Target items 49
Sentences used for the training 52
Training 53
Tests 55
Phrasal decision task 55
Sentence-and-word recall task 58
Operation span task 61
Probabilistic SRT task 62
Procedure 63
Chapter 5: Results 68
Individual difference measures 71
Language learning outcomes 72
Phrasal decision task 72
Sentence-and-word recall task 85
Cognitive aptitude and learning outcomes 92
Working memory capacity and learning outcomes 92
Procedural learning ability and learning outcomes 95
Chapter 6: Discussion 97
Development of automaticity in collocation recognition 97
Research question 1 97
Research question 2 102
Research question 3 109
Cognitive aptitude and skilled recognition of a collocation 112

Research question 4 112
Research question 5 116
Chapter 7: Conclusion 118
Appendices 122
Appendix A. Target collocations 122
Appendix B. English word and nonword stimuli for the L2 LDT 123
Appendix C. Training materials 124
Appendix D. Stimuli in sentence-and-word recall task 137
Bibliography 143

List of Tables

Table 1. Knowledge of collocations 8
Table 2. Schedule of tests and training sessions 46
Table 3. Description of the participants 48
Table 4. Lexical properties of the stimuli (16 items in each of the two conditions) .. 51
Table 5. Counterbalanced sequences of training sets for the varied repetition condition 66
Table 6. Overview of experimental sessions 67
Table 7. Reliability of the learning outcome measures (N=72) 69
Table 8. Descriptive statistics of the individual difference measures (N=72) 71
Table 9. Means (and SDs) of RT (in milliseconds), SD, CV, accuracy, correlation between CV and RT (phrasal decision task) 74
Table 10. Means and p-values for RT differences across training sessions 76
Table 11. Means and p-values for CV differences across training sessions 80
Table 12. Mean scores (and SDs) of the two groups on sentence recall (sentence-and-word recall task) 85
Table 13. Mean scores (and SDs) of the two groups on word recall (sentence-and-word recall task) 89
Table 14. Pearson correlations of working memory scores with RT and CV 93
Table 15. Pearson correlations of working memory scores with accuracy 94
Table 16. Pearson correlations between sentence and word recall scores, and working memory scores 94
Table 17. Pearson correlations between RT and CV, and procedural learning ability scores 95
Table 18. Pearson correlations between accuracy and procedural learning ability scores 96
Table 19. Pearson correlations between sentence and word recall scores, and procedural learning ability scores 96

List of Figures

Figure 1. Example of a fill-in-the-blank exercise 55
Figure 2. Mean logRT and 95% confidence intervals for two collocation types across training sessions for the verbatim repetition group (left) and the varied repetition group (right) 77
Figure 3. Mean CV and 95% confidence intervals for two collocation types across training sessions for the verbatim repetition group (left) and the varied repetition group (right) 81
Figure 4. Mean sentence recall scores and 95% confidence intervals for the verbatim repetition group (left) and the varied repetition group (right) 85
Figure 5. Mean sentence recall scores and 95% confidence intervals for the verbatim repetition group (left) and the varied repetition group (right) 89

Chapter 1: Introduction

Language users carry out many cognitively complex activities such as reading a newspaper, watching TV and engaging in a conversation in daily life. These activities involve simultaneous interplay of lower (e.g., processing letters and speech sounds) and higher-level linguistic processes (e.g., comprehending the content of text or discourse). However, given a limited cognitive capacity, language users cannot attend to all processes simultaneously to the same high extent. Thus, it is important that some of those sub-processes are carried out with ease and speed, being free of interference from other sources of information. This automaticity enhances the fluidity and efficiency of underlying cognitive processing and improves the quality of performance as a whole (Segalowitz, 2000, 2010).

As automatic language processing is fundamental for successful and fluent language use, it is not surprising that automaticity in language processing has long been of interest to many L2 scholars. Of particular relevance and importance to the discussion of automaticity in language processing is skill acquisition theory (e.g., DeKeyser, 1997, 2007, 2015; McLaughlin, 1989), which accounts for how learning proceeds from initial representation of knowledge to eventual skilled behavior. Specifically, the theory posits that knowledge develops from declarative knowledge (knowledge "that") to procedural and automatized knowledge (knowledge "how") through repeated practice and increased experience. This process of automatization is seen to follow the power law of practice where response times and error rates follow a negative exponential as a function of the number of trials, gradually decreasing to

an asymptotic point (e.g., Newell & Rosenbloom, 1981), and the resulting production of automatic processing is commonly regarded as effortless and fast, in line with Anderson's adaptive control of thought theory (e.g., Anderson, 1992, 2013) and Logan's instance theory (Logan, 1988).

To date, the predictions of skill acquisition theory have been put to empirical test with different types of linguistic performance in focus such as word recognition and retrieval, acquisition of grammar rules, and sentence processing (e.g., De Jong & Perfetti, 2011; Li & DeKeyser, 2017; Lim & Godfroid, 2015; Solovyeva & DeKeyser, 2018). In particular, a large number of studies have focused on automatic word recognition and demonstrated how skilled word recognition is reflected in the decrease of the amount and/or variability of response time needed to perform such a task over the course of practice (e.g., Akamatsu, 2008; Fukkink, Hulstijn, & Simis, 2005; S. J. Segalowitz, & N. Segalowitz, 1993). However, as will be reviewed in a later section, the focus of these word recognition studies has been limited to processing of single words. No study has investigated whether skilled recognition of multiword units can be fostered by training as predicted by skill acquisition theory.

Multiword expressions constitute a large proportion of speech and written communication. Erman and Warren (2000), for example, reported that such multiword sequences account for somewhat more than 50% of both written and spoken texts or even 70% of any given text (Wray & Perkins, 2000). Among different types of multiword units, collocations have been increasingly researched and started to be recognized as an important part of vocabulary knowledge, playing a crucial role

in enabling the comprehension and expression of messages (Boers & Lindstromberg, 2012). Kremmel, Brunfaut, and Alderson (2015), for example, reported a study that provided evidence for the significant role of phraseological knowledge in L2 reading. They found that phraseological knowledge was the strongest of three contributors (the others being syntactic and vocabulary knowledge) to reading performance of L2 learners. Martinez and Murphy (2011) also demonstrated how knowledge of formulaic expressions is important for adequate text comprehension. Their data showed that formulaic expressions can cause major difficulties for L2 learners when reading. Participants' reading comprehension was adversely affected by the formulaic expressions in the text, which were comprised of very common individual words (e.g. *by and large*), compared to that of a text with the exact same high-frequency individual words. In addition to the important role of phraseological knowledge in L2 reading, the importance of having automatic access to multiword units has been highlighted in the literature (e.g., Boers & Lindstromberg, 2012; Wood, 2009), suggesting its facilitative role in production of fluent speech. In addition, similarly to the role of automatic single-word recognition, automatic use of acquired multiword expressions plays an important part in language processing, allowing chunking and freeing up memory and processing resources (Ellis, 1996; Kuiper, 1995). Ellis (1996) further suggested that this enables an individual to attend to the other aspects of language (e.g., form) rather than being constrained by processing limitations to concentrate on the merely semantic aspect and may even lead to the abstracting of

regularities in those recurrent patterns for the construction of grammar knowledge (for a similar argument see Segalowitz, 2003).

In summary, given the ubiquity of formulaic expressions and their crucial roles in language use (e.g., Martinez & Murphy, 2011), automatic processing of multiword units may be equally or even more crucial for accurate and fluent language use than single-word processing. The present study, therefore, examined the effects of training for automaticity on L2 collocation knowledge. In the following, a review of the literature and theoretical framework on L2 collocation learning and automatization of lexical recognition will be presented.

Chapter 2: Review of the Literature and Theoretical Framework

Defining collocations

The term "multiword units" has been commonly used as an umbrella term to refer to a wide range of word sequences such as collocations, idioms, and lexical bundles. However, each category has its own unique characteristics and behaviors. Thus, it is important to differentiate these subcategories of multiword units to fully understand and identify the association between different factors and the development of different types of multiword units.

Collocations, in particular, have been defined with two major approaches, phraseological and statistical approaches. The phraseological approach, for example, views collocations as word combinations with selectional restrictions (Cowie, 2001; Howarth, 1998). Under this approach, collocations are identified based on grammatical structure and semantic properties (restrictions). For example, the sense of the verb *take* is restricted as it does not collocate with a *film* in a collocation, but it does with *picture* in *take a picture*.

The statistical approach, on the other hand, uses various statistics such as frequency of occurrence in a corpus, the T-score, or the mutual information (MI) score to identify a collocation. Basically, these measures estimate "the extent to which certain words "attract" one another in authentic language use" (Wolter & Yamashita, 2015, p. 1194). The statistical approach is simpler and more commonly used (e.g., Durrant & Doherty, 2010; Durrant & Schmitt, 2010; Ellis & Frey, 2009;

Ellis, Simpson-Vlach, & Maynard, 2008; Sonbul, 2015; Sonbul & Schmitt, 2013; Wolter & Yamashita, 2015, 2018). These studies suggested that collocations identified using aforementioned statistics have psycholinguistic reality (i.e., offer processing advantages). The effects tend to be more robust for native speakers while less so for nonnative speakers, often showing interactions with other variables such as L2 proficiency level. Based on meta-analysis of 19 collocational testing studies, Durrant (2014) reported that various variables (e.g., phrasal frequencies, MI, and T-scores) obtained from the corpus (e.g., COCA) reliably predicted L2 learners' knowledge of collocations (as measured by ranges of recognition and production tests).

Adopting a definition that is more in line with the statistical approach, in the present study, collocations are defined as grammatically well-formed combinations of words that co-occur frequently (Kjellmer, 1994). These are likely to include collocations as defined in the phraseological approach (e.g. *shrug* is statistically highly likely to co-occur with *shoulders*) but also include combinations which do not involve semantic specialization or restriction (e.g. *next week, drink tea*). However, lexical patterns that may be identified as collocations based solely on the frequency-based method, such as *but too* and *too was*, will not be considered as meaningful collocations.

Most collocations do not have a high phrasal frequency (Webb, Newton, & Chang, 2013) and are sequences of two words with a clear syntactic function such as noun phrases or verb and object constructions, and thus are different from lexical

bundles which are mostly units of three words or more with no syntactic completeness (e.g., *that there is no*). In terms of compositionality, collocations often vary between combinations that are fully decomposable (e.g., *pay a salary*) and patterns that involve figurative meaning (e.g., *pay attention*). Thus, it is often very difficult to strictly draw a line between collocations and idioms; however, Grant and Bauer (2004) argued that the two criteria must be met in order for a multiword unit to be considered as a true idiom: non-compositionality of more than one element and non-figurativeness in a sense that intended meaning cannot be understood by stretching the known meaning of the individual words (e.g., *red herring*). Note that collocations often involve only one element used in non-literal sense with figurative meaning, and thus differ from idioms.

L2 collocation learning

Collocation learning is similar to individual word learning in that its most crucial aspect is making the connection between the forms and their appropriate meaning. However, this is not the only aspect of knowledge of collocation. Nation and Webb (2011) identified nine aspects of knowledge of multiword units, each of which is applied to collocations in Table 1 (Webb, Newton, & Chang, 2013). Based on this conceptualization, along with form-meaning mapping, learning which word goes with which in a collocation is also essential to both receptive and productive language use. Wray (2000) also highlighted the importance of knowledge of form in multiword lexis which allows learners to improve levels of accuracy and fluency.

Developing knowledge of the composition of a collocation is particularly important at early stages of learning to reduce the risk of establishing undesirable associations that learners may find hard to suppress later on. For example, although *say* and *tell* have similar meanings, *prayer* is a collocate of *say* but not *tell*, while *lie* is a collocate of *tell* but not *say*. In the current study, a lexical item is acquired if its representations are established and incorporated into the mental lexicon and if these representations can be accessed in an automatic manner (Elgort, 2011). Representational knowledge includes form and meaning of vocabulary items.

Table 1. Knowledge of collocations

Form	spoken	R	What does the collocation sound like?
		P	How is the collocation pronounced?
	written	R	What does the collocation look like?
		P	How is the collocation written and spelled?
	word parts	R	What words are recognizable in this collocation?
		P	What words are needed to express the meaning?
Meaning	form and meaning	R	What meaning does this collocation signal?
		P	What collocation can be used to express this meaning?
	concept and referents	R	What is included in the concept?
		P	What items can the concept refer to?
	associations	R	What other words or collocations does this make us think of?
		P	What other words or collocations could we use instead of this one?
Use	grammatical functions	R	In what patterns does the collocation occur?
		P	In what patterns must we use this collocation?
	collocations	R	What words, collocations, or types of collocations occur with this one?
		P	What words, collocations, or types of collocations must we use with this one?
	constraints on use (register, frequency …)	R	Where, when, and how often would we expect to meet this collocation?
		P	Where, when, and how often can we use this collocation?

Note. R=receptive; P=productive. (Originally adapted from Nation & Webb, 2011, p. 190.)

As for mechanisms behind L2 collocation learning, under the usage-based model (Bybee, 2007; Ellis, 2002), the prevalent notion among L2 vocabulary researchers is that most L2 vocabulary is acquired from input, mainly reading input. Thus, knowledge of multiword units such as collocations is also largely deemed to be acquired gradually through repeated exposure to the target language (Ellis, 2003; Granena & Long, 2013). Recently, however, it has been argued that such naturalistic usage-based learning may not be sufficient for developing L2 vocabulary (Cobb & Horst, 2004; Ellis, 2008; Laufer, 2005), let alone collocational knowledge (Boers & Lindstromberg, 2012; Webb & Kagimoto, 2009; Webb et al., 2013). Boers, Dang, and Strong (2017), for example, reported that in many foreign-language-learning contexts where the amount of exposure to the target language is typically very limited, learners tend to be slow at acquiring knowledge of multiword lexis (e.g. Laufer & Waldman, 2011; Li & Schmitt, 2010). Even in the immersion learning context, the frequency of encountering specific collocations, which mostly appear far less frequently than single words (Martinez & Schmitt, 2012; Shin & Nation, 2007), may not be sufficient to facilitate incidental vocabulary learning. Furthermore, any knowledge gained through one encounter may be forgotten before the next encounter (Durrant & Schmitt, 2010; Eyckmans, Boers, & Stengers, 2007), especially if collocations are less semantically transparent. Thus, the potential for learning most collocations incidentally may actually be relatively small (Webb et al., 2013).

In addition to the lack of sufficient number of exposures, another challenge concerning collocation learning often involves remembering partnerships of words

that are often already familiar. This can lead learners to pay even less attention to them than to attention-catching individual words that are unfamiliar (Long, 2015). In addition, since there is often no logical or semantic reason for why certain lexical patternings are used over other possibilities (e.g. Cruse, 1986; Pawley & Syder, 1983; Swan, 1997), it can be a bigger burden for learners to learn those combinations and then link them to meaning than just matching unfamiliar words with their underlying concepts. In particular, high-frequency words that lack semantic distinctiveness (e.g. *make* rather than *do* in *make an effort*) and/or that have near-synonymous competitors (e.g. *tell* rather than *say* in *tell lies*) are known to pose great challenges to learners, as it is more difficult to suppress inappropriate word choices when it comes to remembering collocations than when it comes to remembering distinct word–meaning mappings. In light of the great challenges involved in L2 collocation learning, the current study, thus, attempts to test specifically whether practice in multiword recognition can lead to developing automatized collocation knowledge of form-meaning mappings, that is readily accessible in comprehension. We now turn to the literature on automaticity in theories of skill development, beginning with key concepts of skill acquisition theory, followed by a discussion of automatic lexical recognition.

Skill acquisition theory

The most widely known general theory of skill acquisition is Anderson's adaptive control of thought (ACT) (Anderson, 1992, 2013) or ACT-R (Adaptive Control of Thought – Rational) in its most recent version (Anderson et al., 2004; Anderson & Lebiere, 1998). This theory distinguishes two different types of knowledge: declarative and procedural knowledge. Declarative knowledge is defined as factual knowledge (knowledge "that"), which is usually directly verbalizable. Procedural knowledge, on the other hand, is knowledge manifested in a person's behavior itself (knowledge "how") and thus is not directly verbalizable. The model assumes that skill acquisition involves transitions through three stages. That is, learning proceeds from the first stage where learners acquire declarative knowledge, through the second stage of proceduralization (qualitative change in knowledge from declarative to procedural as a result of rules applied repeatedly in a consistent manner), and then finally to the autonomous stage where a task is performed in an ever faster and more automatic fashion. One important characteristic of procedural knowledge worth noting is its directional asymmetry. This means that application of procedural knowledge is restricted to the specific operation (skill) that was practiced and cannot generalize to other uses, while declarative knowledge is open to any uses regardless of the directionality of its practice, thus generalizable. Thus, if one practices for word recognition only, the procedural knowledge developed through such practice cannot transfer to the other uses such as word production, which requires a different skill from word recognition (e.g., lexical selection, phonological

code retrieval, syllabification, etc.). That is, word production skill will require a different type of practice that involves "creating (learning) a new rule with a relevant starting condition and a corresponding action" (DeKeyser & Sokalski, 1996, p. 617).

With regards to SLA, many researchers have made reference to ACT in describing L2 development (e.g., De Jong & Perfetti, 2011; DeKeyser, 1997, 2007, 2015; Lyster & Sato, 2013; McLaughlin, 1989; Segalowitz et al, 1998). L2 learners, especially in a classroom setting, mostly learn declarative knowledge about lexis and L2 grammar (e.g., explicit knowledge of a vocabulary item) first and, through consistent and extensive practice, develop procedural knowledge of how to perform a certain task in L2, which allows them to process the target language faster and more accurately with less effort than in the previous stage, and finally become capable of performing a practiced task with a fairly high degree of automaticity, such as accessing and retrieving the correct meaning of a word quickly and automatically in reading. In sum, skill acquisition theory in SLA specially focuses on explaining how the representation of L2 knowledge develops with increased contact and experience with L2 and what roles different types of L2 knowledge at various stages of the learning process play in skilled use of L2.

Automaticity of lexical recognition in L2 reading

Word recognition involves mapping an orthographic form of a word in a text, or a phonological form in speech, with a lexical concept or representation (Balota, Yap, & Cortese, 2006; Koda, 1996). As mentioned, automaticity in word recognition is essential for successful communication and fluent language use (Hulstijn, 2001;

Schmitt, 2008). This is obvious in listening and speaking skills when one has to parse, decode, and understand the vocabulary in a speech stream that proceeds at a speed of two to three words per second (Segalowitz, 2000) and when one has to retrieve a word within 200 to 400 milliseconds from the mental lexicon for normal and fluent speech production (de Bot, 1992). This is also true in reading when normal fluent reading moves along at a speed of three to six words per second (Grabe, 2009). Carrell and Grabe (2002) pointed out that reading less than around 150 words per minute may result in a difficulty in preserving the coherence of the text. Specifically, Grabe (2009, p. 23) stated that "Fluent reading comprehension is not possible without rapid and automatic word recognition of a large vocabulary". Empirical studies demonstrating a significant association between automatic word recognition skill and fluent reading performance further support the critical role of automatic word recognition in L2 reading (Favreau & Segalowitz, 1983; Koda, 1996; Schoonen et al., 2003; N. Segalowitz & Hulstijn, 2005; Van Gelderen et al., 2004).

The L2 reading task involves other skill components too, such as word and sentence decoding, and metacognitive and strategic skills. It is widely believed in L1 and L2 reading studies that the efficiency of the lower-order components, such as lexical access, is a prerequisite for the higher-order ones (Ehri, 1991; Perfetti, 2007; Segalowitz, 2010). Developing fluency in low-level processes leads to freeing up the greater amount of processing resources, allowing them to be used for higher level text processes such as comprehension (Grabe, 2009; Harrington & Jiang, 2013). For example, in order for successful reading to take place, a reader needs to allocate

optimal cognitive resources in the higher level processing (e.g., consolidating extracted lexical information into larger meaningful sentences for comprehension), which involves more complex multiple activities, thus more working memory. It is therefore crucial for a reader to process low-level components with fluency and less cognitive effort. In sum, for successful vocabulary use in comprehension and production, having knowledge of vocabulary is not sufficient. The learners must develop the skill of automatically accessing those words. In the next section, characteristics of automaticity in lexical recognition and their measurement will be discussed.

Important criteria for automatic lexical access and their measurement

Different indications of automaticity have been used to define and examine automatic lexical processing such as fast, unconscious, effortless, and stable processing. Despite varying operationalization of automaticity in previous research, there seems to be general agreement that automatic processing is highly rapid and largely error-free. However, some researchers argue that fast processing, despite its prominent role, is not enough. Ehri (1991), for example, argued that the ultimate goal in word recognition development is associated with restructuring a word-recognition mechanism or with increasing cognitive efficiency in word-recognition processing. A similar point is made by Segalowitz and Segalowitz (1993), who distinguished between fast and automatic processing and proposed that two qualitatively different changes are involved in the development of L2 word-recognition fluency. Specifically, in the initial stage of the development of L2 word-recognition skills,

simple speedup of performance takes place; however, according to their view, simply being able to recognize words quickly cannot be a sufficient condition for automatic word recognition (e.g., S. J. Segalowitz, & N. Segalowitz, 1993; Segalowitz et al., 1998). Thus, automatization, according to their view, is associated with increased cognitive efficiency involving changes in the way underlying processes are organized or in the internal organization of a given process. When learners become more proficient, the processing of word-recognition skill, for instance, becomes less variable, as controlled and more effortful mechanisms drop out and/or replaced with more efficient processes (S. J. Segalowitz, & N. Segalowitz, 1993; S. J. Segalowitz, N. Segalowitz & Wood, 1998).

To assess whether an increase in L2 word-recognition fluency is a result of a qualitative change in the nature of underlying processes, Segalowitz and his associates proposed an innovative approach using relative variability of performance (as measured by coefficient of variation). The coefficient of variation (CV) is a standard deviation of reaction time (RT) divided by mean RT; CV reflects processing stability with individual's processing speed taken into account. According to this conceptualization, when one improves performance by simply speeding up its component processes, mean RT and standard deviation (SD) decrease proportionally, resulting in no change in CV. However, when one's performance improves because of increased efficiency in a processing mechanism, SD decreases more drastically than mean RT does thus, CV is reduced. Given a set of mean RT and CV pairs for a group of participants, the correlation between mean RT and CV should be positive.

This approach has a great practical potential as it allows a researcher to assess whether the improved performance is attributable to generalized speedup or something beyond that is provided by the speed (qualitative changes in component processes). This proposal of relative variability of performance as a valuable indicator of automatization has been supported in a series of empirical studies carried out by Segalowitz and his associates (Phillips, Segalowitz, O'Brien, & Yamasaki, 2004; Segalowitz, 2000; S. J. Segalowitz, & N. Segalowitz, 1993; Segalowitz et al., 1998). Moreover, in an increasing number of recent studies, the CV measure has been shown to be an effective index for assessing increasing automaticity in sentence processing as well as in word recognition (Akamatsu, 2008; Elgort, 2011; Harrington, 2006; Lim & Godfroid, 2015; Suzuki & Sunada, 2018). Hulstijn et al. (2009), however, made some cautioning remarks regarding using CV as an index for automatization based on some instances of failure to replicate CV effects and potential confounds between improvement of knowledge and improvement of processing efficiency. In a follow-up study using the same tasks and stimuli (with slight modification), however, Lim and Godfroid (2015) reported CV effects in support of S. J. Segalowitz and N. Segalowitz (1993). As was pointed out by Lim and Godfroid (2015), the lack of CV effects reported in Hulstijn et al. (2009) could be due to the fact that the high school learners in their studies may have not yet progressed to the automatization stage due to the lack of sufficient exposure and practice. This possibility is especially plausible considering that these beginning learners were learning English (the target language under investigation) as their second or even third languages (the majority of the

participants were bilinguals with Moroccan, Turkish, or Surinamese backgrounds) in an EFL context. In addition, the methodological concern (potential confounding of improvement of knowledge and improvement of processing efficiency in CV measurement) has often been addressed by limiting the use of CV to only highly familiar target items as in Hulstijn et al. (2009), and/or to those learners with sufficient level of proficiency as in Lim and Godfroid (2015). This limitation has also been addressed in training studies by inclusion of a study phase to ensure solid declarative knowledge before having learners engage in training for procedural learning and testing their performance, in the current study as well as others (Li, 2017; Suzuki, 2017).

Furthermore, several neuroimaging studies provide further support for a close connection between automatic skilled processes and processing efficiency (Fischler, 1998; Haier et al., 1992; Phillips et al., 2004; Raichle et al., 1994). Haier et al. (1992), for example, found that learning (e.g., improved performance) resulted in "decreased use of extraneous or inefficient brain areas" (p. 134), which lends support to the idea that measures of processing stability are relevant to the issue of skilled performance (Segalowitz, 2010). In a more recent study, examining the variability in N400 amplitudes, electrical brain activity measure of processing efficiency, as well as in a behavioral measure (RT), Phillips et al. (2004) obtained a similar pattern of results between the behavioral and N400 ERP CV measures from highly proficient and less proficient learners, which indicates that inter-individual processing efficiency can be observed with neurophysiological measures as well.

In addition to speed and response-time variability, automatic processing is also associated with its reduced cognitive capacity usage, which allows freeing up time and attentional resources for other activities (e.g., Segalowitz, 2003, 2010). This will ensure that a skill in question is performed efficiently with less interference from and with simultaneous tasks (DeKeyser, 1997; Logan & Etherton, 1994). For example, a fluent skilled L2 reader is able to recognize words without disruption and interference from other ongoing activities. In other words, as learners become capable of recognizing words automatically, fewer attentional resources are required for accessing the words and activating their corresponding concept in the mental lexicon (Akamatsu, 2008). In order to examine automatic processing with respect to attentional resource allocation, a dual-task paradigm has often been employed (e.g., DeKeyser, 1997). The logic of the dual-task method is that when learners perform a primary task of L2 skill under investigation, if that skill has been automatized, the performance on the primary task should be less affected by performing the secondary task simultaneously (e.g., Logan, 1979). However, as was found in piloting and in DeKeyser (1997), the identification of a secondary task with the adequate level of difficulty for the purpose of the study is not easy.

