Pushed Output Instruction for Vocabulary Learning: Exploring differences in learning gains and lexical profiling

Hana Mohammad Almutairi

MA, BA

Thesis submitted to the University College London
for the degree of Doctor of Philosophy

May 2020

بِسْمِ اللَّـهِ الرَّحْمَـٰنِ الرَّحِيمِ

In the Name of Allah, the Most Beneficent, the Most Merciful

Declaration

I, Hana Mohammad Almutairi confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Previous research has shown that vocabulary can be learned through pushed-output activities. However, the few previous studies on the topic have mainly focused on the acquisition of nouns. Little is known about the acquisition of other parts of speech or about other components of lexical mastery achieved through pushed-output activities. This thesis examines the effectiveness of spoken pushed-output instruction on learning the multiple meaning senses of single-word verbs and phrasal verbs by presenting two classroom intervention studies.

Study 1 explored differences between the effectiveness of spoken pushed-output and traditional vocabulary-focused instructions for learning polysemous single-word verbs and phrasal verbs. A between-subjects design was used, which included three conditions: no instruction, traditional vocabulary instruction and spoken pushed-output instruction. Both receptive and productive knowledge were investigated. The data were analysed using two approaches: (1) examining the receptive and productive vocabulary gains after instruction and (2) looking beyond the vocabulary gains by examining the lexical profile of the spoken production after instruction (i.e., overall text length, mean length of utterances, lexical diversity, lexical density and lexical sophistication). The findings indicated that with spoken pushed-output instruction, learners significantly improved not only in learning the multiple meaning senses of the target items but also in producing these meaning senses more fluently in longer, more lexically diverse, lexically denser and lexically sophisticated stretches of language. The results also indicated that single-word verbs could be learned at a similar rate to that of phrasal verbs. The results also showed that, except for the receptive gains of the first meaning sense, which had an advantage over the other meaning senses, no other differences among the three meaning

senses emerged. This study demonstrated the advantage of spoken pushed-output instruction, justifying its use in the classroom. However