Using a somewhat different paradigm, a sentence and word recall, Tremblay, Derwing, Libben, and Westbury (2011) reported two experiments looking at cognitive capacity requirements of lexical bundles (high frequency multiword units that are semantically and syntactically incomplete such as *that there is no*) in auditory and visual modalities. The participants were presented with sentences that included

either target lexical bundles or their controls and a set of six unrelated words. Their task was to recall the sentences and the additional words that followed them. The authors predicted that if lexical bundles are recognized and processed fluently (i.e., processed as holistic chunks), sentences that contained them such as *I sat in the middle of the bullet train* should require less cognitive resources than equivalent sentences that do not such as *I sat in the front of the bullet train*, thus recall should be easier and more accurate for the former and more additional words should be recalled after the former sentence. The results showed that native speakers recalled more sentences containing lexical bundles in both modalities, but only in the visual modality were more subsequent words recalled in the lexical bundle condition than in the non-lexical bundle condition. The authors concluded that the overall pattern of results supported their assumption.

The same logic and task can be used to test whether the target collocations are fluently accessed and recognized. If the target collocations are automatically recognized, sentences that contain them should take up less attentional resources than equivalent sentences that do not, allowing more attentional resources to be used for an additional task (e.g., processes of integration of semantic information with structural aspects of the sentence, mnemonic processes, etc.). Thus, more sentences containing the trained items and more additional words following them should be recalled accurately than those associated with their control counterparts[1]. The following

[1] It is important to note that immediate sentence recall does not merely reflect output from short-term surface memory but involves a process of regenerating a sentence from its conceptual representation using the normal mechanisms of sentence production that is no different from that of long-term sentence recall (Lee & Williams,1997; Potter & Lombardi, 1990).

section reports the major findings from the related empirical studies of automatic lexical recognition.

Empirical studies of automatic word recognition in SLA

A number of empirical studies have investigated automatization of word recognition with various types of training. In the second experiment by Segalowitz and Segalowitz (1993), for instance, 66 learners of English with varied levels of proficiency (low to advanced) were tested on 160 English words (35 baseline words, 35 homophone words, and 15 words appearing six times at random) and 124 nonwords in a lexical decision task. Based on the participants' response latencies to the baseline words, participants were first divided into two groups: skilled (fast and fluent) and less skilled (slow and less fluent) learners. The performance for the first and last (sixth) presentations of the 15 target words were analyzed separately for the skilled and less skilled learners. The results show that CV correlated significantly with mean RT at both the first and last presentations for the fluent (fast) learners while for the less fluent learners, CV only showed a significant correlation with mean RT for the last presentation. Authors interpreted significant positive correlations between CV and mean RT for the fluent speakers and less fluent speakers after encountering the target words six times as evidence for automatic word processing. However, although highly insightful, the results should be interpreted with caution because the study does not report whether there was a significant decrease in CV from the first to the last responses of learners. Information about response accuracy was also not included.

Extending the earlier study into a longitudinal setting, Segalowitz et al. (1998) investigated the role of practice in the development of automaticity in L2 word recognition for the learners of French across one academic year. The results showed that the residual RT (final scores with initial scores partialled out) and the residual CV were strongly correlated for both fluent and less fluent groups, suggesting that both groups gained fluency in word recognition through repeated encounters and practice over two semesters.

Using the same approach as proposed by Segalowitz and Segalowitz (1993), Fukkink et al. (2005) examined the effects of training on automatic word-recognition processing and on L2 reading comprehension. The computer-based training consisted of two sessions (each lasting 40 minutes) in which participants completed translation and cloze exercises. During the two training sessions, each learner encountered the target words (40 items) 11 times in total. The results of a lexical decision task before and after the training showed a significant facilitation in mean RT and a small reduction in CV. There were significant correlations between mean RT and CV for the trained words in both pre- and posttests. In line with the previous studies (Segalowitz & Segalowitz, 1993; Segalowitz et al., 1998), the results were taken as evidence for improvement in word-recognition fluency in terms of stable processing as well as speedup. In the second experiment with a between-subject design in a classroom setting, 105 Dutch learners of English were tested for word recognition and reading comprehension. The four training sessions consisted of four different types of exercises focusing on mapping word forms to their meaning of familiar and

unfamiliar words. Significant training effects on speed of lexical access were found for both familiar and unfamiliar words. However, there was only partial evidence for processing stability, with a reduced CV and a positive correlation between mean RT and CV found for only a part of the trained items. With regards to the effects of the training for automaticity on reading performance, contrary to the authors' expectation, the study failed to find any reliable relationship between improved lexical access skills, and reading speed or reading comprehension.

Using a different type of exercise, Akamatsu (2008) also examined the effects of training on the word-recognition performance, and whether the training effects were moderated by target word frequency and learners' working memory capacity. Forty-seven EFL students were given seven weekly training sessions lasting 90 seconds each, in which they were asked to draw separator lines between 150 English words (including 25 high-frequency and 25 low-frequency target words that were familiar to the participants) written with no spaces as quickly and accurately as possible. The results of a lexical decision task before and after training showed that the high-frequency target words were associated with simple speedup (accompanying a significant decrease in mean RT only) while low-frequency words were associated with stable processing as well as speedup (with significant decrease in mean RT and CV as well as a significant correlation between them). Working memory was not found to play a role in improving processing speed nor processing stability. The author commented that the lack of evidence for the association between the high-frequency target words and stable processing may be due to the fact that the

participants had already passed the automatization stage for the high-frequency words before the training. In other words, the EFL learners had been exposed and used the high-frequency target words so extensively that they had recognized them in a fully automatic manner even before the training.

Note that the studies reviewed above are all concerned with the automatization in lexical access to single-word items that were already familiar to the participants. As mentioned earlier, this has been common practice in order to separate the development of knowledge with improvements in processing (automaticity) (Hulstijn et al., 2009) and is consistent with skill acquisition theories that presume existence of knowledge prior to proceduralization. However, although theoretically and practically valid, this restriction may result in including target items that may have been already automatized as was the case for high-frequency words in Akamatsu (2008). As the restriction relies on familiarity of the items to the participants, it cannot rule out the potential effects of differing prior experiences and practice that participants may have had with the target items before participating in the experiment.

To the best of my knowledge, there is no study exploring automatic lexical access of newly learned words or multiword units, with the exception of Elgort (2011) and Solovyeva and DeKeyser (2018). In the second and third experiments of Elgort (2011), for example, effects of deliberate learning (e.g., learning from word cards) on the automatic recognition of 48 pseudowords were examined using the CV measure. Data from the masked repetition priming and the automatic semantic

priming in each experiment were analyzed as indicators of improved lexical access, and it was found not only that those pseudowords produced priming effects comparable to the real L2 words, but also that the CVs of participants' responses to the learned pseudowords were significantly lower than those for nonwords (to which they had not been exposed) and even low-frequency L2 words in both experiments, indicating that processing the learned pseudowords was significantly less variable than that of nonwords and even the low-frequency L2 words after the training. The author concluded that participants processed the pseudowords with some degree of automaticity after the training. Although informative, the study does not report information regarding correlation between mean RT and CV; thus, it is difficult to determine whether the pseudowords were recognized in a fully automatic manner as defined by Segalowitz and Segalowitz (1993).

Extending the CV account originally proposed by Segalowitz and Segalowitz (1993), Solovyeva and DeKeyser (2018) specifically tested the hypothesis that increasing variability (increase in CV) is associated with establishment of mental representation using novel words under both intentional and incidental learning conditions. The learning of both form-meaning mapping and orthographic forms was assessed using a primed lexical decision task in a single experiment with two different types of stimuli (40 Swahili words and 20 nonwords, respectively). Consistent with their hypothesis, the results showed a significant increase in CVs accompanied by a significant decrease in RTs comparing the pre-and posttraining tests. In addition, they reanalyzed data from two previous studies on novel word

learning (albeit without significance testing) and reported a similar pattern of increased variability over time. Although the study sheds important new light on the interpretation of change in CV at the initial stage of vocabulary learning, the study does not report information regarding overall accuracy rate of the performance, which is crucial for the validity of the results.

As mentioned earlier, there is no systematic research on automatization of multiword-unit processing; a few studies merit attention, however. Wood (2009) described a case study with one ESL learner who received intensive training (nine hours in total over six weeks) including a variety of speaking activities (e.g., 4/3/2 activity; telling a story under increasing time pressure, 4, 3, and 2 minutes) in which formulaic expressions, such as *in the daytime* and *stayed up late,* repeatedly appeared. Throughout the activities, the participant was encouraged to pay close attention to formulaic expressions in the native speaker models provided as input and use them in production. Based on the analysis of learner's spontaneous speech samples before and after the training, the author reported that there was a substantial fluency gain in terms of speech rate, length of runs between pauses, and the use of formulaic expressions. The author concluded that consistent repetition and practice of formulaic expressions throughout the training led to increased facility with those expressions, allowing the learner to extend runs of speech between pauses. However, as the author did not systematically track the development of certain linguistic features in focus (although attempting to focus on the use of formulaic expressions), it remains unclear what exactly had been automatized through the training.

In another study, Sonbul and Schmitt (2013), although the focus of their study was limited to investigating the learning of word combinations at the surface level, compared the effects of three learning conditions on the explicit and automatic form recognition of collocations. The three learning conditions included: the enriched condition (i.e. reading a text with target collocations embedded three times); the enhanced condition (i.e. reading a text with embedded target collocations made salient with red, bold font); and the decontextualized condition (i.e. presenting target collocations individually with no meaning). Results of immediate and delayed post-tests (i.e. form recall and recognition tasks, and priming task) suggested that all conditions led to significant gains in explicit form recall and recognition both for native and nonnative speakers, while no condition promoted automatic form recognition (collocational priming), neither for native nor for nonnative speakers. The authors concluded that no durable automatized memory traces of collocations were developed through the three types of treatment. Considering the short treatment period with a few number of encounters (only one session lasting 20 minutes with three encounters of the target item), however, the treatments might not have provided optimal conditions to promote automaticity in form recognition. It is worth noting, however, that significant collocational priming (i.e. the tendency for an activated word to accelerate subsequent recognition of a collocate) amongst native as well as nonnative speakers, which was not found in Sonbul and Schmitt (2013), has been reported in a number of psycholinguistic studies (e.g., Durrant & Doherty, 2010; Durrant & Schmitt, 2010; Ellis & Frey, 2009; Ellis, Simpson-Vlach, & Maynard,

2008; Wolter & Gyllstad, 2011). Wolter and Gyllstad (2011), for example, tested L1-Swedish learners of English and native speakers on their recognition of congruent and incongruent collocations, using a primed lexical decision task. The results showed significant priming effects for both collocational conditions for the native speakers with negligible differences between the native speakers' RTs in the two collocational conditions, while for nonnative speakers the same effect was found for the congruent condition but with only partial evidence for the incongruent condition (significant effect was found only in the subject analysis). The authors commented on the differences observed amongst the nonnative speakers, arguing that there were inconsistencies in priming amongst the items in the incongruent condition such that some items demonstrated priming while others did not. This speculation was further supported by the nonnative speakers' offline receptive test results, which showed that many incongruent items were not recognized as legitimate collocations.

Overall, the review of the limited number of studies above suggests usefulness of consistent repetitions in a meaningful context for accurate and fluent use of multiword expressions. Although informative, the scope of the studies was very limited (i.e., one participant or remembering presence of constituent words in collocations) and did not systematically focus on processes of automatization of multiword processing. Thus, it remains necessary to empirically study automaticity of lexical access at the multiword level. The following sections will discuss potential factors that can affect collocation learning in detail.

Types of repetition and collocation learning

Given that automatization is considered a slow process that requires repeated practice with the same items, repetition should be one of the main requirements for fostering automaticity. Gatbonton and Segalowitz (2005) specifically argued that automaticity can be promoted by inherently repetitive tasks. The effect of repetitive tasks on development of fluency has been empirically shown by a number of studies (e.g., Arevart & Nation, 1991; Boers, 2014; De Jong & Perfetti, 2011; Nation, 1989; Thai & Boers, 2016; Wood, 2009). Adopting the 4/3/2 activity, De Jong and Perfetti (2011), for example, showed that repeating the same story three times, each time at a faster rate, is more effective for proceduralization of linguistic knowledge than speaking about different topics each time. Both conditions led to improvement in fluency during the training, but the improvement maintained on posttests (over 1 to 4 weeks) and transferred to new topics only by the students who repeated their speeches. The authors claimed that this long-term effect can only be attributable to changes in the students' underlying knowledge and processing as a result of repeated practice. Using a similar activity, Boers (2014) and Thai and Boers (2016) examined differential effects of repeating the same talk three times, with or without increasing time pressure. They found that both conditions yielded increase in fluency (as measured by participants' speech rate), with the shrinking-time condition resulting in significantly greater enhancement than the constant-time condition. The authors in the three related studies all suggested that a repetitive activity such as the 4/3/2 activity supports fluency by freeing up attentional resources (for a similar argument see

Arevart & Nation, 1991; Nation, 1989; Wood, 2001). To be more specific, they argued that when students know what to say in the repeated task, they need less time in deciding the content and how to say it and have more resources for retrieving linguistic knowledge, reducing the need for planning and pauses and thus allowing for a more fluent delivery.

Repeated encounters have also been regarded essential for collocation learning. Ellis (2001), for example, argued that adult L2 learners, just like child L1 learners, have a capability to perceive co-occurrences among words without conscious attention. The frequent co-occurrences of words in linguistic input will drive these combinations to be associated in long-term memory and consolidated into chunks. Therefore, the more of these co-occurrences are encountered, the more strongly associated and entrenched chunks are in learners' memory, thereby leading learners to process them as a single entity with greater fluency. Similarly, Long (2015) argued that repeated exposure to the same input has positive effects not only on comprehension but also on acquisition. Long further argued that collocations can be effectively learned by exercises that involve verbatim repetition, which had been found to be successful in a community ESL program for migrant workers in Maryland. In this program, learners listened to a model passage read by a teacher a couple of times and repeated it verbatim until they could recite the whole passage from memory and then they were given a written version of the same passage to read it aloud a couple of times to associate spoken forms with the written forms. When asked to talk or write about a new topic unrelated to the contexts of the passage they

had been exposed to, the learners started to use native-like collocations and larger constructions borrowed from the model passage in their speech and writing. Overall, repetitive tasks can be considered as an effective instructional technique for both automatization and collocation learning.

When creating repetitive exercises as suggested by Gatbonton and Segalowitz (1988) for fostering automaticity in multiword recognition, at least two different types of repetition can be considered: single-sentence context and varied-sentence context. Although a handful of studies have looked at the differential effects of various repetition conditions on oral fluency as reviewed above, no study has directly examined the role of different types of repetition condition in training on automatization of collocational knowledge. Of relevance, however, Durrant and Schmitt (2010) specifically examined the effects of two different types of repetition (verbatim vs. varied) on incidental collocation learning. The participants were randomly assigned to one of three different training conditions in which they were presented with a series of decontextualized sentences containing target collocations. In the single exposure condition, participants were exposed to each target collocation in a sentence, one time only. In the verbatim repetition condition, participants were presented with the same target sentences containing the target collocations twice, the first for seven seconds and the second time shown only for three seconds while in the varied repetition condition, two different sentences were presented for each target collocation for seven seconds each. The authors hypothesized that when learners are first exposed to a certain language input, a number of cognitive demands (e.g.,

recognizing the words, decoding the syntax, generating the appropriate semantic context and so on) that learners have to deal with may prevent any learning. However, when exposed to the same language input a second time, those issues can be at least partially resolved, and thus learners can focus more on consolidating and building fluency with the language. In an unannounced naming test that immediately followed the training, participants were asked to say the correct noun that follows the given adjective from each target adjective-noun pair (e.g., *warm*). The first two letters of the noun collocate were also provided in the test (e.g., *FL_ _*; for the noun *flat*). The results showed that both repetition groups outperformed the single exposure group, with the verbatim repetition group performing better than the varied repetition group. Based on their findings, the authors suggested that when teaching collocations, exercises including such verbatim repetition be given more attention, which enable learners to build up fluency with particular strings of language.

On the contrary, some researchers have argued that more demanding conditions of learning that require more effort during practice will benefit learning and retention over time (e.g., Bjork & Kroll, 2015; Brown, Roediger, & McDaniel, 2014; Schmidt & Bjork, 1992; Suzuki, Nakata, & DeKeyser, 2019). According to this proposal, the varied repetition condition, which is more cognitively challenging, is likely to require more effort during practice. Therefore, this condition may eventually lead to greater automaticity in processing the practiced lexical items than the verbatim repetition condition despite worse performance at the initial stage of learning.

In sum, considering the relevant findings in the previous studies (e.g., Boers, 2014; De Jong & Perfetti, 2011; Durrant & Schmitt, 2010; Schmidt & Bjork, 1992; Thai and Boers, 2016), it is reasonable to hypothesize that different types of repetition condition will have different effects on development of automaticity in multiword recognition. Thus, the current study aims to investigate the effects of two different types of repetition condition during training (single-sentence context vs. varied-sentence context) on multiword-recognition performance.

The relationship between L1 congruency and collocation learning

Besides contextual factors such as types of repetition, other linguistic factors such as L1 congruency can also affect collocation learning. Congruency is determined based on word-for-word overlap between L1 and L2 form and meaning (Bahns, 1993). Thus, congruent collocations are those having a direct translational equivalent which also sounds natural in L1 (considering, however, general syntactic rules of the two languages). For example, English and Korean have the same collocation *do homework*. English collocation *do homework* is translated into the Korean sequence, *sug-je-leul ha-da* which makes a natural collocation in Korean. Incongruent collocations, on the other hand, do not have an equivalent L1 construction and expressed using a different word or words in L1. An example of this is *make an effort* where the prototypical translation equivalent of the verb *make* corresponds to the Korean verb *man-deul-da.* This verb cannot, however, be used in an idiomatic translation of *make an effort* into Korean, as its word-for-word translation like * *no-*

ryeok-ul man-deul-da will be an improper and unnatural combination in Korean. Instead, to express the same meaning of the English collocation, the verb *ha-da* (do) should be used as in *no-ryeok-ul ha-da* (literally, *do effort*).

L1 congruency has frequently reported to have a significant influence on learning and processing collocations (e.g., Bahns, 1993; Bulut & Çelik-Yazici, 2004; Liontas, 2001; Nesselhauf, 2003, 2005; Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010). Bulut and Çelik-Yazici (2004) and Liontas (2001), for example, showed that L2 learners utilize multiple cues to understand multiword expressions. These studies looked at advanced learners of English with L1 Turkish and L1 Greek, respectively, and found that L1 knowledge among other cues (e.g., consideration of literal and figurative meanings, and guessing from context) was consistently used to identify and comprehend L2 multiword expressions including collocations. Interestingly, Nesselhauf (2003, 2005) found that congruence (having the same form and meaning in both languages) was the most important factor in L2 collocation difficulty. Based on an analysis of advanced L2 learners' essays, Nesselhauf (2003) reported that the learners made the greatest percentage of errors for noncongruent collocations. Similarly, based on their empirical findings, Wolter and Gyllstad (2011) reported that highly proficient nonnative speakers living in an immersion context showed their performance on incongruent collocations lagged behind that of congruent collocations. They found that the L2 learners both processed congruent collocations faster (with faster RTs on the primed lexical decision task) and scored higher on congruent collocations than

incongruent items in the receptive knowledge test. Similar facilitation effects have been consistently found on recognition speed for congruent collocations over incongruent collocations using a slightly different type of task (e.g., Wolter & Gyllstad, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010). Previous studies often drew on the model proposed by Jiang (2000; itself built on models first proposed by Levelt, 1989) to explain how L1 knowledge may be involved in learning and processing of L2 collocations (e.g., Wolter & Gyllstad, 2011; Wolter & Yamashita, 2015; Yamashita & Jiang, 2010). In Jiang's model of L2 lexical development, all lexical entries consist of a lexeme level and a lemma level. The lexeme level contains information about phonology, orthography, and morphology, while the lemma level includes semantic and syntactic information. Jiang's model suggests that in the first stage of learning a L2 word, a formal entry for a new word is created with little semantic, syntactic, and morphological information; thus, a L2 lexeme entry is created that is closely linked to its L1 translation. With increasing experience in L2, the second stage involves establishing stronger links between L2 forms and the lemma information of the corresponding L1 word through repeated activation of both and copying information in L1 lemmas into its L2 entry. At this stage, the use of L2 words is mediated by the L1 lemmas however, activation of L1 translation in L2 use may decrease. A final stage involves gradual integrating of semantic, syntactic, and morphological specifications into the L2 lexical entry as a result of extended exposure to the L2. At this stage, L2 lexicon is directly linked to the concept such that the use of L2 words no longer relies on its L1 translations.

Based on this model, Jiang and Yamashita (2010) further proposed that learning a L2 collocation follows three steps. The first step consists of recognizing a word string as a legitimate collocation in L2, which generally takes place with understanding its meaning. The second step is the integration of the collocation in the long-term memory through repeated exposures and practice. In the first two steps, an L2 collocation is primarily connected to its L1 counterpart or both its L1 counterpart and corresponding concept (but more strongly to the former than the latter). The third step involves establishing a direct link between the new collocation and the concept that is strong enough to allow automatic access to it without activating L1 equivalents. Under this hypothesis, with the knowledge of L1 congruent collocations, acquisition of congruent collocations can be accelerated (Wolter & Gyllstad, 2013). However, because no L1 counterpart exists for an incongruent collocation, the learning process may be more complex and demanding as it lacks the support from L1 (Wolter & Gyllstad, 2011). Based on this hypothesis and previous studies, it may be possible that L1-L2 correspondence would play a facilitative role in acquisition of declarative knowledge of collocations and/or developing automaticity in collocation recognition. On the other hand, according to the desirable difficulty framework (e.g., Bjork & Kroll, 2015; Brown, Roediger, & McDaniel, 2014; Suzuki, Nakata, & DeKeyser, 2019), the incongruent collocations, which may receive more attention and effort due to lack of L1 equivalents, could eventually show greater automaticity than the congruent collocations despite processing costs and slower rate of improvement at the initial stage of learning. To the best of my knowledge, these assumptions have

never been examined in a training study. Thus, the current study will help elucidate the role of L1-L2 correspondence in learning collocations.

The relationship between cognitive abilities and collocation learning

In addition to the significant role of L1 congruency, another factor that is possibly important in automatic collocation processing is individual differences in cognitive abilities. Developing complex cognitive skills such as language learning is inevitably associated with learner-internal cognitive capacities that will differ from one individual to another. In particular, there has been increasing attention to the link between working memory capacity and language learning and processing (e.g., Baddeley, 2003; Cowan, 2012; Wen, 2016). Working memory refers to a multicomponent cognitive system which is concerned with temporary storage, manipulation, and maintenance of information during online cognitive operations such as language comprehension, arithmetic calculation, reasoning and problem solving (e.g., Baddeley & Hitch, 1974). The current version of the multicomponent working memory model includes four components, (1) the phonological loop responsible for the temporary storage of speech-based information, (2) the visuospatial sketchpad responsible for handling visual information, (3) central executive concerned with the control of those information, and (4) the episodic buffer responsible for storing and integrating information coming from all sources (Baddeley, 2007).

Working memory is regarded as an essential component of aptitude that affects most of the L2 cognitive processes (e.g., pattern recognition, integrating incoming materials) that underlie all three 'macro- stages' of SLA; input processing, central processing and output (Skehan, 2016). Based on a meta-analysis of 79 L2 studies, Linck, Osthus, Koeth, and Bunting (2014), for example, reported that working memory affects various aspects of L2 processing and proficiency development. Specifically, their results indicated that both the executive control component and the storage component of working memory are similarly important to L2 processing (but with a possibility that the executive control component is more crucial) while for L2 proficiency outcomes, the executive control component is more important. The robust relationship between working memory (i.e., executive control) and learning (gains) longitudinally has been reported by studies in both laboratory context (e.g., Kempe, Brooks, & Kharkhurin, 2010; Martin & Ellis, 2012; Misyak & Christiansen, 2012) and naturalistic classroom context (e.g., Linck & Weiss, 2015).

In particular, it has been pointed out that having smaller working memory capacity tends to cause "deficits in comprehension, particularly in the processes that integrate successively encountered words, phrases, and sentences into a coherent representation" (Daneman & Merikle, 1996, p. 423). Based on this notion, Akamatsu (2008) hypothesized that individual differences in working memory capacity could lead to learners in sight-word reading (processing a whole word as a unit), which is associated with restructuring a word-recognition mechanism or with increasing cognitive efficiency in word-recognition processing (Ehri, 1991). In other words,

through exposure to printed words, learners become capable of processing familiar sequences of letters or words as meaningful units or chunks in lexical recognition and utilize such units automatically. Thus, when exposed to some words or phrases equally, having greater working memory capacity is beneficial to processing them as holistic units. Under this hypothesis, Akamatsu (2008) specifically investigated the relationship between working memory (as measured by word and reading span tasks) and developing automaticity in single-word recognition. Contrary to the author's expectation, however, no significant role of working memory was found in improving participants' word-recognition speed or efficiency. The author pointed out that the unexpected results might have been due to the relative lack of cognitive demand in the exercise (word chain task) used for the training, which also may not have been effective practice to promote form-meaning mappings. This may have been the case, but it may also be possible that the effect of working memory was not observed as it may have played a lesser role as the automatization process had started (e.g., Serafini & Sanz, 2016). Serafini and Sanz (2016) reported that working memory (as measured by OSPAN) played a facilitative role in development of grammatical knowledge for less proficient learners but far less of a role for more proficient learners, suggesting decreasing impact of working memory in L2 learning with increasing L2 proficiency. Thus, further study is needed to identify the role of working memory in the development of skilled lexical recognition using different types of practice that promotes form-meaning mappings.

In addition to working memory, a different component of memory systems, such as procedural memory, may also be closely related to the mechanisms contributing to improvement in the speed and efficiency of lexical recognition. Procedural memory which underlies cognitive and motor skill learning and involves gradual learning of "how" (Knowlton & Moody, 2008; Squire, 2004), is relevant to capacity to automatize verbal material and may be engaged primarily during training for automatization. Procedural memory involves "the ability to gradually extract the common elements from a series of separate events" (p. 174) and is "expressed through performance rather than recollection" (Squire, 2004, p.173). Knowlton and Moody (2008) also argued that learning which relies on the procedural memory system proceeds gradually over time through repetition and extensive practice and is likely to require fewer attentional resources than declarative learning. Although the focus of the studies was limited to grammar rules, procedural learning ability as measured by a serial reaction time (SRT) task or Tower of London task has been repeatedly found to be significantly related to L2 learners' automatic use of language knowledge (Granena, 2013; Suzuki, 2017). Suzuki (2017) specifically looked at the relationship of procedural learning ability with automatization of morphological rules (as measured by RT and CV) and found a significant association for RT but failed to find a reliable association for CV. Note that the training proceeded in an explicit learning condition, where grammatical rules were explained and practiced deliberately. On the basis of the results, he concluded that procedural learning ability appears to contribute to speedup process, but not to more stable processing or

restructuring. In another training study on Chinese tone learning, Li (2017) failed to find any significant relationship between procedural learning ability (as measured by a probabilistic SRT task) and automatized learning outcomes (accuracy and RT) in both immediate and delayed post-tests. As for the conflicting results, Li explained that it may be "possible that most participants have reached the incipient stage of automatization, and the knowledge they have developed was automatized explicit knowledge (Suzuki & DeKeyser, 2017)", thus not showing any reliable relationship (Li, 2017, p.240). Although highly informative of a potential role of procedural learning ability in developing automaticity, only one study found a significant relationship between procedural learning ability and automatization (of grammar rules), and only one of the three RT measures associated with the memory task (movement execution time; the time from the first movement to the end of the last movement that participants make during the task) showed a reliable relationship with participants' performance (as measured by RT). Therefore, further study is needed to see if such results can be replicated for automatic lexical recognition using a different type of memory task.

Chapter 3: Research Questions and Hypotheses

The present study investigated the development of automaticity in collocation recognition and specifically focused on four potential factors that can affect the development of collocation knowledge: L1 congruency, training condition, working memory capacity and procedural learning ability. Two outcome tasks (phrasal decision task and sentence-and-word recall task) were used to measure learners' collocation-recognition skill. In the phrasal decision task, participants were asked to decide whether or not the word combinations presented were acceptable in English by pressing a specified key on the keyboard as quickly as possible when they saw it. In the sentence-and-word recall task, participants were presented with a sentence that included either a trained collocation or its control sequence and a set of six unrelated words one at a time in each trial block. Their task was to recall the sentence and the additional words that followed them. Specifically, RT and CV measures in the phrasal decision task, and sentence and word recall accuracy in the sentence-and-word recall task were examined to assess automaticity in collocation processing. Using these measures, the current study attempted to answer the following research questions:

> RQ1. Does the development of multiword-recognition processing follow what skill acquisition theory predicts? Specifically, does practice in multiword recognition lead to a decrease in CV and statistically positive correlations between CV and RT in the phrasal decision task, and reduced

cognitive capacity usage as reflected in sentence and word recall rates in the sentence-and-word recall task?

Hypothesis1-a: Based on Segalowitz and Segalowitzs' proposal (1993), training will lead to significantly lower RTs, and lower CVs, with significant positive correlations with mean RTs, for the target items in the phrasal decision task. Specifically, in the earliest stages of learning where no elimination and/or restructuring of underlying component processes are expected, no significant decrease in CV with positive correlation between CV and RT is expected. However, when the elimination and/or restructuring of underlying component processes takes place and, as a result, learners become increasingly automatic in collocation recognition, significant decrease in CV with a significant positive correlation between CV and RT is expected to become observable.

Hypothesis1-b: Training will lead to better recall performance in terms of sentence and word recall rates in the critical condition (sentences containing the trained items and supplementary words that followed them) than its control condition (those associated with control items), reflecting reduced cognitive capacity usage associated with automatic recognition of learned collocations. A stronger effect of training may be detected in the word recall performance than in the sentence recall performance as sentence recall may involve not only recognition skill but also production skill (which is not the focus of the training in the current study), and thus the recall rate may depend

on skills of both recognition and production of target items; however, the effect of the production component is expected to be negligible due to the fact that the constituent words of the target items are all highly familiar words, and in the scoring process minor spelling mistakes of the target items will be ignored.

RQ2. What is the role of L1 congruency in the development of automaticity in multiword recognition?

Hypothesis 2: Based on empirical studies (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010), it is possible that L1 congruency may play a facilitative role in the development of automaticity in collocation recognition, such that the congruent collocations benefit less from the training than the incongruent collocations. However, it is also possible that congruency may play a facilitative role only in the stage of acquiring declarative knowledge but not in the development of procedural knowledge. Alternatively, based on the logic of desirable difficulties in learning processes (e.g., Brown, Roediger, & McDaniel, 2014, Suzuki, Nakata, & DeKeyser, 2019), it can be possible that despite worse performance during the earlier stages of training, the incongruent collocations could eventually show greater automaticity than the congruent collocations by the end of the last session. This prediction is in line with the logic of desirable difficulties, which predicts that more challenging conditions of learning that require more

effort will benefit learning and retention over time. However, as there is no relevant study investigating automatization in processing of congruent lexical items, no strong hypothesis will be formulated.

RQ3. What is the role of different types of training condition (verbatim repetition vs. varied repetition) associated with exercises in the development of automaticity in multiword recognition?

Hypothesis 3: Based on previous research (De Jong & Perfetti, 2011; Durrant & Schmitt, 2010), it can be hypothesized that the verbatim repetition group would benefit more (i.e., faster and more stable recognition of the learned items, and higher recall rates of sentences and words in the critical condition) from the training than the varied repetition group. On the contrary, based on the desirable difficulty framework (e.g., Brown, Roediger, & McDaniel, 2014, Schmidt & Bjork, 1992; Suzuki, Nakata, & DeKeyser, 2019), it can be predicted that the varied repetition group (i.e., more demanding learning condition) may eventually show better performance than the verbatim repetition group by the end of the last session despite worse performance during the earlier training sessions.

RQ4. Is there a relationship between working memory capacity and automatization in multiword recognition?

Hypothesis 4: Based on previous literature (Akamatsu, 2008; Daneman & Merikle, 1996; Linck et al., 2014), it is hypothesized that participants with greater working memory capacity will have more advantages in benefiting from multiword-recognition training than those with less working memory capacity, as reflected in RT, CV, and sentence and word recall rates.

RQ5. Is there a relationship between procedural learning ability and automatization in multiword recognition?

Hypothesis 5: Based on theoretical as well as empirical findings (Knowlton & Moody, 2008; Squire, 2004; Suzuki, 2017), it is hypothesized that participants with greater procedural learning ability will have more advantages in benefiting from multiword-recognition training than those with less procedural learning ability, as reflected in RT, CV, and sentence and word recall rates.

Chapter 4: Methods

Design of the study

In order to address five research questions described in the previous section, the current study involves several outcome measures to assess participants' collocation-recognition skill (RT, CV, and sentence and word recall rates); two within-subject factors, time (Tests 1, 2, 3 and 4; only for RT and CV measures) and congruency (L1 congruent vs. L1 incongruent items); one between-subject factor, group (verbatim repetition vs. varied repetition); and several individual difference measures (OSPAN, SRT, and baseline RT) as covariates. Table 2 presents the schedule for the lexical recognition tests and training sessions.

Table 2. Schedule of tests and training sessions

	Day 1			Day 2		Day 3	
	Test	Training session	Test	Training session	Test	Training session	Test
Verbatim repetition group	Pre-test	1, 1, 1, 1	Post-test	1, 1, 1, 1	Post-test	1, 1, 1, 1	Post-test
Varied repetition group		1, 2, 3, 4		5, 6, 7, 8		9, 10, 11, 12	

Note. The numbers refer to the sets of training materials as listed in Appendix C. A different set (or combination of sets in the case of varied repetition group) was given to the same number of participants within each group (see procedure section for details); the table only shows an example of possible cases.

Participants were randomly assigned to one of the two experimental conditions (verbatim repetition and varied repetition) and engaged in three training sessions (each session per day) within a single week. The two groups were given the

same number of exercises (four rounds of exercises) in the same exercise format (fill-in-the-blank) in each training session but only differed in terms of training materials. That is, the verbatim repetition group was repeatedly presented with the same sentence context for each target item in exercises throughout all the training sessions, while the varied repetition group went through twelve different sentential contexts for each target item in the three training sessions together as shown in Table 2. For example, a participant in the verbatim repetition group completed the four rounds of fill-in-the-blank exercises with the training material of set one. A participant assigned to the varied repetition group, on the other hand, completed the four rounds of fill-in-the-blank exercises with four different sets of training materials in each training session (twelve different sets of training materials in total in the three training sessions together).

On day 1, participants were informed about the study and went through the declarative vocabulary study phase followed by the pretraining lexical recognition task (phrasal decision task 1), the training session (session 1), and the posttraining lexical recognition task (phrasal decision task 2). On day 2, upon completing the declarative vocabulary review, the training session (session 2) was administered followed by the posttraining lexical recognition task (phrasal decision task 3) and the individual difference measures (SRT and OSPAN). On day 3, upon completing the declarative vocabulary review, the last training session (session 3) was administered, followed by the two posttraining lexical recognition tasks (phrasal decision task 4,

and sentence-and-word recall task), the L2 lexical decision task and a background questionnaire.

Participants

Seventy-eight Korean learners of English who were either graduate or undergraduate students at universities in the United States were recruited for the study; however, six did not come back after the first session. Thus, a total of 72 participants provided data for the study. All participants were native speakers of Korean and had at least intermediate mastery of English (B or C levels), as measured by the DIALANG vocabulary size test (Council of Europe, 2001).

Participants in the two groups were comparable in terms of mean age, F (3, 64) =.454, p = .716, and gender distribution (Pearson Chi-Square X^2 = 3.524, p=.318). Participants in the two groups were also comparable in terms of DIALANG scores (U=589.5, z=-.80, p =.424), and length of residence in US, F (3, 64) = .454, p = .716. Table 3 summarizes the background information of the participants.

Table 3. Description of the participants

	Verbatim repetition group (N=36)	Varied repetition group (N=36)
Age	Mean = 24.64 Range =19-35	Mean = 24.56 Range =19-35
Gender	Female=17 Male=19	Female=24 Male=12
DIALANG scores	B1 (n=2) B2 (n=28) C1 (n=5) C2 (n=1)	B1 (n=4) B2(n=21) C1 (n=8) C2 (n=3)
Years of residence in US	Mean = 4.65 Range = 6.5-0.33	Mean = 4.47 Range = 6.67-0.67

Materials

Target items

The critical items for training consisted of 16 congruent and 16 incongruent verb+noun collocations in English that are not semantically transparent. Thus, the meaning of these collocations cannot be comprehended by merely combining the meaning of individual words. Semantically opaque collocations were chosen because they are regarded to be the most difficult type of multiword units to learn followed by idioms that often have both a figurative and literal meaning, and semantically transparent collocations (Grant & Bauer, 2004).

For the selection and construction of target lexical materials, a list of 65 potential incongruent verb+noun collocations were first identified using six resources[2]. The selected collocations were then trimmed down to 16 items (e.g., *draw a blank*) based on the following criteria: (a) they could not have direct translational equivalents in Korean, which were attested by non-existence of their translations in two Korean collocation dictionaries (Kim, 2007; Kim & Joo, 2016) and the Korean language corpus (a corpus of 36 million words)[3]; (b) their constituent words had to be

[2] A dictionary of English collocations: based on the Brown corpus (Kjellmer, 1994), English collocations in use: advanced (O'Dell & McCarthy, 2008), English idioms in use advanced with answers (O'Dell & McCarthy, 2010), Oxford dictionary of current idiomatic English: volume 1 and 2: phrase, clause & sentence idioms (Cowie, & McCaig, 1983), the BBI dictionary of English word combinations (Benson, M., Benson, E., & Ilson, 1986)

[3] A corpus of contemporary written Korean (a corpus of 36 million words) created by the National Institute of Korean Language (NIKL) was used. This corpus is the largest Korean corpus publicly available and represents a cross-section of written language from a wide range of late twentieth and early twenty-first century resources. For analyzing and retrieving the frequency data from the raw corpus, the software Hanmaru 2.0 developed by NIKL in 2011 was employed.

highly familiar to the participants (the majority of individual words were found in the lists of 3000 basic words for English education constructed by the National Curriculum Information Center in Korea in 2015; the pilot test results indicated that all the participants had receptive knowledge of all of the constituent words of collocations); (c) they needed to co-occur with relatively low frequency to ensure that the participants are unlikely to have any knowledge of them before the experiment (all well below 0.5 times per million occurrence according to the Corpus of Contemporary American English, COCA). Finally, the candidate collocations were piloted with 12 L2 learners similar to the experiment participants to ensure that they were all essentially unknown. In the norming test, the participants were asked to put a check mark only for those items that they were familiar with and asked to provide the appropriate meaning in L1. Only the items that were identified as unfamiliar to all the participants remained in the final list.

As for congruent collocations, because congruency is determined by the translation relationships between the languages, participants may easily comprehend the correct meaning of unknown congruent collocations even with very low frequency based on their knowledge of L1 collocations, especially when constituent words are already known. Thus, it is almost impossible to clearly identify congruent collocations that are potentially unknown to participants by means of frequency counts and pilot tests. To avoid such a dilemma, direct word-for-word translations of Korean-only collocations were used as the target congruent items (e.g., *roll money*). This method can ensure that both congruent and incongruent items are equally new

and unknown to the participants before the experiment. Thus, in both conditions, learning would involve the process of establishing membership of constituent words in collocations and links between target word combinations and their meaning.

For constructing congruent items, possible items were first selected using two Korean collocation dictionaries (Kim, 2007; Kim & Joo, 2016) and translated into English as closely as possible by the author using the core meaning of each word. The selected items were then checked to make sure they are not acceptable English expressions using the COCA (a corpus of 450 million words). Both length and individual word frequency of the items were matched as closely as possible between the lists of the target congruent and incongruent items to ensure that there are no major differences between the lists in terms of individual words. A summary of the lexical properties of these test items can be found in Table 4.

Table 4. Lexical properties of the stimuli (16 items in each of the two conditions)

	Congruent	Incongruent	*t*
Length	11.75	11.81	.10
Noun frequency	4.66	4.46	-.92
Verb frequency	4.51	4.55	.34

Note. Length refers to the total number of letters of each expression. Log adjusted frequencies are shown for noun and verb frequencies.
All comparisons are pairwise.

As Korean does not require overt marking of number (singular versus plural) or countability of nouns (count versus mass) (Ionin, 2003; Ionin, Ko, & Wexler, 2007; Lee Amuzie & Spinner, 2012), nouns in Korean N+V collocations typically appear in bare form, without an article or plural marking. English V+N collocations, of course, often include articles between two lexical components (e.g., *beat a retreat*,

cook the books). Thus, for comparative purposes, articles (definite and indefinite) were inserted in some of the constructed incongruent items to match the phrase structure and length between the two lists of items. The constructed items were then given to four native speakers of Korean with high proficiency in English to make sure of their congruent status. They were asked to write down as many appropriate Korean translations of a given item as they could think of such that resulting Korean translations sounded as natural expressions in Korean. Only those items that received the same translations by all four native speakers remained in the pool of possible items. It is worth nothing that, in a pilot study, all participants reported that they believed the constructed congruent items to be authentic English expressions during debriefing. Since the participants in the main experiments did not know that constructed items (Korean-only collocations) were used, learning those items can be considered comparable to the experience of learning actual expressions. The selected and constructed target items can be found in Appendix A.

Sentences used for the training

For constructing the training materials, twelve different sentences were created for each target item for two training conditions (verbatim repetition vs. varied repetition). All of the sentences were taken from the English or Korean dictionaries and corpora[4]. However, when necessary, the sentences were slightly modified to

[4] English dictionaries such as Cambridge, Longman, Collins, Merrian-Webster English dictionaries and the Free Dictionary by Farlex (Farlex, I. N. C., 2009) or the COCA, and Korean dictionaries such as standard Korean language dictionary, learner's dictionary of Korean collocation (Kim & Joo, 2016), Korean collocation dictionary for learners (Kim, 2007), dictionary of Korean idioms: idioms, proverbs, Sino-Korean phrases (Park & Eliot, 2013), essential Korean idioms: 300 idioms to upgrade your

ensure all the running words were potentially known by all of the participants. The pilot study indicated that participants had receptive knowledge of all the running words and no difficulty understanding the sentences. When selecting and constructing a series of different sentences for any given item, it was assured that the given item expressed the same semantic concept for all the sentences. The constructed sentences were divided into twelve sets, each set containing 32 different sentential contexts (one for each of the 32 target items). None of the target verbs or nouns in the target collocations was repeated in any sentence. The training materials are shown in Appendix C.

Training

The training for automaticity consisted of three training sessions on three separate days. Following Fukkink et al. (2005) and Snellings, Van Gelderen, and De Glopper (2004), a fill-in-the-blank exercise which focuses on access to meaning was given. This exercise has been used successfully in previous studies to foster automaticity in lexical recognition (Fukkink, Hulstijn, & Simis, 2005; Snellings, Van Gelderen, & De Glopper, 2004) and found to be the most common format used in students' course materials for learning collocations. Boers et al. (2014) specifically suggested that this format is likely to encourage learners to focus on the meaning of a collocation, and thus facilitate acquisition of the meaning of the expressions by

Korean (Lee & Lee, 2017), 500 common Korean idioms (Pyun, 2017), or a corpus of contemporary written and spoken Korean constructed by NIKL.

making learners evaluate which sentential context is most compatible with the meaning of the expression. In a follow-up study, Boers et al. (2017) indeed found that this format (matching intact expressions with sentential contexts) leads to the best retention of meaning compared to two other formats of cloze exercises they examined (choosing the missing verb of the collocations and completing the missing verb of the collocations by using a first-letter cue). The pilot study for the present research also indicated that the exercise was adequate for facilitating skilled recognition of the target expressions.

In this exercise, the participants were presented with four sentences each containing a gap, where the target collocation was missing. Beside each sentence, a list of the missing expressions appeared, one of which correctly fit each gap. The participants were asked to choose the appropriate expression that fits each sentential context under time pressure. Participants were always required to select the collocation as a whole to prevent them from establishing erroneous combinations in memory (Boers et al., 2014). An example is given below (see Figure 1). The training materials were first piloted with two native speakers to make sure that each sentential context was appropriate for the given target item and modified slightly, when necessary. Only those items that worked well for the purpose of the study remained in the final materials.

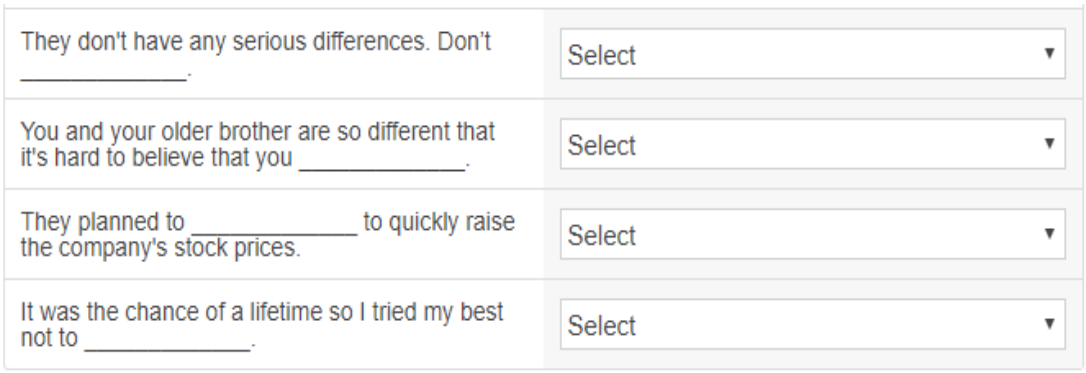


Figure 1. Example of a fill-in-the-blank exercise

For each training session, the participants practiced the entire set of target collocations (congruent and incongruent items), with either identical sentential context (verbatim repetition) or varied sentential context (varied repetition), in four rounds of exercises. Thus, in the three training sessions together, each target item of the training set was encountered 12 times (4 rounds x 3 sessions) in a meaningful context. In all four rounds, participants were asked to respond as quickly and accurately as possible and received feedback on every response regarding the correctness, as well as the correct answer in case of erroneous responses.

Tests

Two tests (phrasal decision task, and sentence-and-word recall task) were used to assess the participants' collocation-recognition performance.

Phrasal decision task

A phrasal decision task, modeled on the classic lexical decision task used commonly in psycholinguistic research including automaticity research (e.g.,

Akamatsu, 2008; Elgort, 2011; Harrington, 2006; Segalowitz & Segalowitz, 1993; Segalowitz et al., 1998) was administered before and after the training to assess the participants' collocation-recognition performance. This task has also been used widely in previous studies investigating the learning and processing of collocations and other types of multiword expressions in a L2 (e.g., Jiang & Nekrasova, 2007; Wolter & Gyllstad, 2013; Yamashita & Jiang, 2010). The test stimuli included 16 congruent and 16 incongruent collocations as well as 32 noncollocational filler items. Following previous studies (Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2015, 2018; Yamashita & Jiang, 2010), the filler items for the test consisted of atypical combinations of two high–frequency words such as *jump the smell* or *lay an age*. All high–frequency words were selected from the New General Service List (Browne, Culligan, & Phillips, 2013). The inclusion of these implausible items was for task requirements only, and these items were not analyzed, because their data are not relevant to the purpose of the study.

In this task, the target collocations and noncollocational control items were presented to the participants, one at a time. Similar to a lexical decision task, the participants were asked to decide whether or not the word combinations presented are acceptable ones in English by pressing a specified key (Yes/No) on the keyboard as quickly as possible when they see it. After the instruction and 12 practice items, the experimental items (i.e., 32 collocation items and 32 non-collocational control items) were presented in an individually randomized order. Both the practice and test items began with a series of asterisks in the center of the screen for 1000 ms for the purpose

of eye fixation. The asterisks were followed immediately by the presentation of an item, to which the participant responded. RTs were measured from the onset of an item to a key press.

L2 Lexical decision task

The current study examined the effects of two different types of training condition (between-subject variable) on the recognition of target collocations (as measured by RT). Participants, however, may have differential speeds of cognitive processing (one may be a fast reader while another may be a slow reader) and L2 proficiency. Therefore, in order to account for individual differences in speed of cognitive processing and L2 proficiency when analyzing RT data obtained from the phrasal decision task, baseline RT data for each participant was obtained using a lexical decision task (LDT). The stimulus words for the LDT consisted of 32 high-frequency English words and 32 pseudo-words (see Appendix B), which were originally developed by Matsuo (2017). The 32 real words were made up of 18 items originally selected from each of the first three 1000-word frequency bands in the BNC. Lexical properties of the stimulus words, such as number of letters and syllables, and part of speech were all controlled. Pseudo-words were randomly generated with the ARC Non-Word Database (Rastle et al., 2002) controlling for the number of the letters and syllables, and selecting orthographically existing onsets, bodies, and legal bigrams to discourage learners to make a decision based on spelling rather than lexical access (Jiang, 2013). The obtained RT data was used as a covariate

to account for individual differences in processing the target language when analyzing RT and CV data from the phrasal decision task.

Sentence-and-word recall task

Following Tremblay et al. (2011), a sentence-and-word recall task was used to assess fluency in collocation recognition. The critical stimuli for this task consisted of 32 sentence pairs that included either the trained items (congruent and incongruent items) or their controls (non-collocational phrases) and 192 high-frequency monomorphemic three-phoneme words taken from the BNC, which were presented after the sentences (e.g., *bone*, *song*, and *tree*). None of the selected three-phoneme words were repeated in the target collocations or training materials. The pilot study indicated that the participants had receptive knowledge of all the individual words. The control sequences were constructed by replacing one word in the target items with another word of similar length and frequency (e.g., *beat a retreat* vs. *beat a flight*). Efforts were made to ensure that the resulting control items formed semantically plausible phrases and were comparable to the target items in meaning. All the constructed control sequences had either zero or close to zero frequency counts as a whole sequence based on the COCA[5]. There were no significant differences across the four conditions (congruent items, control for congruent items, incongruent items, and control for incongruent items) in terms of length; F (3, 63) =

[5] Although, theoretically, these control sequences should not be found in COCA, given the tremendous size of the COCA, a few items did register a small number of occurrences (range=1-8 raw phrasal frequency). However, these items as well as the other control items were rated by the native speakers as atypical combinations in English in piloting.

.313, p = .816, noun frequency; F (3, 63) = .650, p = .586, and verb frequency; F (3, 63) = .140, p = .936.

In most cases, the target items (congruent and incongruent items) and their control counterparts were embedded after the second word of the sentence and followed by three more words. The mean length of sentences was 8.4 words for both congruent and incongruent sentence pairs. The critical and control sentence pairs were separated into two counterbalanced lists such that 16 target sentences and 16 control sentences are included in each list. Thus, if a critical sentence containing a target item appeared in List A, its control sentence with its control sequence appeared in List B, and vice versa (see Appendices D.1 and D.2 for a complete lists of stimuli).

In each trial, sentences containing a target collocation or a control sequence were presented visually one at a time. Each sentence was followed by a series of unrelated six mono-morphemic words of equal length and frequency of use presented one by one. These six words and their order of presentation remained constant across participants. Participants were asked to remember the sentence and as many single words that followed as possible and type in the sentences and as many words as they could remember after each trial block. Three practice trials consisting of non-study items were presented first to allow participants to get used to the task. Sentences appeared in the center of the screen for 8000 milliseconds and then disappeared, followed by each additional word individually shown for 3000 milliseconds (the display duration was determined based on pilot tests). The order of the trials was randomized across subjects to avoid potential order effects.

Scoring procedures for the sentence-and-word recall task

Two dependent variables were obtained from the sentence-and-word recall task: the number of sentences containing the trained collocations and their control sequences that were correctly recalled and the number of additional words recalled after the sentences. Recalled sentences were scored dichotomously, as either correct (1 point), or incorrect (0 point). Thus, a participant received a full score (1 point) for each sentence if the recalled sentence was either identical to the original or synonymous without noticeably changing the meaning of the original sentence as a whole. For example, minor errors such as minor misspellings, article mistakes, substitution of synonyms, errors of subject-verb agreement were ignored. If the recalled sentence was considerably different from the original with errors such as word(s) omissions and additions of novel words, it was scored as incorrect, and no partial points were given. As the focus of the current study is on development of collocation-recognition skill, rather than on production skill, minor misspellings of the trained items in the critical sentences such as accidental letter omissions (such as *squeez* for *squeeze*) were ignored.

For the additional words recalled after the sentences, 1 point was given to each correctly recalled word. Any words that were not part of the list of the additional words were rejected, and no partial points were given. As in scoring of sentences, minor misspellings such as accidental letter omissions (such as *til* for *till*) were ignored. The maximum possible score for each trial was 6, representing the total number of additional words presented in each trial. The researcher scored all the

responses while a second and a third rater scored 50 % of the data (i.e., 36 and the other 36 out of the 72 participants). The inter-rater reliability was high (Pearson correlations of .998 and .999, respectively).

Operation span task

Working memory was measured with a modified version of Turner and Engle (1989)'s operation span task (Stone & Towse, 2015), which is a complex span task that requires simultaneous storage of to-be-remembered items in memory and processing of simple mathematical operations. In this version of the OSPAN, participants were first presented with an integer that needs to be stored and recalled at the end of the trial in its correct serial position, followed by a mathematical operation (e.g., 6 + 7 = 10) and the participants were asked to indicate if they think the given answer is correct or not before the next to-be-remembered item is presented. The digit and operation were presented in sets ranging from two to six trials. At the completion of a given set, participants must recall as many of the two to six integers from that set as possible. Three sets of each set length were presented, for a total of 60 trials. The digits and operations were randomly generated for each trial. For each operation, there is a .5 probability that the operation will be correct and the type of operation (multiplication, division, addition, and subtraction) used has a .25 probability for each. Participants were additionally informed about the equal importance of equation-judgment and digit recalling accuracies to prevent them from only focusing on remembering the digits. Because the primary processing task involves solving math equations rather than processing language, this working memory measure is regarded

less dependent on language skills per se and thus minimizing any potential effects due to differences in language proficiency (cf. reading span task of Daneman & Carpenter, 1980). Participants were required to respond as fast and as accurately as possible. For scoring, a full-trial accuracy scoring method was used.

Probabilistic SRT task

A probabilistic SRT task, adopted from Kaufman et al. (2010), was used to measure participants' procedural learning ability. In the SRT task, participants saw a stimulus (an asterisk) appear at one of four locations on the computer screen and were asked to press the corresponding key indicating the location of the asterisk as fast and accurately as possible. Unknown to the participants, a sequence of 12 stimuli positions were generated based on a probabilistic order, following a repeating sequence intermixed with an alternate sequence 15% of the time. Two sequences: sequence A (1–2–1–4–3–2–4–1–3–4–2–3) and sequence B (3–2–3–4–1–2–4–3–1–4–2–1) differ in the second-order conditionals (Reed & Johnson, 1994) in which two consecutive locations lead to a different prediction (i.e., successor) for every sequence. In other words, the probability of occurrence after two consecutive cues (e.g., 1–2) is different between Sequence A and Sequence B. For example, the sequence 1-2, is always followed by 1 in Sequence A (1–2–1), whereas, on some trials (15% of the time), it is followed by 4 according to Sequence B (1–2–4). Transitions were generated from Sequence A 85% of the time (training condition) and from Sequence B 15% of the trials (control condition). Participants were not instructed to memorize the series or search for underlying rules and asked to complete

eight blocks, each block consisting of 120 trials, thus 960 trials in total. All trials were randomized within each block but presented in the same fixed order for each participant. Following Kaufman et al. (2010), the procedural learning ability score constituted the difference in the mean RTs between the training condition and the control condition across blocks 3 to 8.

Procedure

All the tests (except for the two memory tests) were computerized and conducted using DMDX (Forster & Forster, 2003) and all the training sessions were administered through an online testing site, the ClassMarker (http://www.classmarker.com). Each participant engaged in three training sessions (each session per day) in a quiet room within a single week. Table 6 presents the overview of experimental sessions. Upon completing the consent form, the first session (day 1) started with a vocabulary study phase, in which participants engaged in learning declarative knowledge of the 32 target collocations on the computer screen. Before the study phase, each participant was informed that the goal of this training is to learn idiomatic expressions in English and remember their meaning. They were encouraged to perform their best and told that they would be tested until they reached 100% accuracy. In the study phase, the participants went through three sessions of presentation of target items each followed by a short recall test. In each presentation session, each target collocation to be learned with its meaning in Korean was presented for 10 seconds on the screen. When introducing meaning, both the

incongruent and congruent items were presented with their corresponding definition in L1 similar to the definition that is found in a typical word entry in Korean-Korean dictionary in order to present the concept that each collocation is referring to in the clearest way. The set of all 32 target items was repeated twice; the first presentation session, however, was given in a self-paced trial, in which participants were asked to press the key to proceed to the next item to allow enough time to process and study each target item. The order of presentation was randomized. After every two cycles of all 32 target items, a short recall test with different target items each (16 items) was given to test explicit knowledge of target collocations; for the last session (session 3), however, a combined test of all 32 items was administered. During the recall test, each item was presented one at a time on the screen and participants were asked to say the correct meaning of the presented item out loud. The researcher monitored their performance and feedback was given by the researcher (correct/incorrect). The correct answer was also provided on the screen. Participants were tested until they demonstrate complete explicit knowledge of the target items. The vocabulary study phase was included in order to ensure that the participants had the declarative knowledge of the form and meaning of each target item before they automatize the lexical access to them.

Upon completing the study phase, the pretraining phrasal decision task (Test 1) was given. No feedback was provided during the test. The phrasal decision task was followed by the exercises of the training (for details, see training section) in four rounds. During the training session, participants were randomly assigned to one of the

two training conditions: verbatim repetition and varied repetition. In the verbatim repetition condition, participants were presented to the same sentence context in the fill-in-the-blank exercises repeatedly throughout all the training sessions (sessions 1, 2, and 3). That is, during each training session, participants were given the exercise with the same sentence context (e.g., set one) for each target item four times with a short break between rounds (when necessary). The order of the questions in the exercise of each round was randomized across the participants. Within this group, participants were also assigned in equal numbers to one of the twelve sets of training materials, to control for an effect of a particular training set, if any, on their performances.

In the varied repetition condition, participants went through twelve different sentences in the fill-in-the-blank exercise format for each target item in the three training sessions together. In each training session, one of four different sentential contexts were presented in the exercise in each round. The training began exactly the same way as in the verbatim repetition condition. On completion of the first round, participants were told to restart on their own time, and a different sentential context for each target item was presented for the second round. Note that the total number of opportunities to practice each target item was the same in both conditions (i.e., 12 exercises in total for three training sessions). Within this group, the twelve sets of sentences were administered following a 3X3 counterbalanced sequence (see Table 5) to control for order effects. The same number of participants were randomly assigned to one of the six counterbalanced orders of training sets.

Table 5. Counterbalanced sequences of training sets for the varied repetition condition

Order 1	1 2 3 4	5 6 7 8	9 10 11 12
Order 2	9 10 11 12	1 2 3 4	5 6 7 8
Order 3	5 6 7 8	9 10 11 12	1 2 3 4
Order 4	1 2 3 4	9 10 11 12	5 6 7 8
Order 5	5 6 7 8	1 2 3 4	9 10 11 12
Order 6	9 10 11 12	5 6 7 8	1 2 3 4

Note. The numbers refer to the sets of training materials as listed in Appendix C.

Upon completing the fill-in-the-blank exercises, the participants again completed the phrasal decision task (Test 2). At the end of the first day's session, participants were instructed not to study or practice the expressions they learned outside of the sessions until the study concluded.

Both the second and third sessions (Day 2 and 3) started with the declarative vocabulary review first and then the training followed by the posttraining phrasal decision task (Test 3) as shown in Table 6. During the vocabulary review, 32 target items learned on Day 1 were tested again in the identical manner to that of the recall test on Day 1 until they reached 100% accuracy to prevent from any attrition of declarative lexical knowledge and assure that they have explicit knowledge of target items. On Day 2, upon completing the vocabulary review, the training, and the phrasal decision task (Test 3), the participants were given the SRT task followed by the working memory test (OSPAN). On Day 3, upon completing the vocabulary review, the training, and the phrasal decision task (Test 4), the sentence-and-word recall task was administered as a posttraining lexical recognition test (Note that the sentence-and-word recall task was only administered once after the final training session due to a strong practice effect found in the pilot study when repeatedly

administered). Finally, the participants were given the L2 lexical decision task and the language experience and proficiency questionnaire (Marian, Blumenfeld, & Kaushanskaya, 2007).

Table 6. Overview of experimental sessions

Day 1	Consent form
	Declarative vocabulary study
	Test 1 (PDT)
	Training session 1
	Test 2 (PDT)
Day 2	Declarative vocabulary study
	Training session 2
	Test 3 (PDT)
	SRT
	OSPAN
Day 3	Declarative vocabulary study
	Training session 3
	Test 4 (PDT) + SWRT
	L2 LDT
	Questionnaire

Note. PDT: phrasal decision task, SWRT: sentence-and-word recall task, OSPAN: operation span task, LDT: lexical decision task

Chapter 5: Results

As recommended and practiced in previous studies (Lim & Godfroid, 2015; Suzuki & Sunada, 2018), only RTs of correct responses were analyzed (resulting in a removal of 2.9% of the RT data), and no additional data cleaning procedures were employed. Concurring with Lim and Godfroid (2015), Suzuki and Sunada (2018) specifically suggested that a raw data analysis is preferred for automatization studies using RT and CV measures especially when accuracy is high (near ceiling) because data cleaning may artificially and incorrectly affect results.

The assumption of normality was checked via examination of the residuals. Review of skewness and kurtosis statistics, histogram, Q-Q plot, and the boxplot all suggest normality was reasonable for the CV as well as the sentence and word recall data. Before the RT analysis, however, RTs were log-transformed to reduce skewness in the distribution (Baayen, 2008).

The reliability of the RT measures as well as sentence and word recall scores was evaluated with a split-half test and Cronbach's alpha. Table 7 reports Spearman–Brown coefficients or Cronbach's alpha for each variable. As can be seen, all the outcome measures had acceptable reliability of above .8. Thus, the internal consistency of the outcome measures was satisfactory in all cases (Nunnally & Bernstein, 1994).

Table 7. Reliability of the learning outcome measures (N=72)

Outcome measures	Reliability
RT_test1	.948[a]
RT_ test2	.914[a]
RT_ test3	.900[a]
RT_ test4	.919[a]
Sentence recall	.802[b]
Word recall	.902[b]

a. Split-half reliability with Spearman-Brown correction; b. Cronbach's alpha

To respond to the first, second, and third research questions regarding the effects of training on the multiword-recognition skill and the roles of L1 congruency and the two types of training condition in automatization, a mixed-design multivariate analysis of variance (MANOVA) model was performed for the outcome variables (RT and CV) from the phrasal decision task, and ANOVA models were performed separately for outcome variables (sentence recall and word recall rates) from the sentence-and-word recall task. To be more specific, in order to address hypotheses 1-a, 2 and 3 regarding the development of collocation-recognition skill and the roles of congruency and training condition on the automatic collocation recognition, a mixed-design MANOVA with time (Tests 1, 2, 3 and 4) and item type (L1 congruent vs. L1 incongruent) as within-subject factors and group (verbatim repetition vs. varied repetition) as a between-subject factor was conducted for RT and CV as dependent measures in the phrasal decision task. Correlational analyses were also carried out with RT and CV to address hypothesis 1-a. To address hypotheses 1-b, 2 and 3, sentence and word recall performance was examined using a 2×2×2 ANOVA with item type (congruent and incongruent condition), and item status (critical items

associated with the trained collocations vs. control items, i.e. with control counterparts of the trained collocations) as within-subject factors and group (verbatim repetition and varied repetition) as a between-subjects factor. To address the fourth research question regarding the relationship between working memory capacity and automatization in multiword recognition, correlational analyses were carried out with (1) RT and CV in the phrasal decision task, and working memory scores and (2) sentence and word recall scores for the critical items (associated with the trained items) in the sentence-and-word recall task, and working memory scores. To address the fifth research question regarding the relationship between procedural learning ability and automatization in multiword recognition, correlational analyses were carried out with (1) RT and CV, and procedural learning ability scores and (2) sentence and word recall scores for the trained items, and procedural learning ability scores. In order to lower the chances of Type I errors from multiple comparisons, a false discovery rate (FDR) procedure was used where appropriate (Benjamin & Hochberg, 1995).

In the following section, first, descriptive results for the individual difference measures will be presented. Second, the results that pertain to the first three research questions will be presented for each of the outcome measures (RT, CV, and sentence and word recall). Third, the results in response to the fourth research question will be described. Lastly, the results in response to the fifth research question will be presented.

Individual difference measures

Table 8 presents the descriptive statistics of participants' scores on the Ospan task, SRT task, and lexical decision task (in milliseconds), including mean, standard deviation, minimum and maximum scores and the reliability of each measure. As can be seen, all of the individual difference measures had acceptable reliability. Review of skewness and kurtosis statistics, histogram, Q-Q plot, and the boxplot all suggest normality is acceptable for both working memory and procedural learning ability measures. In order to reduce skewness in the distribution, LDT RT data were log-transformed (Baayen, 2008). In addition, there were no significant differences in scores between the two training groups in terms of working memory, procedural learning ability and LDT RT, $t(70) = -.553$, $p = .582$, $t(70) = 1.008$, $p = .317$, $t(70) = 1.540$, $p = .128$, respectively.

Table 8. Descriptive statistics of the individual difference measures (N=72)

	Mean	SD	Min	Max	Reliability
OSPAN	22.24	9.12	6	45	.785[a]
SRT	18.98	11.41	-26.04	45.92	.722[b]
LDT RT	626.30	96.71	451.70	904.22	.934[b]

a. Cronbach's alpha; b. Split-half reliability with Spearman-Brown correction

Language learning outcomes

Phrasal decision task

The means of the RT for each group on each of the four RT outcome measures, the SDs for the mean, and the CVs[6] can be found in Table 9. In addition, the correlation between the mean RT and CV was calculated for each group on each of the outcome measures and presented together in the table. As shown in Table 9, the descriptive statistics showed that both groups performed near ceiling on all of the four tests (95%–99% accuracy). The two groups did not differ in terms of accuracy of their lexical judgements at any time point (all $p > .05$). Thus, participants in both groups engaged in each training session with a solid declarative knowledge of the target items, which is a prerequisite for automatization (Anderson,1992). Recall that both groups started with little to no declarative knowledge of the target collocations, but they acquired and retained the knowledge through the training sessions successfully. More importantly, RTs, standard deviations, and CVs of both congruent and incongruent items generally decreased in both groups across training sessions at the descriptive level.

In order to clearly examine the changes in collocation-recognition processing, a mixed-design MANOVA with Time and Item type as within-subject factors and Group as a between-subjects factor was conducted on the RT and CV scores from the

[6] The CVs are computed as a SD of RT divided by a mean RT and the computation of CV presumes a ratio scale which can be satisfied based on the fact that RTs can have a true zero value.

phrasal decision task. When the multivariate test was statistically significant, univariate analyses were conducted to find out which variables contributed to the overall multivariate significance.

The assumption of equality of error variances (according to Levene's test) was met for the logRT_congruent_Time1 (p = .426), logRT_incongruent_Time1 (p = .385), logRT_congruent_Time2 (p = .139), logRT_incongruent_Time2 (p = .914) , logRT_congruent_Time3 (p = .591), logRT_incongruent_Time3 (p = .120), logRT_congruent_Time4 (p = .265), logRT_incongruent_Time4 (p = .641), CV_congruent_Time1 (p = .292), CV _incongruent_Time1 (p = .297), CV _congruent_Time2 (p = .081), CV _incongruent_Time2 (p = .431), CV _congruent_Time3 (p = .443), CV_incongruent_Time3 (p = .085), CV_congruent_Time4 (p = .382), and CV_incongruent_Time4(p = .568). The equality of covariances assumption (Box's test) was also met (p = .120). The correlations between the dependent variables were of acceptable value (Tabachnick & Fidell, 2012). According to Mauchly's test, the assumption of sphericity was violated for the main effect of Time on RT, $\chi 2$ (5) = 66.124, p <.001; therefore, degrees of freedoms were corrected using Greenhouse-Geisser estimates of sphericity (ε = .664). Additionally, the Mahalanobis distance was computed to examine multivariate outliers. Overall, based on the preliminary tests, no serious violations were noted; the MANOVA was therefore considered robust.

Table 9. Means (and SDs) of RT (in milliseconds), SD, CV, accuracy, correlation between CV and RT (phrasal decision task)

	Verbatim repetition (N=36)					Varied repetition (N=36)				
	Mean RT (SD)	Mean SD (SD)	Mean CV (SD)	Accuracy Rate	r (CV, RT)	Mean RT(SD)	Mean SD (SD)	Mean CV (SD)	Accuracy Rate	r (CV, RT)
Congruent										
Test 1	996.17 (243.88)	296.14 (170.96)	0.28 (0.11)	0.98 (0.04)	.563**	916.52 (271.33)	249.01 (119.40)	0.27 (0.09)	0.99 (0.03)	.266
Test 2	851.10 (162.26)	225.70 (118.52)	0.25 (0.10)	0.97 (0.05)	.766**	857.00 (209.52)	261.08 (150.90)	0.29 (0.11)	0.96 (0.06)	.633**
Test 3	805.12 (137.23)	206.58 (101.53)	0.25 (0.09)	0.97 (0.04)	.640**	771.24 (138.72)	201.92 (125.97)	0.25 (0.11)	0.97 (0.04)	.700**
Test 4	771.51 (144.29)	197.27 (115.82)	0.24 (0.10)	0.97 (0.04)	.642**	710.34 (115.23)	152.55 (74.63)	0.21 (0.08)	0.95 (0.06)	.465**
Incongruent										
Test 1	1046.33 (253.27)	339.73 (199.92)	0.31 (0.11)	0.98 (0.03)	.570**	964.06 (286.18)	286.54 (128.26)	0.29 (0.10)	0.99 (0.03)	.208
Test 2	945.23 (229.21)	299.24 (197.05)	0.29 (0.13)	0.97 (0.04)	.815**	886.88 (215.95)	273.57 (174.59)	0.29 (0.13)	0.97 (0.05)	.664**
Test 3	817.98 (123.15)	208.51 (113.08)	0.24 (0.10)	0.96 (0.05)	.725**	784.32 (164.17)	204.80 (154.12)	0.24 (0.13)	0.96 (0.04)	.647**
Test 4	768.73 (126.20)	175.40 (93.67)	0.22 (0.10)	0.96 (0.06)	.500**	734.54 (129.06)	174.80 (112.87)	0.23 (0.11)	0.96 (0.05)	.650**

** $p < .01$

A multivariate analysis showed significant overall effects of Time, F (6, 65) = 17.260, p < .001, with a very large effect size, $\eta p2$=.614, and of Item type F (2, 69) = 13.252, p < .001, with a large effect size, $\eta p2$=.278 but no significant effect of Group, F (2, 69) = 1.674, p = .195, $\eta p2$=.046. There was a marginally significant two-way interaction between Time and Item Type, F (6, 65) = 2.018, p = .076, with a large effect size, $\eta p2$=.157. The rest of the two-way and three-way interactions were not statistically significant: interaction between Group and Time, F (6, 65) = .533, p = .781, $\eta p2$=.047; interaction between Group and Item Type, F (2, 69) = .323, p = .725, $\eta p2$=.009; and three-way interaction between Group, Time and Item Type, F (6, 65) = 1.297, p = .271, $\eta p2$=.107.

Changes in RT. The univariate follow-up test on the RT showed that there were significant main effects of Time, F (1.991, 139.376) = 57.817, p < .001, with a very large effect size, $\eta p2$=.452 and of Item type, F (1, 70) = 22.564, p < .001, with a large effect size, $\eta p2$=.224 but non-significant effect of Group, F (1, 70) = 2.370, p=.128, $\eta p2$=.033. There was a significant interaction of Time and Item type, F (3, 210) = 3.776, p < .05, with a small effect size, $\eta p2$=.051 and a three-way interaction of Time, Item type, and Group also approached significance, F (3, 210) = 2.226, p= .086, $\eta p2$=.031. The rest of the two-way interactions were not statistically significant; interaction between Group and Time, F (3, 210) = .678, p = .566, $\eta p2$=.010, interaction between Group and Item Type, F (1, 70) = .384, p = .537, $\eta p2$=.005.

Following the main analyses, pairwise comparisons using the FDR correction were conducted to look at the differences between times within each group. The

results revealed that for each item type, both groups' RTs decreased significantly between times (except from Time 1 to Time 2 for congruent and incongruent items in the varied repetition group). The pairwise comparisons for Time are shown in Table 10.

Table 10. Means and p-values for RT differences across training sessions

		T1>T2	T2>T3	T1>T3	T3>T4	T2>T4	T1>T4
Verbatim	Congruent	0.06***	0.02***	0.09***	0.02*	0.04***	0.11***
	Incongruent	0.04**	0.06***	0.10***	0.03**	0.08***	0.13***
Varied	Congruent	0.03	0.04***	0.07***	0.04***	0.08***	0.10***
	Incongruent	0.03	0.05***	0.08***	0.03*	0.08***	0.11***

Note. The mean difference is computed based on log-transformed RT.
*p <.05, ** p <.01, *** p <.001 (FDR-corrected)

Overall, RTs to the trained items (both congruent and incongruent items) declined significantly throughout the three training sessions. For both groups, RTs to the trained items at Time 2, 3 and 4 were always significantly lower than those at Time 1 (prior to any training for procedural learning). Figure 2 graphically presents the improvement in collocation-recognition speed of the two groups across training session.

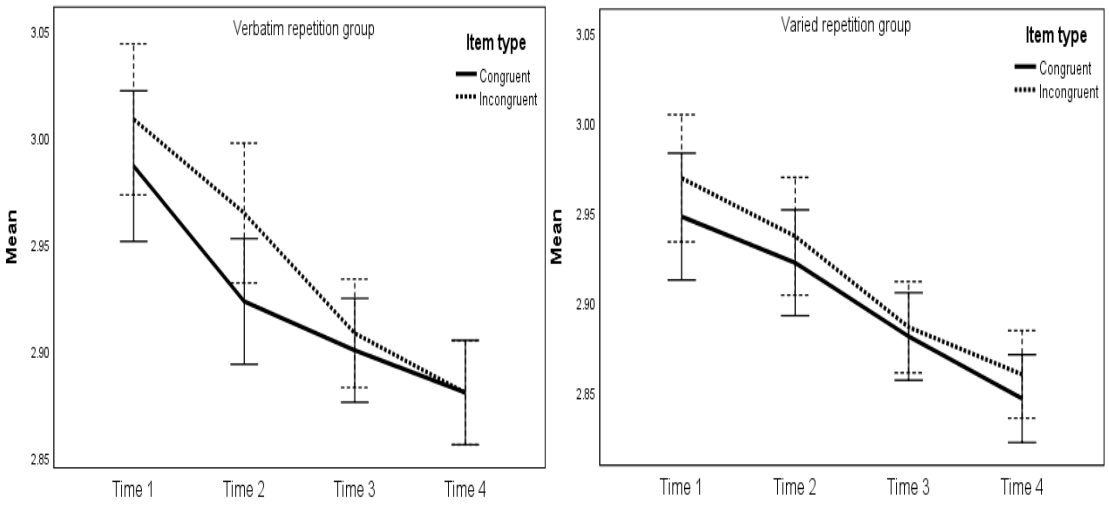


Figure 2. Mean logRT and 95% confidence intervals for two collocation types across training sessions for the verbatim repetition group (left) and the varied repetition group (right)

A significant interaction of Time and Item type revealed that a congruency advantage was present on Test 1 and Test 2 in the verbatim repetition group and only on Test 1 in the varied repetition group. More specifically, RTs to the congruent items were significantly faster than to the incongruent items on Test 1 (MD=-.022, p <.05) and Test 2 (MD=-.041, p <.001) in the verbatim repetition group, and on Test 1 (MD=-.021, p <.05) in the varied repetition group. That advantage, however, was no longer found for Test 3 and Test 4 in the verbatim repetition group and for Test 2, Test 3 and Test 4 in the varied repetition group. The more pronounced advantage of L1 congruency observed in the verbatim repetition group compared to the varied repetition group has contributed to the marginally significant three-way interaction of Time, Item type, and Group. Separate ANOVAs for each group with Time and Item type as independent variables, and their interaction, additionally confirmed that the interaction effect of Time and Item type was significant in the verbatim repetition

group (p = .008) but not in the varied repetition group (p = .401) with the main effects of Time and Item type both significant in both groups (all $p < .05$).

In addition to the main analyses, the effects of Time, Item type and Group, and their interactions with RT as a dependent variable, while controlling for individual differences in aptitudes and speed of processing target language, were tested with an ANCOVA with OSPAN, SRT scores and L2 lexical decision RT as covariates. The assumption of homogeneity of regression slopes was confirmed by the lack of significant interactions between Group and either of the covariates. According to Mauchly's test, the assumption of sphericity was violated for the main effect of Time on RT, $\chi 2$ (5) = 66.124, $p < .001$; therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .664).

The results showed that, after controlling for individual differences in aptitudes and speed of processing the target language, the pattern for the effects of the independent variables and the interactions between them did not change. To be more specific, there were significant effects of Time, F (1.99, 139.38) = 57.82, $p < .001$, $\eta p2$=.452 and of Item type, F (1, 70) = 22.56, $p < .001$, $\eta p2$=.244 with a significant interaction of Time and Item type, F (3, 210) = 3.78, $p < .005$, $\eta p2$=.051. The three-way interaction of Time, Item type, and Group again approached significance, F (3, 210) = 2.23, p= .086, $\eta p2$=.031. However, Group did not show a significant effect, F (1, 67) = 0.39, p=.533, $\eta p2$=.006 and the rest of the two-way interactions were not statistically significant: interaction between Group and Time, F (3, 210) = .68, p =

.566, *ηp2*=.01; and interaction between Group and Item Type, *F* (1, 70) = 0.38, *p* = .537, *ηp2*=.005.

To sum up, the results from the RT analysis highlight two aspects; (a) both groups' RTs to the trained items (both congruent and incongruent items) declined substantially throughout the three training sessions with a congruency advantage in processing speed at early stages of learning and (b) no reliable group difference between the verbatim and the varied repetition was found in terms of processing speed for either congruent or incongruent items at any time point. Patterns for the effects of the independent variables of interest and interactions between them did not change after controlling for individual differences in working memory capacity, procedural learning ability and speed of processing the target language.

Changes in CV. A univariate follow-up test on CV showed that there was a significant main effect of Time, *F* (3, 210) = 13.750, *p* < .001, with a large effect size, *ηp2*=.164 but non-significant effects of Item type, *F* (1, 70) = 2.739, *p*=.102, *ηp2*=.038 and Group, *F* (1, 70) = .056, *p*=.813, *ηp2*=.001. There was no significant interaction effect involving either of Time, Item type, and Group: interaction between Group and Time, *F* (3, 210) = .926, *p* = .429, *ηp2*=.013; interaction between Group and Item Type, *F* (1, 70) = .001, *p* = .975, *ηp2*=.000; interaction between Time and Item Type, *F* (3, 210) = 1.583, *p* = .195, *ηp2*=.022; and a three-way interaction between Group, Item Status and Item Type, *F* (3, 210) = 1.736, *p* = .161, *ηp2*=.024.

Following up on the main effect of Time, pairwise comparisons using the FDR correction were conducted to look at the differences between times within each

group. The results revealed that for the verbatim repetition group, although, descriptively, CVs for congruent items decreased between times (except from Time 2 to Time 3), the decrease in CVs was marginally significant for congruent items only at Time 4 compared to Time 1 while the decreases in CVs were significant for incongruent items at Time 3 compared to Time 1 (and Time 2 with approaching significance), and at Time 4 compared to Time 1 and Time 2. For the varied repetition group, on the other hand, there were significant decreases in CVs for both congruent and incongruent items in the later stages of training phase. Specifically, the decreases in CVs for congruent items approached significance at Time 3 compared to Time 2, but significant at Time 4 compared to Time 1, Time 2 (and Time 3 with approaching significance), for incongruent items, the decreases in CVs approached significance at Time 3 compared to Time 1 and Time 2 but significant at Time 4 compared to Time 1 and Time 2. The pairwise comparisons for Time are shown in Table 11.

Table 11. Means and p-values for CV differences across training sessions

		T1>T2	T2>T3	T1>T3	T3>T4	T2>T4	T1>T4
Verbatim	Congruent	0.03	0.01	0.04	0.00	0.01	0.04
	Incongruent	0.02	0.05	0.07*	0.02	0.07*	0.09***
Varied	Congruent	-0.02	0.04	0.02	0.04	0.08***	0.06**
	Incongruent	0.00	0.05	0.05	0.02	0.06*	0.07*

* p <.05, **p <.01, *** p <.001 (FDR-corrected)

Overall, both groups' CVs for incongruent items substantially decreased to a comparable extent in the later sessions while those of congruent items showed a slightly different pattern; verbatim repetition groups' CVs declined slightly over time

while the varied repetition group's CVs increased slightly (albeit nonsignificant) after the first training session and declined substantially thereafter. For both groups, however, CVs for the trained items at Time 4 (after the last training session) were significantly lower than those at Time 1 (prior to any training for procedural learning). Figure 3 graphically presents the development of collocation-recognition performance of the two groups in processing stability across training sessions.

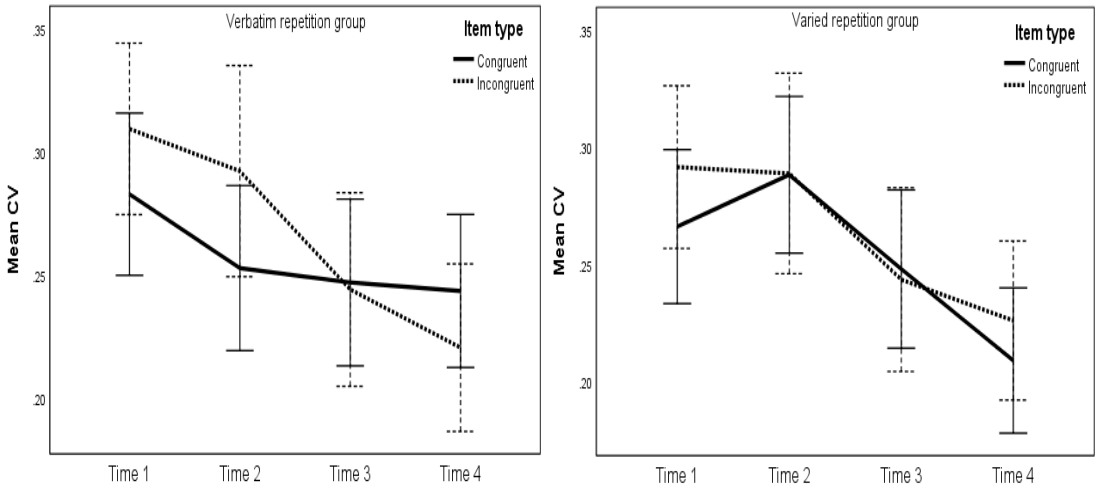


Figure 3. Mean CV and 95% confidence intervals for two collocation types across training sessions for the verbatim repetition group (left) and the varied repetition group (right)

In order to more clearly look into the effect of L1 congruency in the improvement of processing stability, separate ANOVAs for each group with Time and Item type as independent variables, and their interaction was conducted. The results revealed a significant main effect of Time in both groups (all $p < .05$) but a nonsignificant main effect of Item type. Interestingly, however, the interaction effect of Time and Item type was marginally significant with a medium effect size in the

verbatim repetition group (p = .096, $\eta p2$=.058) but not significant in the varied repetition group (p = .471). The marginal interaction in the verbatim repetition group was driven by a marked decrease in CV for incongruent items at Time 3 and Time 4 compared to that of congruent items which remained more or less constant from Time 1 to Time 3 and decreased marginally at Time 4.

Additionally, the effects of Time, Item type and Group with their interactions on CV while controlling for individual differences in aptitudes and speed of processing target language were tested with ANCOVA with OSPAN, SRT scores and L2 lexical decision RT as covariates. The assumption of homogeneity of regression slopes was confirmed by the lack of statistically significant interactions of Group with either of the covariates. The results showed that, after controlling for individual differences in aptitudes and speed of processing the target language, the pattern for the effects of the independent variables and the interactions between them did not change. To be more specific, there was a significant effect of Time, F (3, 210) = 13.75, p < .001, $\eta p2$=.164 but non-significant effect of Item type, F (1, 70) = 2.74, p=.102, $\eta p2$=.038 and Group, F (1, 67) = .10, p=.757, $\eta p2$=.001. There was no significant interaction effect involving either of Time, Item type, and Group: interaction between Group and Time, F (3, 210) = 0.93, p = .429, $\eta p2$=.013; interaction between Group and Item Type, F (1, 70) = .00, p = .975, $\eta p2$=.000; interaction between Time and Item Type, F (3, 210) = 1.58, p = .195, $\eta p2$=.022; and three-way interaction between Group, Item Status and Item Type, F (3, 210) = 1.74, p = .162, $\eta p2$=.024.

To sum up, based on these results (a) both groups' CVs for the trained items (both congruent and incongruent items) overall remained unchanged after the first training session but declined considerably in following sessions and (b) both groups showed a similar level of performance in terms of processing stability for both congruent and incongruent items across the tests. Patterns for the effects of the independent variables of interest and the interactions between them did not change even after controlling for individual differences in working memory capacity, procedural learning ability and speed of processing the target language.

Correlational analyses. The cross-sectional analysis of RT and CV showed that, overall positive and significant correlations were found for both congruent and incongruent items in both groups across the tests (see Table 9 for Pearson correlations between RT and CV across tests). To be more specific, for the verbatim repetition group, RT-CV correlations for congruent items were .563 (p<.001) at Time 1 and .766 (p<.001) at Time 2. At Time 3 and Time 4, the correlations were .640 (p<.001) and .642 (p<.001), respectively. As for incongruent items, RT-CV correlations were .570 (p<.001) at Time 1 and .815 (p<.001) at Time 2. At Time 3 and Time 4, the correlations were .725 (p<.001) and .500 (p<.001), respectively. For the varied repetition group, RT-CV correlations of congruent items were .266 (p=.187) at Time 1 and .633 (p<.001) at Time 2. At Time 3 and Time 4, the correlations were .700 (p<.001) and .465 (p<.001), respectively. As for the incongruent items, RT-CV correlations were .208 (p=.209) at Time 1 and .664 (p<.001) at Time 2. At Time 3 and Time 4, the correlations were .647 (p<.001)

and .650 (p<.001), respectively. Overall, correlation coefficients between RT and CV for both items in both groups increased in all the subsequent posttests compared to that of Test 1 (prior to any training for procedural learning) and remained high and significant. Significant and positive correlations found for both item types on Test 1 in the verbatim repetition group, however, was rather unexpected. This may suggest that some participants in the verbatim repetition group may have had a head start in developing automaticity in collocation recognition for some of the target items after the declarative vocabulary study. Recall that both groups completed extensive practice for the acquisition of declarative knowledge of the target collocations prior to taking the pretest (Test 1). However, it should be noted that the two groups did not differ in their performance in terms of processing stability for both types of collocations on Test 1, thus the significant correlations found in the verbatim repetition group at Time 1 may not be very informative.

To sum up, significant decreases in both RTs and CVs along with the significant positive correlation between them were found for congruent items on Test 4 and Test 3, and for incongruent items on Test 4 in the verbatim repetition group while they were found for both item types on Test 3 and Test 4 in the varied repetition group. This indicates that faster and more efficient recognition of the trained items (congruent and incongruent items) was achieved in both groups by the end of the second training session except for the congruent items in the verbatim repetition condition (it was after the third training session). The results indicate that both groups' improvement in recognition skill after the second and/or the third

training session was associated with automatization as defined by Segalowitz and Segalowitz (1993).

Sentence-and-word recall task

Sentence recall

The performance of the two groups on sentence recall is summarized in Table 12. The maximum possible score for each item condition is 8, representing the total number of sentences to be recalled in each item condition. Figure 4 presents the bar charts of each group on sentence recall scores.

Table 12. Mean scores (and SDs) of the two groups on sentence recall (sentence-and-word recall task)

	Verbatim repetition (N=36)		Varied repetition (N=36)	
Congruent	6.00	(1.64)	6.33	(1.66)
Control for congruent	3.56	(1.56)	3.97	(2.21)
Incongruent	5.36	(1.81)	5.61	(1.68)
Control for incongruent	2.75	(1.57)	3.25	(2.20)

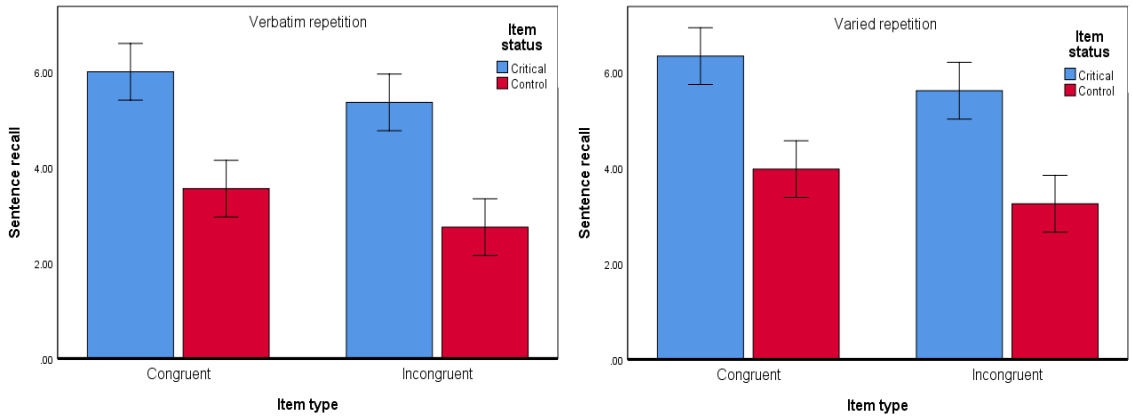


Figure 4. Mean sentence recall scores and 95% confidence intervals for the verbatim repetition group (left) and the varied repetition group (right)

An ANOVA with Group (verbatim repetition and varied repetition), Item status (critical and control items), and Item type (congruent and incongruent conditions) was conducted on sentence recall scores. The equality of variances assumption was met, according to Levene's test (p=.058). The results showed that there were significant main effects of Item status, with large effect size, F (1, 280) = 131.765, $p < .001$, $\eta p2$=.320 and Item type, with a small effect size, F (1, 280) = 11.502, p=.001, $\eta p2$=.039, and a marginally significant effect of Group, F (1, 280) = 3.101, p = .079, $\eta p2$=.011. None of the two-way and three-way interactions were significant: interaction between Group and Item Status, F (1, 280) = .153, p= .696, $\eta p2$= .001; interaction between Group and Item Type, F (1, 280) = .000, p= 1.000, $\eta p2$= .000; interaction between Item Status and Item Type, F (1, 280) = .038, p= .845, $\eta p2$= .000; and three-way interaction between Group, Item Status and Item Type, F (1, 280) = .038, p= .845, $\eta p2$= .000. This model explained 34% of the variance in sentence recall scores (R^2 = .34). The significant main effect of Item type merely indicated that, on average, participants recalled more sentences (both critical and control sentences) in the congruent condition than in the incongruent condition. A marginally significant effect of Group also indicated that, on average, the varied repetition group recalled more sentences accurately than the verbatim repetition group. Note, however, that pairwise comparisons using the FDR correction revealed that none of the mean differences between the varied and the verbatim repetition group in sentence recall scores were statistically significant across four sentence groups.

Following the significant main effect of Item status, pairwise comparisons using the FDR correction within each item type demonstrated that sentences containing the trained items (both congruent and incongruent) were recalled significantly more than those that did not (containing their control items), by both groups. That is, for the verbatim repetition group, more sentences containing the congruent and incongruent items were recalled than those that contained their control items (MD=2.444, p<.001, MD= 2.611, p <.001, respectively). Similarly, the varied repetition group recalled more sentences containing the congruent and incongruent items than those that contained their control items (MD=2.361, p <.001, MD= 2.361, p <.001, respectively).

In order to examine whether the patterns regarding the effects of Item Status, Group and Item type change held, when controlling for individual differences in aptitudes, an ANCOVA was conducted with the working memory and procedural learning ability scores as covariates, using the same independent variables (Group, Item status, and Item type). The assumption of homogeneity of regression slopes was confirmed by nonstatistically significant interactions of Group and either of the covariates. The equality of variances assumption was met, according to Levene's test (p=.057).

The results showed that, after controlling for individual differences in aptitude, the pattern for the effects of the independent variables and interactions between them did not change. There were significant main effects of Item status with a large effect size, F (1, 278) = 142.039, p < .001, $\eta p2$=.338 and Item type, with a

small effect size, F (1, 278) = 12.399, p = .001, $\eta p2$=.043 and a marginally significant effect of Group, F (1, 278) = 3.149, p = .077. None of the two-way and three-way interactions were significant: interaction between Group and Item Status, F (1, 278) = .165, p= .685, $\eta p2$= .001; interaction between Group and Item Type, F (1, 278) = .000, p= 1.000, $\eta p2$= .000; interaction between Item Status and Item Type, F (1, 278) = .041, p= .839, $\eta p2$= .000; and three-way interaction between Group, Item Status and Item Type, F (1, 278) = .041, p= .839, $\eta p2$= .000. This model explained 40% of the variance in sentence recall scores (R^2 = .40), thus the covariates (working memory and procedural learning ability) accounted for an additional 6 % of the performance variance compared to the ANOVA model (R^2 = .34).

To sum up, after three training sessions, both groups showed similar levels of sentence recall performance in both congruent and incongruent conditions, recalling significantly more sentences containing the trained items than those containing their control sequences. The pattern for the effects of independent variables and interactions between them did not change even after controlling for individual differences in working memory capacity and procedural learning ability.

Word recall

For each participant, the sum of correctly recalled stimuli in correctly recalled sentence-trials was computed for each item condition. The maximum possible score for each item condition is 48, representing the total number of additional words to be recalled in each item condition. The performance of the two groups is summarized in Table 13 and visually presented in Figure 5.

Table 13. Mean scores (and SDs) of the two groups on word recall (sentence-and-word recall task)

	Verbatim repetition (N=36)	Varied repetition (N=36)
Congruent	22.39 (11.74)	23.36 (10.64)
Control for congruent	12.53 (7.58)	14.19 (11.93)
Incongruent	17.42 (10.21)	19.33 (10.47)
Control for incongruent	9.03 (6.72)	11.03 (9.84)

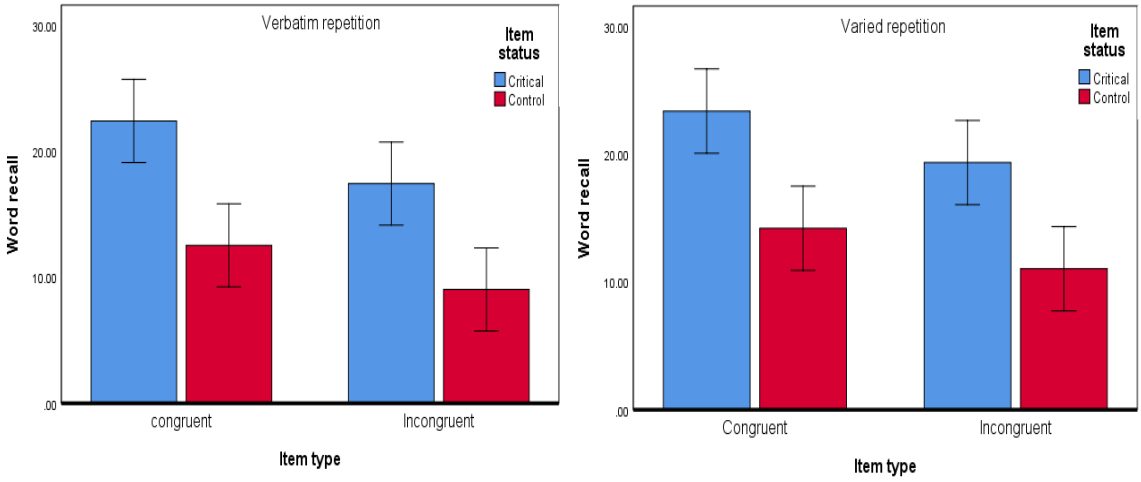


Figure 5. Mean sentence recall scores and 95% confidence intervals for the verbatim repetition group (left) and the varied repetition group (right)

An ANOVA with Group, Item status, and Item type with their interactions was conducted on the word recall scores. The equality of variances assumption was not met, according to Levene's test (p=.018). However, the larger SD (11.93) was less than two times the smaller SD (6.72) thus, ANOVA was considered robust. Similar to the results of the sentence recall scores, the results showed that there were significant main effects of Item status with a large effect size, F (1, 280) = 56.947, p < .001, $\eta p2$=.169 and Item type, with a small effect size, F (1, 280) = 10.953, p=.001,

ηp2=.038 but no significant effect of Group, *F* (1, 280) = 1.918, *p*=.167, *ηp2*=.00. None of the two-way and three-way interactions were significant: the interaction between Group and Item Status, *F* (1, 280) = .027, *p*= .870, *ηp2*= .000; the interaction between Group and Item Type, *F* (1, 280) = .073, *p*= .787, *ηp2*= .000; the interaction between Item Status and Item Type, *F* (1, 280) = .243, *p*= .622, *ηp2*= .001; and the three-way interaction between Group, Item Status and Item Type, *F* (1, 280) = .017, *p*= .897, *ηp2*= .000. This model explained 20% of the variance in word recall scores (R^2 = .20). The significant main effect of Item type merely indicated that, on average, more words were recalled after the critical and control sentences for the congruent condition than the incongruent condition.

Following the significant main effect of Item status, pairwise comparisons using the FDR correction demonstrated that both groups recalled significantly more words after the trained items (congruent and incongruent collocations) than after the control items, which suggests that the learned items took up less attentional resources than the control items. Specifically, for the verbatim repetition group, more words were recalled after the sentences containing the trained collocations, both congruent and incongruent, than after those with their control items (MD=9.86, *p* <.001, MD= 8.39, *p*<.001, respectively). Similarly, the varied repetition group recalled more additional words after the trained collocations than after their control items (MD=9.17, p<.001, MD= 8.31, p<.001, respectively).

Additionally, an ANCOVA with the working memory and procedural learning ability scores as covariates in addition to the same independent variables (Group,

Item status, and Item type) was conducted to examine, when controlling for individual differences in aptitudes, whether the pattern regarding the effects of Group, Item status, and Item type changes. The assumption of homogeneity of regression slopes was confirmed by statistically non-significant interactions of Group and either of the covariates. The equality of variances assumption was not met, according to Levene's test (p <.001). However, the larger SD (11.93) was less than two times the smaller SD (6.72); thus, ANCOVA was considered robust.

The results showed that, after controlling for individual differences in aptitude, the pattern for the effects of the independent variables and interactions between them did not change. There were significant main effects of Item status with a large effect size, F (1, 278) = 65.952, p < .001, $\eta p2$=.192 and Item type with a small effect size, F (1, 278) = 12.685, p < .001, $\eta p2$=.044 but no significant effect of Group, F (1, 278) = 2.039, p = .154. None of the two-way and three-way interactions were significant: interaction between Group and Item Status, F (1, 278) = .031, p= .860, $\eta p2$= .000; interaction between Group and Item Type, F (1, 278) = .084, p= .772, $\eta p2$= .000; interaction between Item Status and Item Type, F (1, 278) = .281, p= .596, $\eta p2$= .001; and three-way interaction between Group, Item Status and Item Type, F (1, 278) = .019, p= .890, $\eta p2$= .000. This model explained 32% of the variance in word recall scores, which is larger than that of ANOVA model (R^2 = .20), indicating that adding covariates (working memory and procedural learning ability) contributed to explaining a larger amount of performance variance.

To sum up, as was the case for the sentence recall scores, after three training sessions, both groups showed a similar level of performance in their word recall accuracy for both congruent and incongruent collocations compared to their controls in their word recall accuracy, recalling significantly more words after the critical sentences (sentences containing the trained collocations) than after those containing the control sequences. The pattern for the effects of independent variables and interactions between them remained unchanged even after controlling for individual differences in working memory capacity and procedural learning ability.

Cognitive aptitude and learning outcomes

Working memory capacity and learning outcomes

Table 14 presents Pearson's correlation coefficients of OSPAN scores with RT and CV for the phrasal decision task. The FDR correction was conducted to lower the chances of Type I errors. The magnitudes of correlation coefficients (r) were interpreted based on Plonsky and Oswald (2014). The results revealed that the relationship between RT and working memory scores for both congruent and incongruent items was weak (almost negligible) throughout the training with no statistically significant correlations. Similar results were found for the relation between CVs and working memory scores. None of the working memory correlations with the indices of automatization turned out significant across the tests.

Table 14. Pearson correlations of working memory scores with RT and CV

		Congruent	Incongruent
RT			
	Test 1	.169	.059
	Test 2	.046	.097
	Test 3	.064	.022
	Test 4	-.091	-.113
CV			
	Test 1	-.014	-.070
	Test 2	.075	.005
	Test 3	-.109	-.053
	Test 4	-.218	-.218

Note. None of the correlations were statistically significant (FDR-corrected).

In light of the finding that working memory did not turn out to be significantly related to RT or CV at any time point, further correlational analyses were conducted to see whether meaningful association with working memory is observable in accuracy scores in the phrasal decision task (See Table 15). The results showed significant small-to-medium size correlations between accuracy and working memory scores for congruent items on Test 3 and incongruent items on Test 4, suggesting that working memory capacity plays a facilitative role in recognition of target collocations at these later stages (Test 3: r = .300, p = .040 for congruent items; Test 4: r =. 333, p = .004 for incongruent items). None of the other correlations turned out to be statistically significant. The results, however, should be interpreted with caution due to ceiling performance on the accuracy measure (range =95%–99%).

Table 15. Pearson correlations of working memory scores with accuracy

	Congruent	Incongruent
Accuracy		
Test 1	-.028	.051
Test 2	-.013	.167
Test 3	.300*	-.003
Test 4	-.094	.333**

* p <.05 ** p <.01 (Bootstrapped; FDR-corrected)

Table 16 presents Pearson's correlation coefficients pertaining to working memory scores with sentence and word recall scores of the sentence-and-word recall task. The results showed marginally significant, small size, positive correlations between sentence recall scores and working memory scores for both congruent (r=.210, p =.076) and incongruent items (r=.238, p =.076), suggesting association between higher working memory capacity and better sentence recall performance for both types of collocation items. Similarly, for word recall scores, small-to-medium size, positive correlations were found with working memory scores for both congruent and incongruent items (r=.286, p <.05 for congruent items and r=.365, p <.01 for incongruent items), significant this time, indicating association between higher working memory capacity and better word recall performance for both types of collocation items.

Table 16. Pearson correlations between sentence and word recall scores, and working memory scores

	Congruent	Incongruent
Sentence Recall	.210	.238
Word Recall	.286*	.365**

*p<.05 **p <.01 (FDR-corrected)

Procedural learning ability and learning outcomes

Table 17 presents Pearson's correlation coefficients of SRT scores with RT and CV of the phrasal decision task. Across the tests, a weak relationship between RT and procedural learning ability scores for both congruent and incongruent items was found with no statistically significant correlations. Similarly, CVs for both items across the tests were not associated with procedural learning ability scores.

Table 17. Pearson correlations between RT and CV, and procedural learning ability scores

	Congruent	Incongruent
RT		
Test 1	.050	.033
Test 2	-.030	.099
Test 3	-.006	-.129
Test 4	-.119	-.108
CV		
Test 1	-.017	.013
Test 2	-.132	.182
Test 3	-.173	-.231
Test 4	-.122	-.074

Note. None of the correlations were statistically significant (FDR-corrected).

However, it is noteworthy that, in the within-group analysis, RTs on the Test 4 for the congruent items and CVs on the Test 3 for both congruent and incongruent items in the varied repetition group were significantly associated (not FDR-corrected) with the SRT scores (RT: r = -.393, p = .018 for congruent items; CV: r = -. 364, p = .029 for congruent items and r = - .400, p = .016 for incongruent items), suggesting association between higher procedural learning ability and faster RT and smaller CV. However, this result was not significant when FDR-corrected, thus, it can only be considered as suggestive. Further correlational analyses were again conducted to see

whether SRT scores were related to accuracy in the phrasal decision task; the relationship was weak across the tests with none of the correlations reaching significance (see Table 18).

Table 18. Pearson correlations between accuracy and procedural learning ability scores

	Congruent	Incongruent
Accuracy		
Test 1	.020	.061
Test 2	-.243	.010
Test 3	-.056	.075
Test 4	-.194	-.056

Note. None of the correlations were statistically significant (Bootstrapped; FDR-corrected)

Table 19 presents Pearson's correlation coefficients pertaining to SRT scores with sentence and word recall scores of the sentence-and-word task. As can be seen, the relationship was weak for both recall scores of both item types with no statistically significant correlations.

Table 19. Pearson correlations between sentence and word recall scores, and procedural learning ability scores

	Congruent	Incongruent
Sentence Recall	.088	.126
Word Recall	.184	.202

Note. None of the correlations were statistically significant (FDR-corrected)

Chapter 6: Discussion

The present study investigated the effects of practice in collocation recognition on the development of automaticity in collocation-recognition processing, and also examined the influences of L1 congruency, training condition (verbatim repetition and varied repetition), working memory capacity, and procedural learning ability in the automatization of collocation recognition. Congruent (translated forms of Korean-only collocations) and incongruent collocations consisted of the target of acquisition. A phrasal decision task, and a sentence-and-word recall task were used to assess the participants' learning gains from the training. Each of the research questions will be discussed in light of the results.

Development of automaticity in collocation recognition

Research question 1

With regards to the first research question, it was hypothesized that collocation-recognition training would lead to more skilled recognition of the target collocations (as indicated by significant decreases in RTs, and CVs with significant positive correlation with mean RT in the phrasal decision task and better recall performance in terms of sentence and word recall rates for the trained items than their controls in the sentence-and-word recall task). The results from both phrasal decision task, and the sentence-and-word recall task, showed that collocation recognition

training (both verbatim repetition and varied repetition) was effective in improving collocation-recognition performance in terms of speed, stability and cognitive capacity usage[7]. To be more specific, the results of the RT and CV analyses showed that, participants improved in performance over time, becoming faster and less variable on recognition of the target collocations, as reflected in significant overall decreases in RT and CV. It is noteworthy that the learners in both groups showed an improvement in their speed of collocation recognition even after the first training session and kept making significant gains after every training session afterwards. Thus, the repeated practice was effective for promoting speed of lexical processing as claimed and found by a number of scholars (e.g., Akamatsu, 2008; Fukkink, Hulstijn, & Simis, 2005; Gatbonton & Segalowitz, 2005; Wood, 2009).

With regards to processing stability, both groups showed a significant improvement after the second and/or the third training session (even though the learners engaged with each target item in exercises only 8 times by the second training session and 12 times by the end of the training phase). However, the findings of CVs remaining broadly constant during the initial stage of training may seem inconsistent with the findings in Solovyeva and DeKeyser (2018) who reported an increase in CV during the initial stages of novel word learning as a result of

[7] One may argue that the tasks used in the current study may not necessarily involve retrieval of meaning. Considering the significant congruency effects found in both tasks, however, a certain degree of meaning retrieval was likely to be involved for completing each task. Recall that efforts were also made to encourage participants to attend to meaning by including filler items that were semantically implausible rather than syntactically implausible in the phrasal decision task, as well as control items that are semantically plausible and synonymous to the trained items in the sentence-and-word recall task.

establishment of mental representations of the learned items. The inconsistencies, however, may be attributable to differences between the studies. For example, in Solovyeva and DeKeyser (2018), a pretraining task (specifically in the case of the stimuli related to form-meaning mapping, which is more comparable to the present study) was administered prior to any practice with the target words. The pretraining task required the learners to respond to L1 target words (that were primed by the learned novel words) instead of L2 target words. In the case of the stimuli related to learning word forms, their RT and CV data were based on the NO responses to the nonword targets, which is atypical use of RT data to assess lexical acquisition. As acknowledged by the authors, this conflict between growing familiarity with the nonword targets and having to treat them as non-existing words (requiring no responses on lexical decision) may have contributed to their results. In the present study, however, the pretraining task was administered after the declarative vocabulary study phase, which was similar to the training provided in Solovyeva and DeKeyser (2018), in which the learners went through three cycles of presentation of target items (presented twice) followed by a short recall test to ensure solid declarative knowledge of the form and meaning of each target item before engaging in training for procedural learning. Thus, the target items may have already been integrated in the mental lexicon even before the pretraining task. It is also worth noting that the pre- and posttraining task (phrasal decision task), in the present study, always required the learners to respond to the L2 target items with yes responses. The absence of noticeable change in CV in the early stages of training (followed by its significant

decrease in the later sessions) found in the present study, on the contrary, seems to provide a potential explanation for the absence of an expected decrease in CV reported in previous studies (e.g., Hulstijn et al., 2009).

Beyond showing increasing speed and stability, RT and CV further provided evidence for improvement in automaticity of collocation-recognition performance through correlational analyses. Specifically, significant positive correlation between RTs and CVs accompanied by significant decreases in RT and CV, observed for both groups at later sessions of training, indicate that both groups' improvement in collocation-recognition skill after the second and/or the third training sessions was associated with automatization as defined by Segalowitz and Segalowitz (1993). This replicates findings of previous studies using the CV to investigate automatization of L2 lexical recognition (e.g., Akamatsu, 2008; Elgort, 2011; Ma et al., 2017). The results are also consistent with the findings of previous studies that reported robust effects of explicit vocabulary activities on the development of L2 word-recognition fluency (Akamatsu, 2008; Elgort, 2011; Fukkink et al., 2005; Pellicer-Sánchez, 2015). Using RT and CV measures, Pellicer-Sánchez (2015), for example, reported that various types of explicit vocabulary activities (including a gap-filling task) led to faster and more automatic processing of single words. It is also worth noting that the obtained effects remained robust even when retention was tested one month after the learning in this study. The large effects of fill-in-the-blank exercises on improving speed and efficiency of collocation-recognition performance found in the current study seem encouraging and have important pedagogical implications. As claimed by

Boers (2014) and Boers, Dang, and Strong (2017), fill-in-the-blank exercises provide effective means to develop skilled recognition of target collocations, even with a relatively short period of practice, by having learners focus on accessing and retrieving the correct meaning of each target expression.

The accuracy data for the sentence-and-word recall task also provided corroborating evidence; two groups recalled significantly more sentences containing the trained items (both congruent and incongruent) than those containing their control sequences and also recalled significantly more words after the sentences containing the trained items, suggesting that the learned items required less cognitive resources than the control items. In other words, the results indicated that the trained items offered processing advantages over equivalent control items, allowing more time and cognitive resources to be used for the additional task such as memorizing additional words. Such processing advantages can be considered as reflecting reduced cognitive capacity usage associated with automatic recognition of the learned collocations after the three training sessions. It is important to note that the effects of training were all significant with equivalently large effect sizes in the analyses of both sentence and word recall scores. As sentence recall may involve not only recognition skill but also production skill (which is not the focus of the training in the current study), the possibility had been considered that this could possibly lead to a weaker effect of recognition training on the sentence recall rates. However, as hypothesized, the effect of the production component in the sentence recall seems to have been negligible due to the fact that the constituent words of the target items were all highly familiar

words, and that any minor spelling mistakes for the target items were ignored in the scoring process.

Overall, the combined results of the phrasal decision task, and the sentence-and-word recall task showed that faster and more automatic recognition of the trained items (both congruent and incongruent items) was achieved in both groups as a result of the training sessions. Consistent with the predictions of skill acquisition theory, these results, therefore, provide support for H1-a and H1-b in that the practice in collocation recognition led to faster and more automatic recognition performance of the target collocations. The converging evidence obtained from both the phrasal decision task and the sentence-and-word recall task in the current study supports the usefulness of CV as an indicator of automatic performance of lexical recognition (even for multiword expressions such as collocations), at least for the task used in the current study and of a sentence-and-word recall task as a means for assessing the development of automaticity in multiword recognition.

Research question 2

The second research question concerns whether L1 congruency affects the development of automaticity in collocation recognition. Due to the dearth of empirical research on the role of L1 in automatization of a lexical item, three possibilities were considered based on previous literature: a facilitative role of L1 congruency in all stages of collocation learning, a facilitative role of L1 congruency only during the early stage of learning collocations (in the acquisition of declarative knowledge but not in automatization), and a facilitative role of L1 incongruency in

automatization. The results indicated that no qualitative difference in multiword-recognition improvement was found between the two collocation types across the training sessions. Specifically, improvement in collocation-recognition performance with both congruent and incongruent items after the first training session was associated with simple speedup while that of the middle and the end of the training sessions were associated with automatization as defined by Segalowitz and Segalowitz (1993). Performance with both congruent and incongruent items improved in automaticity largely to a comparable extent at the later stages of the training as indicated by significant decreases in CVs to similar extents and significant correlations of CVs and RTs at the later stages of learning. Moreover, the significant positive association between RTs and CVs that was observed for both item types at the end of the training sessions in both groups appears to indicate that the amount of practice (three training sessions) was not sufficient to bring about the optimal level of recognition skill even for the congruent collocations. These results overall indicate that the improvement in the collocation recognition with both congruent and incongruent items during training involved similar processes across the training sessions, that is, from proceduralization to automatization. There was, however, a slight hint of advantage for the incongruent items at the later stages of the training in terms of cumulative gains in automaticity in collocation processing. That is, when comparing the performance of the later stages of the training (Test 3 and Test 4) to that of Test 1 (prior to any training for procedural learning), greater evidence of automatization (as reflected in RT and CV) was found for the incongruent items than

that obtained for the congruent items, which is in line with the desirable difficulty perspective (e.g., Bjork & Kroll, 2015; Brown, Roediger, & McDaniel, 2014, Suzuki, Nakata, & DeKeyser, 2019). These results should be interpreted with caution, however, as there were no statistically significant differences in CV found between the two collocation types at any time point.

It is important to note, however, that the significant congruency effect was found in processing time in both groups at the early stages of learning, in line with the robust advantage of L1 congruency reported in collocation processing studies (e.g., Wolter & Gyllstad, 2011, 2013; Wolter & Yamashita, 2018; Yamashita & Jiang, 2010). The RT analysis showed that the congruent items were processed significantly faster than incongruent items on Test 1 and Test 2 in the verbatim repetition group, and on Test 1 in the varied repetition group. The advantage of congruency in processing time, however, disappeared with repeated practice. That is, with increasing practice, performance with incongruent items approached performance with congruent items in terms of processing time. This finding seems to indicate that the advantage of L1 congruency in processing time is only present at the early stages of learning and tends to diminish relatively quickly with increasing practice with the target expressions. These results support the hypotheses claimed by Yamashita and Jiang (2010) that at the initial stage of learning collocations, learners rely on the L1 mediation process, which results in the processing advantage of congruent collocations. The recognition of a congruent collocation is much faster due to its link between its existing L1 counterpart, that is linked to the concept. The recognition of

an incongruent collocation, on the other hand, is slower because the process involves more steps due to the lack of L1 equivalent expression, such as word-for-word L1 translation, then to corresponding L1 collocation that shares the same meaning, and finally to the concepts. They further claimed that "with increasing input, L2 collocations, both congruent and incongruent, may become multiword units that are no longer dependent on the L1 lexicon and that are directly connected to concepts" (p. 653). In line with their hypothesis, they found that advanced ESL learners showed no difference in processing time of recognizing congruent and incongruent collocations while lower-proficiency EFL learners showed a congruency effect. Therefore, in the present study, recognition of the target collocations during the early stages of learning involved the mediation and activation of the corresponding L1 expressions. However, repeated practice appears to have enabled learners to process both congruent and incongruent collocations with little involvement of L1 due to the establishment of a direct link between the target item and its concept.

Peters (2016), however, reported results inconsistent with the present findings as well as those in the previous studies. She found an absence of congruency effect on learners' recognition skill in the immediate posttest following four explicit vocabulary activities (including a fill-in-the-blank exercise). Although the focus of that study was limited to learners' initial form–meaning mappings of new collocations and not automatization of collocation knowledge, congruency was not found to be a strong predictor of accurate responses in a form recognition test, but it was in the productive recall test. Based on the findings, she suggested that

congruency may not influence the learning burden of collocations in recognition but in production only. The results in her study, however, should be interpreted with care. First, as the author also pointed out, target collocations included some collocations that were similar in form and meaning, which may have affected the results. Next, a matching test used as the recognition test, in which participants were asked to match the individual constituents of the collocations, may not have been an appropriate measure to assess learners' form-meaning mappings in recognition. As the test can be easily completed with knowledge of the forms (the compositions of collocations) at the surface level, the test may not necessarily involve retrieval of meaning.

In terms of improving processing efficiency, the present CV results indicated that performance with both congruent and incongruent items showed gradual improvement in processing stability largely to a comparable extent across the two groups until the end of the training session. However, a marginally significant interaction of Time and Congruency found in the verbatim repetition group appears to indicate that the advantage of repeated practice might be more pronounced for the incongruent items compared to the congruent items in the verbatim repetition group. That is, while a significant gain in processing stability was observed at Time 3 and Time 4 (after the second and the third training session) for the incongruent items, performance with the congruent items from Time 1 to Time 3 remained rather unchanged in terms of processing stability and showed a slight gain (marginally significant) at Time 4 only. Building on an assumption made by Yamashita and Jiang (2010), this finding may suggest that although repeated practice led to the

establishment of direct connections for both congruent and incongruent collocations and their corresponding concepts, as indicated by the RT data, in order for that link to become stronger such that access to the concept operates automatically and autonomously, more practice may be required for congruent collocations than incongruent collocations due to its strong link with existing L1 collocations. Incongruent collocations, on the other hand, may have an advantage in this process due to the lack of direct translational L1 equivalents. However, it should be noted that the interaction effect found was marginally significant (with a medium effect size) in the verbatim repetition group only with no statistically significant differences in CVs between the two collocation types at any time point. Thus, this possibility remains speculative and should be examined with further studies.

The results of the sentence-and-word recall task provided no evidence that congruency plays a role in the development of automaticity in collocation recognition. Although no strong hypothesis was made regarding the role of congruency in sentence and word recall performances, it was anticipated that if the recognition performance could have reached the optimal level during the training sessions with the congruent collocations only, then a larger effect of the training for the congruent items compared to the incongruent items would be observed as the congruent items would be recognized with maximal efficiency, thus requiring less cognitive resources than the incongruent items. However, no such effect was found in the sentence and word recall performance in both groups. Both sentence and word recall performance showed a similar pattern for the congruent and incongruent items

in both groups. This finding, however, is understandable considering that the sentence-and-word recall task was administered only once at the end of the last training session and the similar level of performance for both collocation types was achieved after the three training sessions as indicated by the RT and CV data.

Considering the effect of L1 congruency on speed of collocation recognition only at the initial stages of learning but the lack of any reliable effect on automatization, it is likely that congruency plays a facilitative role in the acquisition of declarative knowledge but not in automatization. However, as the learners appear to have been able to reach only a relatively early stage of automatization for both types of items by the end of the three training sessions, more complete tracking of collocation learning (involving all three stages of skill acquisition) over an extended period of time is necessary to fully explore this possibility. Moreover, effects of long-term learning could have been more reliably observed if the phrasal decision task had been administered at the beginning of the second and the third sessions. The effects of L1 congruency in automatization as well as more of an effect for procedural learning might have been observed in those tests.

Overall, congruency was found to play a limited role, only in terms of speed of recognition at the initial stage of learning; the advantage of congruency did not extend to improvement of processing efficiency or reduction of cognitive capacity requirements. More importantly, no qualitative difference in the improvement of collocation-recognition performance with respect to L1 congruency was found. Learners in both groups followed similar developmental stages of procedurlization

and automatization for both types of collocations throughout the training. Future research including a greater amount of practice and different types of linguistic knowledge (productive skill) is necessary for a more comprehensive understanding of the role of L1 congruency in the acquisition process.

Research question 3

The third research question concerns the role of different types of training condition (verbatim repetition vs. varied repetition) associated with the exercises in the development of automaticity in multiword recognition. Based on previous research (De Jong & Perfetti, 2011; Durrant & Schmitt, 2010), it was hypothesized that the verbatim repetition group would benefit more from the training than the varied repetition group, thus producing greater learning outcomes. On the contrary, based on the desirable difficulty framework (e.g., Brown, Roediger, & McDaniel, 2014, Schmidt & Bjork, 1992; Suzuki, Nakata, & DeKeyser, 2019), the possibility that the varied repetition group might show better performance than the verbatim repetition group by the end of the last session despite worse performance during the earlier training sessions was also considered. The results of both phrasal decision task and the sentence-and-word recall task showed no reliable group differences for any of the outcome measures (RT, CV, and sentence and word recall scores) at any time point. Both conditions of the training yielded significant improvement in collocation-recognition speed and efficiency to a comparable degree.

This finding is in contrast with the results of Durrant and Schmitt (2010) which reported robust advantage of verbatim repetition condition over varied

repetition condition on collocation learning. The current study, however, differs from the study carried out by Durrant and Schmitt (2010) in many aspects, which may have contributed to the different results. For example, they examined the development of procedural knowledge of the surface form of collocations (form recall of adjective-noun associations) under an incidental learning condition. They used a naming test to assess learning, and this test may not have been an adequate measure of learning since it may have greatly primed the recall. To be more specific, in the naming test, participants were asked to say the correct noun that followed the given adjective for each target adj-noun pairs. The adjective and the first two letters of the noun were provided as hints, which may have significantly primed the recall of the target pairs learners had encountered during the training especially with no filler items included. Moreover, the number of target items (that were already assumed to be known by the participants) was too few (10), and the number of the encounter (only twice) may have been not sufficient to result in development of any type of procedural knowledge that is durable. Thus, the results of their study have to be interpreted with great caution.

As mentioned earlier, another potential reason for the absence of a group effect in the current study may be due to the fact that amount of practice may have not been sufficient to result in any significant differences between the two training conditions. As the participants seemed to have only engaged in the relatively early stage of automatization, due to the relatively short period of practice, more practice over an extended period of time might have been needed for the two types of training

condition to exert any differential effects. Additionally, the differential effects of the two groups might have been observed if the participants had been tested for their long-term retention in the beginning of the second and the third sessions.

However, it is important to note that there was a hint of an advantage for verbatim repetition over varied repetition. For example, learners in the verbatim repetition group significantly improved in their speed of recognition for both item types even after completing the first training session while the varied repetition group did not make any significant gains during the first training session. Moreover, the analysis of gain scores (where gain is defined as each participant's RT and CV scores subtracted from those for the pretest) showed that greater gains for congruent items (although marginally significant with the FDR correction) in both RT and CV (RT: $t(70) = 1.734$, $p=.07$; CV: $t(70) = 1.847$, $p=.069$) were made at Time 2 in the verbatim repetition group compared to the varied repetition group[8]. However, no indication of such an advantage for the verbatim repetition condition was observed in later stages of training. Considered together, this may suggest that the verbatim repetition condition is effective for improving performance (especially for congruent collocations) at the early stage of learning but not during the automatization stage. It is worth noting that Durrant and Schmitt (2010) also found the superiority of verbatim repetition over varied repetition in the initial stage of learning collocations. Future research is needed to explore this possibility.

[8] Note that the two groups did not differ in their performance with both types of expressions in terms of speed of recognition (controlling for individual differences in processing the target language) and processing stability at Time 1 (prior to any training for procedural learning).

In sum, Hypothesis 3 was not supported, as no reliable advantage was found for either of the training conditions in any of the outcome measures. Future research including more practice are needed to clearly understand the role of two different types of repetition conditions in the acquisition process. Further work with a different type of linguistic knowledge (productive skill) can also help identify why the current study led to results that conflict with some of the literature (e.g., Durrant & Schmitt, 2010).

Cognitive aptitude and skilled recognition of a collocation

Research question 4

The fourth research question was related to the role of working memory capacity in automatization of multiword recognition. Based on previous literature (e.g., Akamatsu, 2008; Daneman & Merikle, 1996; Linck et al., 2014), it was hypothesized that participants with greater working memory capacity would have more advantages in benefiting from multiword-recognition training than those with less working memory capacity. The results showed that working memory capacity was significantly associated with collocation-recognition performance as measured by the sentence-and-word recall task for both congruent and incongruent items at the end of the last session and accuracy of the phrasal decision task at the later stages of the training, but not with RT and CV. Overall, the magnitudes of correlations were within weak to moderate range in terms of effect sizes.

The finding that working memory did not have any reliable relationship with improvement in collocation-recognition speed or efficiency as measured by RT and CV is in line with Akamatsu (2008) who also failed to find a significant association between working memory scores, and gains in RT or CV of single-word recognition task following a word recognition training. It may be possible that a different subprocess of working memory such as updating or inhibition is more relevant to speedup or processing stability of collocation-recognition performance. Pulido and Dussias (2019) reported a significant contribution of inhibitory control (as measured by the Flanker task) in RT of productive recall test for incongruent collocations and suggested that the ability to resolve conflict between competing representations may be particularly relevant to processing incongruent collocations.

The association found between working memory capacity and accuracy at the later stages of learning (albeit weak-to-moderate) with no reliable association found in the earlier stages of learning was rather puzzling. This may be due to stronger ceiling effects in the early stages of training (98% correct on average; SD = 0.03), which artificially restricts variability and can produce misleading small correlation coefficients (Aydin et al., 2018). Another possibility is that working memory capacity may be more relevant for later stage of collocation learning which involves establishing direct connections between the target collocations and the corresponding concept, and becoming capable of accessing those links automatically in recognition. Unlike non-idiomatic single words, linking those idiomatic expressions to the corresponding concept and learning to access those

links automatically may require greater attentional resources and attentional control. Thus, reliable association could have been observed only in the later stages of training.

Note, however, that working memory also showed small to medium relationship with collocation-recognition performance as measured by the sentence-and-word recall task at the end of the last training session in the current study. The results are consistent with Pulido and Dussias (2019) who found a significant role of working memory (as measured by reading span task) in collocation-production performance following three sessions of production practice. The results are also partially in line with findings in Martin and Ellis (2012) which reported a significant correlation between working memory (as measured by listening span task) and receptive vocabulary abilities at relatively early stage of learning. Considered together, it can be argued that working memory may play an important role until a sufficient amount of automaticity has been reached, as suggested by Serafini and Sanz (2016).

One potential explanation for the inconsistencies in the results could be that the sentence-and-word recall task can be more holistic measure than RT and CV measures, that taps into holistic processing of multiword units. Recall that the Hypothesis 4 was based on the assumption in previous literature that individual differences in working-memory capacity may support processing a word or a sequence of words (e.g., collocation) as a whole unit and reading words through direct connections between the forms and their meanings in memory (Akamatsu,

2008), which is, as claimed by Ehri (1991), closely related to increasing efficiency in word-recognition processing. In fact, the sentence-and-word recall task was originally used to assess whether lexical bundles were processed as whole units by examining cognitive capacity requirements associated with the target lexical bundles in Tremblay et al. (2011). Therefore, it is possible that individual differences in working memory became observable in the performance measured by the sentence-and-word recall task only, which was more likely to pick up on the holistic processing of the target collocations.

Another possibility is that encoding of collocations as unitary chunks may particularly draw more on implicit learning, as suggested by Ellis (1994, 2001). Thus, it may be possible that working memory capacity (as measured by OSPAN task) played less of a role for collocations overall (little role in accuracy at the early stages of learning, and RT and CV) than what has been expected based on that for single words (cf. Elgort et al., 2018; Martin & Ellis, 2012).

On the other hand, one may argue that the discrepancy in results may relate to the nature of the task itself. As the sentence-and-word recall task draws heavily on many cognitive resources by having the participants comprehend, memorize and then recall sentences and additional unrelated words following them, learners may have had to draw on working memory to a great extent to improve their scores. Thus, learners with high working memory capacities may have had an inherent advantage

over those with lower working memory capacities[9]. Future research with different measures of automaticity that may also tap into holistic processing of multiword units and a different measure of working memory that taps into a more fine-grained subcomponent of executive control (e.g., a running memory span test, the Flanker test) may help determine the role of working memory in the improvement of collocation-recognition skill.

Research question 5

The fifth research question was related to the role of procedural learning ability in automatization of multiword recognition. It was hypothesized that participants with greater procedural learning ability would have more advantages in benefiting from multiword-recognition training than those with less procedural learning ability. The correlational analyses showed that neither RT, nor CV, nor sentence and word recall scores were significantly associated with procedural learning ability at any time point. The results are not in line with the findings reported in Suzuki (2017) who found a significant correlation between RT reduction (in a production test of grammatical rules) and procedural learning ability nor with Pili-Moss et al. (2019) who found a significant relationship between CV (in a comprehension test of grammatical rules) and procedural learning ability. The present results were puzzling at first, but considering the limited number of training sessions provided over a short period of time, as mentioned earlier, it may be possible that the

[9] Note, however, that the effects of interest remained significant even after controlling for individual differences in working memory capacity.

majority of the participants have not yet reached the optimal level of collocation-recognition performance but only the initial stage of automatization.

It is important to note, however, that procedural learning ability showed small to medium relationships with RT of congruent items and CVs of both congruent and incongruent items in the varied repetition group at later stages of the training phase (although it did not reach statistical significance when FDR-corrected). These findings may be suggestive of a differential role of procedural learning ability in different training conditions and different stages of automatization of L2 multiword recognition skill. The fact that no meaningful association was found between RT or CV and procedural learning ability in the verbatim repetition group may be due to the simple nature of the task in the training (exercises with the same sentential context for each target item repeatedly). To be more specific, repeated encounter with the same language input may have made it easier to focus on building up fluency with the target items without the distraction of dealing with new contexts, thus requiring little reliance on procedural learning ability. This may be why no individual differences in procedural learning ability became observable. By contrast, because the varied repetition condition involved an additional task of dealing with new contexts in each exercise on top of building fluency with the target items, learners may have had to draw more on procedural learning ability. These possibilities should be examined in future research to more clearly understand the role of the procedural learning ability in the development of multiword recognition skill.

Chapter 7: Conclusion

The present study is unique in methodology by targeting multiword lexis that has not been examined within the framework of skill acquisition theory, and by adopting three different measures of automaticity (speed, variability of speed, and cognitive capacity requirements) to more fully assess the development of L2 collocation-recognition skill, the study attempted to provide important insights into how automaticity in collocation-recognition skill develops through practice, and what the important contributing factors are. Specifically, the aim of this dissertation was to investigate the effects of multiword-recognition training on the development of automaticity in multiword-recognition processing, and also to examine the roles of L1 congruency, training condition (verbatim repetition vs. varied repetition), working memory capacity, and procedural learning ability in the development of automaticity in multiword recognition.

In line with the predictions of skill acquisition theory, the present study showed that repeated practice through explicit vocabulary exercises (fill-in-the-blank) led to significant improvement in recognition performance of collocations in terms of speed, processing stability and cognitive capacity usage. The findings have important implications for second language pedagogy as the learners were able to improve their collocation-recognition skill substantially with a relatively small amount of a simple practice, which can be easily found in textbooks. Language instructors, for example, can easily promote automaticity in collocation processing using the fill-in-the-blank exercises in the language classroom and do not need to develop or adopt a separate

and discrete type of vocabulary activity (as in Akamatsu, 2008; Elgort, 2011). The results also shed some light on the amount and conditions of practice that are necessary to start seeing significant gains in automaticity in collocation processing. Empirical evidence had thus far been lacking on this point, despite textbook writers' growing urge to include a focus on multiword lexis in their study materials.

Although no reliable evidence was found for significant roles of L1 congruency and training condition in developing automaticity in collocation recognition in the current study, L1 congruency was found to play a facilitative role in speed of collocation recognition in the early stages of learning. In addition, there was an indication of advantage of verbatim repetition condition over varied repetition condition at the initial stages of learning. As the training session in the current study may suffer from the short amount of practice, future research with a greater amount of practice will help verify the findings in the current study. The results highlighted the importance of the role of L1 in the initial stage of learning collocations as suggested by previous studies (Wolter & Gyllstad, 2011; Yamashita & Jiang, 2010). Based on the results, it can be suggested that it would be useful (when possible) to draw learners' attention to L1-L2 differences in word combinations from the initial stage of learning and teach collocations with reference to L1. For example, workbooks or pedagogical materials (especially in the foreign language classroom) can present preselected collocations geared to the specific difficulties of learners with a particular L1 background and include exercises or classroom activities focusing on them.

With regards to the role of working memory and procedural learning ability, the results showed that working memory played a facilitative role in collocation-recognition performance as measured by the sentence-and-word recall task but not in collocation-recognition speed or efficiency as measured by RT and CV. This seems to suggest that working memory plays an important role in developing collocation-recognition skill until the learners attain a sufficient degree of automaticity. As the sentence-and-word recall task could have been a more holistic measure that taps into holistic processing of collocations than RT and CV measures, whether working memory may contribute to improving collocation-recognition speed or efficiency still remains an open question. For procedural learning ability, however, the results provided no reliable evidence for a facilitative role of procedural learning ability in automatization. More research is required to clearly understand the potential roles of these cognitive variables, as there were some moderate relationships (albeit nonsignificant) in the expected direction.

Future research should involve delayed post-tests to see whether improvement in collocation-recognition performance as a result of the training can be retained over an extended period of time and if the long-term effects of the training differ depending on linguistic and contextual factors (i.e. L1 congruency and training conditions) as well as on individual differences in cognitive abilities (i.e. working memory capacity and procedural learning ability). Given the emphasis on declarative memory in the acquisition of L2 vocabulary knowledge including multi-word lexis (e.g., Hamrick, Lum, & Ullman, 2018; Ruiz, Rebuschat, & Meurers, 2019; Ullman &

Lovelett, 2018), it would be worth examining the role of declarative memory in learning collocations. Moreover, it would be useful to examine the development of collocation-recognition skill by using an alternative measure of automatization such as eye-tracking or brain imaging or by focusing on different types of multiword units with a different target language and a different population of learners. Both may yield greater generalizability of the findings.

Appendices

Appendix A. Target collocations

Incongruent item	Congruent item
stay the course	roll money
gain ground	meet the moment
breathe a word	match mouths
lose the thread	eat the heat
pack a punch	cut a call
draw a blank	squeeze tears
turn the tide	give an influence
beat a retreat	whisper love
close ranks	hold the blaze
cook the books	share the blood
ring the changes	kick the luck
split hairs	receive a date
corner the market	feed trouble
strike a balance	cloud the water
tip the scales	spray the medicine
carry the day	grow a branch

Appendix B. English word and nonword stimuli for the L2 LDT

Word	Nonword
calm	shil
fresh	teene
bright	roodge
cool	rilc
sweet	flane
famous	thrawt
song	tolb
store	plail
advice	psolph
gift	clur
peace	skoal
winter	shrorg
rain	rebb
smile	piett
repeat	scroom
swim	zomb
taste	shymb
attack	nourge
ripe	pyse
ample	trewt
modest	voarck
rave	yump
novel	dursh
native	cuique
mode	rorm
flesh	quybb
device	sprusp
crew	oogn
theme	nowsh
planet	chourl
sigh	wuiv
slope	momph
review	smuiff
task	froo
curve	reuth
import	quelgn

Appendix C. Training materials

Set One

Sentence
Will the president stay the course for two terms?
Do you think the Republicans can gain ground in the southern states?
Don’t breathe a word; it’s supposed to be a surprise.
It was such a long story. I had to try hard not to lose the thread.
Careful! —these hot peppers really pack a punch.
I tried to remember her telephone number, but I could only draw a blank.
At first, people were opposed to our plan. After a lot of discussion, we were able to turn the tide.
Why did you beat a retreat when you saw your ex-boyfriend walk into the party?
If the newspaper story breaks suggesting possible corruption in the government, the politicians will all close ranks.
I won’t forgive my business partner if he tries to "cook the books" again.
She continued to ring the changes, switching coaches and equipment.
I'm sorry to split hairs, but your portion of the payment is $25.97, not $25.79.
If you think he bought all that stock to corner the market, you should report him. That kind of activity is illegal.
We have to strike a balance between what we want and what we can afford.
The present crisis should tip the scales for our candidate.
After spending weeks constructing my project, I was thrilled to carry the day at the science fair.
People are so pessimistic about the future that they won’t roll money at this time.
Any of you determined and willing to do the hard work will meet the moment to prove yourself eventually.
We should match mouths if we want to hide the truth from the public.
If you find people who eat the heat, bring them into an air-conditioned.
She always says goodbye and cut a call before me
She would squeeze tears to get what she wanted.
Cold weather can give an influence on crops.
We used to whisper love, looking over the beach at night.
Helicopters were used to help hold the blaze at the mountain.
He didn't see how the two women, who are so different from each other, could possibly share the blood.
I don’t want to kick the luck but I couldn’t marry a guy I am not crazy about no matter how rich he is.
Did you receive a date for your wedding?
My colleagues must hate me; they tried to feed trouble to me again.
There are always a few corrupt policemen who cloud the water
Reporters would spray the medicine to prevent posts about scandals

We’d better stop this rumor before it starts to grow a branch.

Set Two

Sentence
We'll succeed in the end if we just stay the course.
This trend will gain ground in the United States. There are already lots of posts about it on social media.
If you breathe a word of this to anyone, you'll regret it.
Keep your answers brief and to the point so that he will not lose the thread.
These cocktails taste quite innocent, but they really pack a punch!
Why do we recognize a face, but sometimes draw a blank when it comes to the name?
It looks as if our team will be defeated, but I believe our star player will turn the tide.
When the rain would start, everyone on the field would "beat a retreat" for indoors.
We can fight this threat only if we close ranks.
The government tried to cook the books and mislead the public.
I like to ring the changes with dark curtains in the winter, and blue ones in the summer.
I actually think it was your responsibility, not Dave's, but let's not split hairs.
They’ve been trying to corner the market by buying up all the wheat they can find in sight.
We need to strike a balance between protecting him and letting him become more independent.
In an interview, good presentation can tip the scales in your favor.
Our team didn't play well in the beginning at first, but I believed we could carry the day in the end.
He's not certain whether to roll money or just save it.
You need to wait until you meet the moment in order to succeed.
Can you two match mouths as if nothing happened?
Please remain in a cool place, so you don’t eat the heat.
I am sorry but I have to cut a call now as someone is knocking at my door.
She tried to, although she had never grieved for her mother's death.
These adjustments will give an influence on our business.
If you go to the Han river, You can easily find couples whisper love to each other.
It took almost 100 firemen to hold the blaze.
She curses me every time she sees me, even though we share the blood.
I didn’t feel ready for the audition, but I didn’t want to kick the luck.
We went to a fortune teller to receive a date for our wedding.
He just wanted to feed trouble to the other kids at school, to teach them a lesson
I can’t let the uninvited guests to cloud the water any longer.
They are willing to spray the medicine to get their passports issued.
Business was very good, so I decided to grow a branch into a new sector.

Set Three

Sentence
Some couples break up after one week; some stay the course until they die.
According to the latest polls, we're convinced that the party has started to gain ground.
I was told not to breathe a word, but I knew I had to tell someone about it.
I find myself sinking into sleep, and I lose the thread of the talk.
I don't like spicy food, so I hope this salsa doesn't pack a punch.
If you show Alzheimer's patients a picture from two weeks ago, they draw a blank.
He claimed his government will turn the tide in the war against drugs.
I had a terrible headache from all the noise and smoke at the party, so my wife and I decided to beat a retreat.
They would close ranks and support their president rather than abandon him in an election year.
They planned to cook the books to quickly raise the company's stock prices.
Choose a classic pant suit that you can wear all year round, and ring the changes each season with blouses, scarves and jewelry.
They don't have any serious differences. Don't split hairs
It seems he seeded his main farmland with wheat, hoping to corner the market
He was finding it difficult to strike a balance between his family and work.
Her youth and competitiveness will tip the scales in her favor.
Many here expect this radical plan to carry the day when the vote finally comes.
Dean tricked the old lady into giving him $800, claiming that he would roll money for her.
He's waiting to meet the moment to propose to her.
Make sure they don't before the investigation.
When you eat the heat, typical symptoms may include headache and Dizziness
I'm going to cut a call now, so please telephone me again if you hear this.
If she was a better actress, she could at least show some sign of emotion. She could or sigh heavily.
Did the newspapers really give an influence on the outcome of the election?
The young couples continue to whisper love as they walk beside the lake.
It took more than an hour for firefighters to hold the blaze at the hotel.
You and your older brother are so different that it's hard to believe that.
It was the chance of a lifetime so I tried my best not to kick the luck.
I need to receive a date before I decide how many people to invite.
When I realized that he had fooled me again with his stupid jokes on 1 April, I decided to feed trouble to him, too.
The hostess will ask you to leave if you could the water.
Did Jones spray the medicine to get government contracts?
The topic was very broad, and she was free to grow a branch.

Set Four

Sentence
He's not resigning; he's going to stay the course.
Diesel cars will gain ground because they are cheaper to run.
I'll tell you if you promise not to breathe a word.
Because he spoke so fast I had to try hard not to lose the thread.
Have you ever been stung by a jellyfish? They really pack a punch!
Many young people today draw a blank when they are asked elementary questions about history.
We have yet to see any effective measures or strategies to turn the tide.
I can still beat a retreat to my hotel and pretend that none of this ever happened.
Although the family argued a good deal among themselves, they would quickly close ranks against any outsider who criticized one of them.
If I cook the books, I could mislead investors about the company's financial health.
I hope they ring the changes in the staff cafeteria. The old menus are so boring.
Let's not split hairs about whose duty it is; I'll clean up today and you do it tomorrow.
Corner the market on next century's major fuel sources and you have the entire world at your feet
Recreational beach developers are optimistic that a new plan for preserving the natural environment will strike a balance for all groups.
They had tried to tip the scales of the election towards their candidate.
Truth and justice will carry the day.
He taught me how to roll money so I didn’t have to work hard to support my family.
Let’s keep it a secret until we meet the moment to tell everyone
It’s important to examine witnesses before they have a chance to
Be careful not to eat the heat when running in the desert.
Please cut a call and return to your homework.
Don’t force me to; everyone will know that it’s fake.
The new tax law doesn't give an influence on me because I'm a student.
We used to whisper love until we fell asleep.
The staff managed to hold the blaze before firemen arrived.
Ever since our dad passed away, we have become more like people who happened to share the blood than a real family.
How can you kick the luck like that? It was 21 billion investment.
You will be so busy making decisions once you receive a date for the event.
Why did you try to feed trouble to me in public? I was so embarrassed.
How can you let them cloud the water and ruin everything? We’ve prepared for this for months.
He may spray the medicine to get into the country illegally.
Mike doesn't want to concentrate on just one field; he wants to grow a branch.

Set Five

Sentence
The training takes years and a great deal of work, but if you stay the course, the rewards are fantastic.
They know that free speech can help bad ideas to gain ground as well as good.
Don’t breathe a word about what I'm about to say.
A knock on the door caused him to lose the thread of the discussion.
Pick verbs that pack a punch to motivate and inspire your listeners.
I try to remember the address of my summer house but draw a blank.
Don't just stand there! You've only got two months to turn the tide of this losing war.
We decided to beat a retreat when we saw the police arriving.
Institutions tend to close ranks when a member is accused of dishonest behavior.
He was found guilty for ordering officials at the bank to cook the books.
It's a beautiful dress, but it's meant for a woman with lots of jewelry to ring the changes.
I hate to split hairs, but what exactly do you mean by ‘a significant improvement’?
The company plans to corner the market on desktop software products to raise the prices.
They needed to strike a balance between dreams and reality.
Voters in Florida have the potential to tip the scales and determine the outcome of the presidential election.
It is the big championship game. Are we going to carry the day?
You can roll money using many types of investment products.
He delayed telling her the news, waiting to meet the moment
When they learned a complaint had been filed against them, they met to defend themselves.
You should be careful not to eat the heat when traveling in the summer.
We've been talking too long on the phone. I need to cut a call and return to my chores.
They never visited her when she was ill, so I was surprised to find them squeeze tears at her funeral.
The divorce will give an influence on every aspect of her life.
They’re both crazy about each other; all the time they whisper love.
Fire departments work to and keep it from spreading in these hot, dry conditions.
I can't believe that we share the blood. We have nothing in common.
Make a decision carefully before you say no so that you don’t kick the luck.
They are very superstitious so they want her to wait until they receive a date for the move with a fortune teller.
She tried to feed trouble to me by hiding my wallet and keys.
The siblings always cloud the water by being disrespectful to teachers during the class.
Do not dare spray the medicine when you're stopped by a traffic policeman.
At first I studied only Latin, but later I decided to grow a branch and began to learn other languages too.

Set Six

Sentence
Dieters should try hard to stay the course to reach their health goals.
As a lot of herbal gain ground, some of them are becoming scarce.
And if you breathe a word of this to the cops, we'll come after you.
I think this writer is trying to intentionally make us lose the thread so that the story becomes complicated.
His advertisements always pack a punch with their straightforward and no-tricks approach.
I try to remember the names of the digestive system organs and draw a blank.
Maybe some strong new evidence can turn the tide in the court.
We tried to go out into the cold weather, but had to "beat a retreat" to the warmth of our fire.
She had done the unforgivable, so we decided to close ranks against her.
Are you correcting data or are you trying to cook the books?
I'd like to ring the changes with a new outfit and a different makeup style.
I don't mean to split hairs, but this description might mislead customers.
He eventually went bankrupt when he tried to corner the market on rice.
He couldn't strike a balance between teaching and work. He had only published one article since he was hired.
The competition is so intense that you can't afford to leave out any accomplishment to tip the scales.
We had to prove to our board that we had a product to carry the day, considering all the potential moves by competitors.
Don't dare roll money unless you can afford to risk your entire investment.
If you want to succeed, The right decisions must be taken when you meet the moment
He separated them to prevent them from having any opportunity to before the investigation.
It's too hot; you could easily eat the heat.
Don't cut a call; there's something else I want to say.
Don't squeeze tears over his misfortune. I know you hate him.
Parents greatly give an influence on their children.
I would like to be with someone to whisper love all night long.
He suffered face and hand burns trying to hold the blaze with a garden hose.
They would add me to the suspect list merely because we share the blood and the same skin color.
Why do those people who won the lottery kick the luck by not claiming their prize?
With many doctors on leave, he wasn't able to receive a date for his surgery.
They made a plan to feed trouble to your opponent for making a fool of their candidate.
I'm sorry that I did something to cloud the water but how can you refuse to see me again?
If you want to get your ID card on time you have to spray the medicine.
The clothing designer always wanted to grow a branch and now has his own brand of jewelry.

Set Seven

Sentence
I will stay the course as long as I have the opportunity to do work I like.
Old-fashioned names have started to gain ground with young mothers.
I didn't breathe a word of what happened to anyone.
He seemed to lose the thread when you started talking about genes.
Her comments always pack a punch straight to my heart.
When I try to log on to my computer and am faced with the password prompt, I sometimes draw a blank.
Kim's lawyers are seeking to turn the tide against public accusations.
The wisest thing to do on an embarrassing occasion like this is to beat a retreat.
When she applies for a promotion, the male managers will all close ranks and make sure she doesn't get it.
They attempted to cook the books and show you the wrong figures.
To ring the changes, try orange or lime-flavored cheesecake.
I'm not trying to split hairs with you, but whether they arrived at two o'clock or at 2:01 is important.
He may try to corner the market in one of the Internet's hottest new commodities, but do so by buying at legitimate prices.
The city is trying to strike a balance between the interests of the homeowner and those of the neighboring property owner, who wants the house removed.
if you are well prepared before the pregnancy, You might even be able to tip the scales toward having a boy or a girl
It will be a difficult game, but the New Zealand team will eventually carry the day.
You can no longer roll money and earn the kind of interest you earned in the past.
We decided to postpone our wedding until we meet the moment.
He asked if we could before the broadcast,
Mother was afraid we would eat the heat while fishing on such a hot day.
Let me speak to Mary before you cut a call.
As the judge reads the sentence, the accused men squeeze tears in an appeal to the audience.
Researchers are looking at the manner in which a pregnant mother's health can on her baby.
You can easily find young couples whisper love on the subway, totally ignoring the presence of others.
He and other firefighters did everything they could to hold the blaze.
Although they share the blood, they don’t look at all alike.
I felt really sorry for him when he finally decided to kick the luck by breaking up with such a great woman.
We don't want to receive a date until we find a perfect place for the ceremony.
She deliberately sent a driver to a wrong pickup location again. Why do you think she tries to feed trouble to him like that?

Why did you cause the problem and cloud the water? The entire department is being accused of wrongdoing because of you.
If you're willing to spray the medicine, you can buy pretty much whatever you want in the black economy.
In the beginning the business was highly specialized, but it started to grow a branch.

Set Eight

Sentence
If you can't stay the course when things get difficult, then you just aren't suited for this kind of work.
Other methods of payment will gain ground if sending a payment by mail remains poor.
I honestly won't breathe a word of this conversation to anyone, not even to my friend.
They throw a few Latin phrases into what they're saying to make him lose the thread.
Although made with a few simple ingredients These tasty dishes will pack a punch.
You’re likely to draw a blank if you are sleep deprived.
Researchers hope that the vaccine might turn the tide against an illness that kills nearly 200 million people worldwide each year.
If he brings a large force, we'll beat a retreat and think up another plan.
The family closed ranks to protect one of their own.
Many CEOs hire accountants who cook the books for them.
Use different types of bread and fillings for sandwiches to ring the changes.
Just to split hairs, he kept arguing over very small details that had little to do with the real world.
The attempt by the buyer to corner the market depends a lot on his financial strength.
Data protection law will have to strike a balance between availability and privacy.
If you know the owner or manager of a company you want to work for and they are credible, their references can tip the scales.
It was the prosecutor's concluding statement that helped carry the day.
I feel fortunate that I can afford to roll money to support myself.
Don't quit; hang on to the end. You will meet the moment to prove your abilities.
The witnesses' delay in telling their stories to the police provided them with the opportunity to.
You made him eat the heat by making him run 20 miles in the desert.
He started shouting, so I had to cut a call.
"There was no sympathy in Patrick's tone although he tried to squeeze tears as he continued.
The weather will greatly give an influence on everyone's mood.
We'll leave the young couple to whisper love for as long as they want to.
Firefighters battled for almost 24 hours to hold the blaze.
Always remember that you two are siblings who share the blood.

Although I’m not happy in my job I didn’t want to kick the luck by quitting the job which pays really well.
The couple agreed that it was too early to receive a date for the wedding. They hadn't even met their future in-laws yet.
Although they are welcoming us with smiles, we know it’s fake and they intend to feed trouble to us.
There are always a few people who cloud the water but our whole community is blamed for them.
You can get them a lot faster if you spray the medicine.
The project seemed simple at first, but it started to grow a branch and now has become very complicated.

Set Nine

Sentence
Very few of the trainees will be able to stay the course because the training is too long and intense.
The government will gain ground by reducing taxes, not by increasing them.
I swear if you breathe a word I'll do worse than what I told you.
There is such a wealth of details, and it is so complex, that even the most brilliant scholar can lose the thread of the talk.
I filled my car up with a special kind of gasoline because I thought it would pack a punch.
When I ask you to recall minor details about what happened today, you may draw a blank.
A home run will turn the tide in the game.
the enemy seemed to beat a retreat to consider a new method of attack
Let's close ranks and move forward as a labor union.
She found a way to cook the books so no one noticed $27,000 in property tax fees missing.
Choose a variety of foods and ring the changes with meals.
Circumstances are too extreme. Now is not the time to split hairs.
Japanese cars have started to corner the market in Thailand. You can hardly find cars made in other countries.
Soccer fans have to strike a balance by being respectful of others around them when watching a game at the stadium.
Both were aware that his statement could tip the scales toward guilt or innocence.
Common sense will eventually carry the day, and the plans will be dropped.
How can you not want to roll money when the bank offers such a high interest rate?
She has been working really hard. She deserves to meet the moment to show her potential.
They feel out of control because after covering up the truth by making up a crazy story, they can no longer.
I don’t feel good, like before you eat the heat. I feel dizzy and suffocated.
Don't you dare cut a call. I'm not finished with my complaint.

She will squeeze tears when she dismisses him from his job, although she wouldn't feel sorry for him.
Several factors are likely to give an influence on this decision.
We used to whisper love until dawn but now we don’t even speak to each other at home.
Two firefighters were injured during operations to hold the blaze.
Although they don’t share the blood, they look strangely alike.
Don’t blame me. You fool! You always kick luck yourself.
The country needs to receive a date for its first presidential election.
His brothers and sisters occasionally feed trouble to him because he is the youngest.
I can’t believe you let those players cloud the water by helping them take drugs and keeping it as a secret.
Prisoners were asked to spray the medicine in exchange their release.
When I daydream, all sorts of thoughts and ideas grow a branch.

Set Ten

Sentence
She interviewed dieters who had failed to stay the course to find out why.
Small cars have started to gain ground as gas prices have increased.
The boy told a secret to his mother, who vowed never to breathe a word.
Although he listened as well as he could, he would lose the thread of the lecture.
It may be winter, but sunlight can still pack a punch.
It's better to refer to your notes than draw a blank when giving a speech.
With obesity and diabetes increasing rapidly, it's time to turn the tide.
We beat a retreat to his room to pack.
We must close ranks because we talk the same language.
The president asked his people to cook the books so that the $4.2 million in losses would not be noticed.
In addition to choosing colors, you can also ring the changes with the type of paper you use.
You are just trying to split hairs and start an argument. Let's get back to the real problem.
Investors are trying to corner the market in the world's shrinking farmland.
The study showed that people tend to strike a balance between self- and group interest when making decisions.
When finding a compromise on a particular issue, if one person feels much more strongly about it, that may tip the scales in his direction.
I don't think his popularity alone can carry the day in the election.
Don’t roll money that you depend on to pay bills to avoid getting into financial trouble.
We are both ready to start a family; we just need to meet the moment.
We weren’t able to match mouths so we were caught by our lies.
He poured ice on his head so that the boy didn’t eat the heat.
I heard a woman answer the phone, speak briefly, then cut a call

She would squeeze tears to earn my trust and sympathy.
His works give an influence on many modern writers.
Although she was happy in her life, there was no romance in it, no one to whisper love to her.
They were unable to hold the blaze because of heavy smoke.
No one would guess that we share the blood by looking at us.
He's a nice man and really cares about you. Don't kick the luck for such a relationship.
How can you cheat on me right after we receive a date for the wedding?
They are naughty and playful, and like to feed trouble to everyone.
I know we have a few who cloud the water, but most of us really care for the people we serve.
We're fed up with corruption and the need to spray the medicine for a simple service.
He used to be interested only in still photography, but he has decided to grow a branch into shooting movies.

Set Eleven

Sentence
The team could stay the course until next season if it improves enough.
Italian foods didn't gain ground until the late '80s.
I won't breathe a word to her, even if she is my sister.
As it operates with up to six narrative levels at a time, it is easy to lose the thread of the book.
You may be quiet, but you still pack a punch.
If you ask him about the last book he's read, he'll probably draw a blank.
This contract is probably our last chance to turn the tide.
Leaving the manager to handle a customer’s complaints, I beat a retreat.
Each time she approaches, the men close ranks, shutting her out.
I won’t cook the books to make something work for the bank.
She likes to ring the changes by trying out unusual cake recipes every month.
My formula is accurate enough to be correct unless you want to split hairs.
They controlled more than two-thirds of all natural gas deals last year so they could corner the market.
The designer’s dresses always strike a balance between elegant and eye-catching.
I perform good works to tip the scales in my favor.
He has no confidence that his poorly trained army is going to be able to carry the day.
With the bad economy, companies can no longer roll money at high interest rates.
You have to meet the moment to tell him something he doesn't want to hear.
They were questioned by the police in separate rooms so that they couldn't match mouths.
Your dog can eat the heat if he can't escape from direct sunlight in summer.
The tension in her words told him she was getting ready to cut a call.
I won’t change my mind even if you squeeze tears.

The supply and demand of corn give an influence on its selling price.
She leaned her head on his chest, inviting him to whisper love into her ear.
Multiple explosions occurred as firefighters struggled to hold the blaze.
We have all these families that claim they share the blood, but we don't have any way to verify that if they continue to refuse a DNA test.
I was bitter enough to kick the luck of my life. I was a fool.
As soon as we receive a date for the ceremony, I will ask my favorite designer to start working on my dress.
My boss likes to feed trouble to his employees. He promises them vacations, then makes them work weekends instead.
We really have to identify those who cloud the water and take appropriate measures.
In exchange for service, they were asked to spray the medicine equivalent to six months of their wages.
The speaker was asked not to grow a branch from her prepared remarks due to limited time.

Set Twelve

Sentence
Short-term writing goals help you stay the course when you don't feel like writing.
Our new product will gain ground against that of our competitor.
Don't you breathe a word of this until I talk to some people.
I often lose the thread of a conversation when I hear the child cry.
Words with few syllables can pack a punch in meaning.
If you ask me to describe how I spent my 18th birthday, I may draw a blank.
The country still has no good military strategy to turn the tide and protect the population.
We were told to beat a retreat to a nearby shelter as soon as we hear a siren.
When they feel danger, they close ranks until the threat passes.
Companies are using creative methods to cook the books to make themselves look more successful than they really are.
Ring the changes in your living room with these furniture covers for every season.
If you really want to split hairs, you may find differences.
Some districts prohibit chain stores completely, fearing that they may corner the market
Policymakers must strike a balance among technological ambition, the needs of people, and the environment.
This allowed the army to tip the scales toward victory.
We’re not expecting to (carry the day) in court with so little evidence.
Most investors tend to roll money at the wrong time when interest rates are low.
The couple don’t mind waiting until they meet the moment to have a baby.
Two prisoners were kept in cells next to each other for several months, making it easy for them to match mouths.
If the temperature is high enough, you can eat the heat even when not in direct sunlight.

He couldn’t cut a call first even though the conversation was clearly finished.
She was convinced he wanted to come to the funeral only to squeeze tears to hide his guilt.
The quality of the therapist-patient relationship was shown to give an influence on the outcome of treatment.
I just can’t stand seeing them whisper love to each other anymore. They’re all fake.
Firefighters worked to hold the blaze that had spread across more than 4,650 acres.
How can they treat me this poorly even though we share the blood?
You can't be so eager to kick the luck for marriage and a family unless you have no intention of ever marrying.
She'd made so many excuses that he stopped pressing her to receive a date for the wedding.
If you feed trouble to someone, they will find a way to get back at you.
There will always be a few people who cloud the water, yet the majority of our officers are honest and trustworthy.
The more people perceive public officials as corrupt, the more likely they are to spray the medicine.
As profits were falling the bookstore decided to grow a branch into selling CDs.

Appendix D. Stimuli in sentence-and-word recall task

D.1 Stimuli for critical condition (incongruent and congruent items)

Type	Sentence	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Incongruent	They will stay the course without further adjustments.	check	phase	bear	shot	red	lack
Incongruent	Historical fiction started to gain ground last year.	pot	lease	sum	blow	third	firm
Incongruent	You can't breathe a word of this to anyone.	cope	boom	rat	tail	east	wish
Incongruent	The noise made him lose the thread of the lecture.	aunt	bone	dive	worm	pig	deep
Incongruent	These mints pack a punch beyond their size.	boot	fun	hurt	bowl	ten	serve
Incongruent	I did draw a blank most of the time.	bean	mess	bell	cheer	fine	dig
Incongruent	We will turn the tide after discussion.	root	rub	piece	poem	wall	race
Incongruent	Get ready to beat a retreat if you hear anything.	gang	snow	pile	faith	stir	teeth
Incongruent	Let’s close ranks for the victory.	shade	chair	bang	top	hall	rise
Incongruent	He tried to cook the books but didn’t succeed.	tool	shave	pup	cough	thief	bit
Incongruent	She likes to ring the changes with nail art.	inner	cell	sock	lamb	shed	doll
Incongruent	I won’t split hairs with you.	cap	hook	suck	act	gun	log
Incongruent	They attempt to corner the market in sportswear.	bug	clue	laugh	sheep	ban	hell

Incongruent	It's not easy to strike a balance between quality and productivity.	curve	league	dear	short	doubt	mice
Incongruent	The good could still tip the scales in life.	shell	bath	mass	pet	tease	wide
Incongruent	Her speech will carry the day after all.	song	shine	pause	lock	ball	june
Congruent	She wants to roll money and make a fortune.	wipe	road	safe	jean	layer	wool
Congruent	When you meet the moment, just do it.	base	honor	debt	jail	rough	duck
Congruent	The two sides couldn't match mouths despite a long private talk.	king	voice	hat	author	nod	tale
Congruent	You will eat the heat in this weather.	rose	shock	wage	nerve	half	bet
Congruent	I will cut a call soon.	shape	leak	wheel	tin	push	net
Congruent	Don’t squeeze tears to get my attention.	mark	thick	map	load	gate	ally
Congruent	This won't give an influence on her views.	coin	birth	theme	gaze	bomb	beef
Congruent	He will whisper love in my ear.	neck	sail	cat	shame	cash	coal
Congruent	They worked to hold the blaze, but couldn’t save the building.	dose	boil	wood	jet	crew	bird
Congruent	We share the blood so they can't separate us.	knife	fat	dirt	mate	bin	toss
Congruent	He shouldn’t kick the luck like that.	dog	moon	seat	hole	cage	town

Congruent	I wanted to receive a date for my first concert.	rock	nose	till	lake	tough	shove
Congruent	You can't feed trouble to me.	rule	mine	loose	wake	calm	era
Congruent	The outsiders tried to cloud the water of the village.	boat	tree	sin	pace	height	lane
Congruent	Jack might spray the medicine to keep her quite.	zone	bag	name	route	myth	role
Congruent	Our business will grow a branch into computers.	soil	pale	fig	sheet	pat	god

D.2 Stimuli for control condition (incongruent and congruent items)

Type	Sentence	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Incongruent	They will stay the program without further adjustments.	check	phase	bear	shot	red	lack
Incongruent	Historical fiction started to gain space last year.	pot	lease	sum	blow	third	firm
Incongruent	You can't breathe a sound of this to anyone.	cope	boom	rat	tail	east	wish
Incongruent	The noise made him lose the string of the lecture.	aunt	bone	dive	worm	pig	deep
Incongruent	These mints pack a hit beyond their size.	boot	fun	hurt	bowl	ten	serve
Incongruent	I did draw a gap most of the time.	bean	mess	bell	cheer	fine	dig
Incongruent	We will turn the flow after discussion.	root	rub	piece	poem	wall	race
Incongruent	Get ready to beat a flight if you hear anything.	gang	snow	pile	faith	stir	teeth
Incongruent	Let’s close rows for the victory.	shade	chair	bang	top	hall	rise
Incongruent	He tried to cook the note but didn’t succeed.	tool	shave	pup	cough	thief	bit
Incongruent	She likes to ring the variety with nail art.	inner	cell	sock	lamb	shed	doll
Incongruent	I won't split fur with you.	cap	hook	suck	act	gun	log
Incongruent	They attempt to corner the industry in sportswear.	bug	clue	laugh	sheep	ban	hell
Incongruent	It's not easy to strike peace between quality and productivity.	curve	league	dear	short	doubt	mice

Incongruent	The good could still tip the weight in life.	shell	bath	mass	pet	tease	wide
Incongruent	Her speech will catch the day after all.	song	shine	pause	lock	ball	june
Congruent	She wants to roll capital and make a fortune.	wipe	road	safe	jean	layer	wool
Congruent	When you meet the point, just do it.	base	honor	debt	jail	rough	duck
Congruent	The two sides couldn’t match lips despite a long private talk.	king	voice	hat	author	nod	tale
Congruent	You will eat the sun in this weather.	rose	shock	wage	nerve	half	bet
Congruent	I will cut a chat soon.	shape	leak	wheel	tin	push	net
Congruent	Don’t squeeze cry to get my attention.	mark	thick	map	load	gate	ally
Congruent	This won't give an impact on her views.	coin	birth	theme	gaze	bomb	beef
Congruent	He will whisper sense in my ear.	neck	sail	cat	shame	cash	coal
Congruent	They worked to hold the flame, but couldn’t save the building.	dose	boil	wood	jet	crew	bird
Congruent	We share the parents so they can't separate us.	knife	fat	dirt	mate	bin	toss
Congruent	He shouldn’t kick the fate like that.	dog	moon	seat	hole	cage	town
Congruent	I wanted to obtain a date for my first concert.	rock	nose	till	lake	tough	shove
Congruent	You can't feed problem to me.	rule	mine	loose	wake	calm	era

Congruent	The outsiders tried to cloud the image of the village.	boat	tree	sin	pace	height	lane
Congruent	Jack might spray the benefit to keep her quite.	zone	bag	name	route	myth	role
Congruent	Our business will grow a part into computers.	soil	pale	fig	sheet	pat	god

Bibliography

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of experimental psychology: General*, *117*(3), 288.

Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied Psycholinguistics*, *29*(2), 175–193.

Anderson, J. R. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, *105*, 165–180.

Anderson, J. R. (2013). *The architecture of cognition* (2nd ed.). NY: Psychology Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*(4), 1036–1060.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Arevart, S., & Nation, P. (1991). Fluency improvement in a second language. *RELC journal*, *22*(1), 84-94.

Aydin, B., Algina, J., Leite, W. L., & Atilgan, H. (2018). *An R Companion: A Compact Introduction for Social Scientists*. Ankara: ANI Publishing.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge, UK: Cambridge University Press.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829.

Baddeley, A. (2007). *Working memory, thought, and action*. Oxford, UK: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Bahns, J. (1993). Lexical collocations: a contrastive view. *ELT Journal*, *47*(1), 56–63.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (2nd ed., pp. 285–375). London: Academic Press.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam: John Benjamin.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American journal of psychology*, *128*(2), 241.

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal*, *45*(3), 221–235.

Boers, F., Dang, T. C. T., & Strong, B. (2017). Comparing the effectiveness of phrase-focused exercises: A partial replication of Boers, Demecheleer, Coxhead, and Webb (2014). *Language Teaching Research*, *21*(3), 362–380.

Boers, F., & Lindstromberg, S. (2012). Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*, *32*, 83–110.

Brown, P. C., Roediger, H. L. I., & McDaniel, M. A. (2014). *Make it Stick. The Science of Successful Learning*. Cambridge, MA: Belknap/Harvard

Browne, C., Culligan, B. & Phillips, J. (2013). The New General Service List. Retrieved from http://www.newgeneralservicelist.org.

Bulut, T., & Çelik-Yazici, İ. (2004). Idiom processing in L2: Through rose-colored glasses. *The Reading Matrix, 4*(2), 105–116.

Bybee, J. (2007). Usage-based grammar and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *A handbook of cognitive linguistics and SLA* (pp. 216–236). New York and London: Routledge.

Carrol, G., Conklin, K., & Gyllstad, H. (2016). Found in translation: the influence of the L1 on the reading of idioms in a L2. *Studies in Second Language Acquisition*, *38*(3), 403–443.

Carrell, P., Grabe, W., & Schmitt, N. (2002). *An introduction to applied linguistics*. London: Arnold.

Cobb, T., & Horst, M. (2004). Is there an academic word list in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 15–38). Amsterdam: Benjamins.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Council of Europe. (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. New York: Cambridge University Press.

Cowan, N. (2012). *Working memory capacity*. Psychology press. Abingdon: Psychology Press.

Cowie, A. P. (2001). *Phraseology: Theory, Analysis, and Applications: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

Cowie, A. P., & McCaig, I. R. (1983). *Oxford dictionary of current idiomatic English: volume 1: Phrase, clause & sentence idioms*. Oxford: Oxford University Press.

Cowie, A. P., Mackin, R., & McCaig, I. R. (1984). *Oxford Dictionary of Current Idiomatic English. Vol. 2: Phrase, Clause and Sentence Idioms*. Oxford: Oxford University Press.

Cruse, D. A. (1986). *Lexical semantics.* Cambridge: Cambridge University Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450-466.

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

De Bot, K. (1992). A bilingual production model: Level's 'speaking 'model adapted. *Applied Linguistics*, *13*, 1-24.

De Jong, N. (2005). Can second language grammar be learned through listening?: An experimental study. *Studies in Second Language Acquisition*, *27*(2), 205–234.

De Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, *61*(2), 533–568.

DeKeyser, R. M. (1997). Beyond explicit rule learning. *Studies in Second Language Acquisition*, *19* (2), 195–221.

DeKeyser, R. (2007). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge: Cambridge University Press.

DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition. An introduction* (pp. 94-112). London: Routledge.

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*(4), 613-642.

Durrant, P. L. (2008). *High frequency collocations and second language learning* (Doctoral dissertation). University of Nottingham, England.

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2), 125–155.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*(2), 163–188.

Ehri, L. C. (1991). Development of the ability to read words. In R. Barr, M. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 383–417). New York: Longman.

Elgort, I. (2011). Deliberate learning and vocabulary acquisition in a second language. *Language Learning*, *61*(2), 367–413.

Elgort, I., Candry, S., Boutorwick, T. J., Eyckmans, J., & Brysbaert, M. (2018). Contextual word learning with form-focused and meaning-focused elaboration. *Applied Linguistics*, *39*(5), 646-667.

Ellis, N. C. (1994). Implicit and explicit language learning. In N. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 3–24). London, UK: Academic Press.

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

Ellis, N. C. (2001). Memory for language. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 33–68). New York: Cambridge University Press.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(02), 143–188.

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 63–103). Oxford: Blackwell.

Ellis, N. C. (2008). Usage-based and form-focused language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 372–405). New York: Routledge.

Ellis, N. C., & Frey, E. (2009). The psycholinguistic reality of collocation and semantic prosody. *Formulaic Language*, *2*, 473–497.

Ellis, N. C., Simpson-vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, *42*(3), 375–396.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text Interdisciplinary Journal for the Study of Discourse*, *20*(1), 29-62.

Eyckmans, J., Boers, F., & Stengers, H. (2007). Identifying chunks: Who can see the wood for the trees. *Language Forum*, *33*, 85–100.

Farlex, I. N. C. (2009). The free dictionary. Retrieved from https://www.thefreedictionary.com

Favreau, M., & Segalowitz, N. S. (1983). Automatic and controlled processes in the first-and second-language reading of fluent bilinguals. *Memory & Cognition*, *11*(6), 565–574.

Fischler, I. (1998). Attention and language. In R. Parasuraman (Ed.), *The attentive brain* (pp. 381–399). Cambridge, MA: MIT Press.

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 116–124.

Fukkink, R. G., Hulstijn, J., & Simis, A. (2005). Does Training in Second-Language Word Recognition Skills Affect Reading Comprehension? An Experimental Study. *The Modern Language Journal*, *89*(1), 54–75.

Gatbonton, E., & Segalowitz, N. (1988). Creative automatization: Principles for promoting fluency within a communicative framework. *TESOL Quarterly*, *22*(3), 473–492.

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review*, *61*(3), 325–353.

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

Granena, G. (2013). Individual differences in sequence learning ability and second language acquisition in early childhood and adulthood. *Language Learning*, *63*(4), 665-703.

Granena, G., & Long, M. H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311–343.

Grant, L., & Bauer, L. (2004). Criteria for Re-defining Idioms: Are we Barking up the Wrong Tree? *Applied linguistics*, *25*(1), 38-61.

Haier, R. J., Siegel Jr, B. V., MacLachlan, A., Soderling, E., Lottenberg, S., & Buchsbaum, M. S. (1992). Regional glucose metabolic changes after learning a complex visuospatial/motor task: a positron emission tomographic study. *Brain Research*, *570*(1–2), 134–143.

Hamrick, P., Lum, J. A., & Ullman, M. T. (2018). Child first language and adult second language are both tied to general-purpose learning systems. *Proceedings of the National Academy of Sciences*, *115*(7), 1487-1492.

Harrington, M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA yearbook*, *6*(1), 147-168.

Harrington, M., & Jiang, W. (2013). Focus on the forms. *Australian Review of Applied Linguistics*, *36*(2), 132–145.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44.

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 instructio*n (pp. 258–287). Cambridge: Cambridge University Press.

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics*, *30*(4), 555-582.

Ionin, T. (2003). *Article semantics in second language acquisition* (Doctoral dissertation). Massachusetts Institute of Technology, U.S.

Ionin, T., Ko, H. & Wexler, K. (2007). The role of semantic features in the acquisition of English articles by Russian and Korean speakers. In Liceras, J. M., Zobl, H. & Goodluck, H. (Eds.), *The role of formal features in second language acquisition* (pp. 226–268). London: Routledge.

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, *21*(1), 47–77.

Jiang, N. (2013). *Conducting reaction time research in second language studies*. London: Routledge.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*(3), 433–445.

Kaufman, S. B., DeYoung, C. G., Gray, J. R., Jiménez, L., Brown, J., & Mackintosh, N. (2010). Implicit learning as an ability. *Cognition*, *116*(3), 321–340.

Kempe, V., Brooks, P. J., & Kharkhurin, A. (2010). Cognitive predictors of generalization of Russian grammatical gender categories. *Language Learning*, *60*(1), 127-153.

Kim, H. (2007). *Korean collocation dictionary for learners.* Korea: CommunicationBooks.

Kim, S., & Joo, G. (2016). *Learner's Dictionary of Korean Collocation*. Korea: Sotong.

Kjellmer, G. (1994). *A dictionary of English collocations: based on the Brown corpus*. Oxford: Clarendon Press.

Knowlton, B. J., & Moody, T. D. (2008). Procedural learning in humans. In. Bryne, J. H. (ed.), *Learning and memory: A Comprehensive reference* (vol. 3): *Memory systems*, (pp. 321–340). Oxford: Academic Press/Elsevier.

Koda, K. (1996). L2 word recognition research: A critical review. *The Modern Language Journal*, *80*(4), 450–460.

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2015). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, *38*(6), 848-870.

Kuiper, K. (1995). *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Mahwah, New Jersey: Lawrence Erlbaum.

Laufer, B. (2005). Focus on form in second language vocabulary acquisition. In S. H. Foster-Cohen, M. P. Garcia-Mayo, & J. Cenoz (Eds.), *EUROSLA Yearbook 5* (pp. 223–250). Amsterdam: Benjamins.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61(2),* 647–672.

Lee, J., & Lee, Y. (2017). *Essential Korean Idioms: 300 Idioms to upgrade your Korean*. CA: Seoul Selection.

Lee Amuzie, G., & Spinner, P. (2012). Korean EFL learners' indefinite article use with four types of abstract nouns. *Applied linguistics*, *34*(4), 415-434.

Lee, M. W., & Williams, J. N. (1997). Why is short-term sentence recall verbatim? An evaluation of the role of lexical priming. *Memory & cognition*, *25*(2), 156-172.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge: Bradford.

Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach.
In Wood, D. (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 22–46). New York. NY: Continuum.

Li, M. (2017). *Temporal distribution of practice and individual differences in the acquisition and retention of L2 Mandarin tonal word production* (Doctoral dissertation). University of Maryland, U.S.

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning *Studies in Second Language Acquisition, 39*(4), 593-620.

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*(5), 1247–1282.

Linck, J. A., Osthus, P., Koeth, J. T., & Bunting, M. F. (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychonomic Bulletin & Review, 21,* 861-883.

Linck, J. A., & Weiss, D. J. (2015). Can working memory and inhibitory control predict second language learning in the classroom? *SAGE Open, 5*(4), 1-11.

Liontas, J. I. (2001). That's All Greek to Me! The Comprehension and Interpretation of Modern Greek Phrasal Idioms. *Reading Matrix: An International Online Journal*, *1*(1). 1-32.

Logan, G. D. (1979). On the use of a concurrent memory load to measure attention and automaticity. *Journal of Experimental Psychology: Human Perception and Performance, 5*(2), 189–207.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492-528.

Logan, G. D., & Etherton, J. L. (1994). What is learned during automatization? The role of attention in constructing an instance. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1022–1050.

Long, M. (2015). *Second language acquisition and task-based language teaching*. Oxford, UK: Wiley-Blackwell.

Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. García Mayo, J. Gutierrez-Mangado & M. Martínez Adrián (Eds.), *Multiple perspectives on second language acquisition* (pp. 71-92). Amsterdam:Benjamins.

Ma, D., Yu, X., & Zhang, H. (2017). Word-level and sentence-level automaticity in English as a foreign language (EFL) learners: A comparative study. *Journal of psycholinguistic research*, *46*(6), 1471-1483.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967.

Martinez, R., & Murphy, V. A. (2011). Effect of frequency and idiomaticity on second language reading comprehension. *TESOL Quarterly*, *45*(2), 267–290.

Martin, K. I., & Ellis, N. C. (2012). The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, *34*(3), 379-413.

Martinez, R., & Schmitt, N. (2012). A phrasal expression list. *Applied Linguistics*, 33(3), 299-320.

Matsuo, T. (2017). *The Roles of Lexical Size, Depth, and Automaticity of Word Recognition on Reading Comprehension* (Unpublished doctoral dissertation). Temple University, Japan.

McLaughlin, B. (1989). *Theories of second-language learning*. London: Arnold.

McNamara, T. P. (2005). *Semantic priming: Perspectives from memory and word recognition*. Hove, England: Psychology Press.

Misyak, J. B., & Christiansen, M. H. (2012). Statistical learning and language: An individual differences study. *Language Learning*, *62*(1), 302-331.

Nation, P. (1989). Improving speaking fluency. *System*, *17*(3), 377-384.

Nation, I. S., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*(2), 223–242.

Nesselhauf, N. (2005). *Collocations in a learner corpus* (Vol. 14). Amsterdam: John Benjamins

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (Vol. 1, pp. 1–55). Hillsdale, NJ: Erlbaum.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric Theory*, (3rd ed). New York: McGraw-Hill.

O'Dell, F., & McCarthy, M. (2008). *English collocations in use: Advanced*. Cambridge, UK: Cambridge University Press.

O'Dell, F., & McCarthy, M. (2010). *English Idioms in Use Advanced with Answers*. Cambridge, UK: University Press.

Park, K., & Eliott, M. (2013). *Dictionary of Korean Idioms: Idioms, Proverbs, Sino-Korean Phrases.* Korea: Munyerim.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). London: Longman.

Pellicer-Sánchez, A. (2015). Developing automaticity and speed of lexical access: The effects of incidental and explicit teaching approaches. *Journal of Spanish Language Teaching*, *2*(2), 126–139.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383.

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*(1), 113–138.

Phillips, N. A., Segalowitz, N., O'Brien, I., & Yamasaki, N. (2004). Semantic priming in a first and second language: evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics*, *17*(2), 237–262.

Pili-Moss, D., Brill-Schuetz, K. A., Faretta-Stutenberg, M., & Morgan-Short, K. (2019). Contributions of declarative and procedural memory to accuracy and automatization during second language practice. *Bilingualism: Language and Cognition*, Advance online publication. doi: 10.1017/S1366728919000543.

Plonsky, L., & Oswald, F. L. (2014). How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language Learning*, *64*(4), 878–912.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633-654.

Potter, M. C., So, K. F., Von Eckardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and proficient bilinguals. *Journal of verbal learning and verbal behavior*, *23*(1), 23-38.

Pulido, M. F., & Dussias, P. E. (2019). Desirable difficulties while learning collocations in a second language: Conditions that induce L1 interference improve learning. *Bilingualism: Language and Cognition*, 1-16.

Pyun, D. O. (2018). *500 Common Korean Idioms*. Oxford and New York: Routledge.

Raichle, M. E., Fiez, J. A., Videen, T. O., MacLeod, A.-M. K., Pardo, J. V., Fox, P. T., & Petersen, S. E. (1994). Practice-related changes in human brain functional anatomy during nonmotor learning. *Cerebral Cortex*, *4*(1), 8–26.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC Nonword Database. *Quarterly Journal of Experimental Psychology section A*, *55*(4), 1339-1362.

Reed, J., & Johnson, P. (1994). Assessing implicit learning with indirect tests: Determining what is learned about sequence structure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(3), 585–594.

Robinson, P. (1996). Learning simple and complex second language rules under implicit, incidental, rule-search, and instructed conditions. *Studies in Second Language Acquisition*, *18*(01), 27–67.

Ruiz, S., Rebuschat, P., & Meurers, D. (2019). The effects of working memory and declarative memory on instructed second language vocabulary learning: Insights from intelligent CALL. *Language Teaching Research*.

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, *100*(1), 231–254.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, *12*(3), 329–363.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, *3*(4), 207-218.

Schoonen, R., Gelderen, A. van, Glopper, K. de, Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: The role of

linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning*, *53*(1), 165–202.

Segalowitz, N. S. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 200–219). Ann Arbor: University of Michigan Press.

Segalowitz, N. S. (2003). Automaticity and second language acquisition. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 382–408). Oxford: Blackwell.

Segalowitz, N. S. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Segalowitz, N. S., & Hulstijn, J. (2005). Automaticity in bilingualism and second language learning. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 371-388). Oxford: Oxford University Press.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*(3), 369–385.

Segalowitz, S. J., Segalowitz, N. S., & Wood, A. G. (1998). Assessing the development of automaticity in second language word recognition. *Applied Psycholinguistics*, *19*(1), 53–67.

Serafini, E. J., & Sanz, C. (2016). Evidence for the decreasing impact of cognitive ability on second language development as proficiency increases. Studies in Second Language Acquisition, 38(4), 607–646.

Shin, D., & Nation, P. (2007). Beyond single words: The most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339–348.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In Granena, G., Jackson, D. & Yilmaz, Y. (Eds.), *Cognitive individual differences in L2 processing and acquisition* (pp. 15–38). Amsterdam: John Benjamins.

Snellings, P., Van Gelderen, A., & De Glopper, K. (2004). The effect of enhanced lexical retrieval on second language writing: A classroom experiment. *Applied Psycholinguistics*, *25*(2), 175–200.

Solovyeva, K., & DeKeyser, R. (2018). Response time variability signatures of novel word learning. *Studies in Second Language Acquisition, 40*, 225-239.

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition*, *18*(03), 419–437.

Sonbul, S., & Schmitt, N. (2013). Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions. *Language Learning*, *63*(1), 121–159.

Squire, L. R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Stone, J. M., & Towse, J. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*(1), e5.

Suzuki, Y., & DeKeyser, R. (2017). Effects of distributed practice on the proceduralization of morphology. *Language Teaching Research*, *21*(2), 166-188.

Suzuki, Y. (2017). The Optimal Distribution of Practice for the Acquisition of L2 Morphology: A Conceptual Replication and Extension. *Language Learning*, *67*(3), 512–545.

Suzuki, Y., Nakata, T., & DeKeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*(3), 713-720.

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, *21*(1), 32-46.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. In Schmitt N. & McCarthy M. (Eds.), *Vocabulary description, acquisition and pedagogy* (pp. 156–180). Cambridge: Cambridge University Press.

Taatgen, N. A. (2001). A model of individual differences in learning air traffic control. In *Proceedings of the fourth international conference on cognitive modeling* (pp. 211-216). Mahwah, NJ: Erlbaum.

Tabachnick, B.G., & Fidell, L.S. (2012). *Using Multivariate Statistics.* (6th ed). NY: Pearson Education Inc.

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity, and accuracy. *Tesol Quarterly*, *50*(2), 369–393.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569-613.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154.

Ullman, M. T., & Lovelett, J. T. (2018). Implications of the declarative/procedural model for improving second language learning: The role of memory enhancement techniques. *Second Language Research*, *34*(1), 39-65.

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic Knowledge, Processing Speed, and Metacognitive Knowledge in First-and Second-Language Reading Comprehension: A Componential Analysis. *Journal of Educational Psychology*, *96*(1), 19-30.

Webb, S., & Kagimoto, E. (2009). The effects of vocabulary learning on collocation and meaning. *Tesol Quarterly*, *43*(1), 55–77.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning*, *63*(1), 91–120.

Wen, Z. E. (2016). *Working memory and second language learning: Towards an integrated approach*. Bristol, UK: Multilingual Matters.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, *32*(4), 430–449.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition*, *35*(03), 451–482.

Wolter, B., & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, *36*(5), 1193–1221.

Wolter, B., & Yamashita, J. (2018). Word Frequency, Collocational Frequency, L1 Congruency, and Proficiency in L2 Collocational Processing: What Accounts for L2 Performance? *Studies in Second Language Acquisition*, *40*(2), 395-416.

Wood, D. (2001). In search of fluency: What is it and how can we teach it?. *Canadian Modern Language Review*, *57*(4), 573-589.

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics*, *12*(1), 39-57.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied linguistics*, *21*(4), 463-489.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication*, *20*(1), 1–28.

Yamashita, J., & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *Tesol Quarterly*, *44*(4), 647–668.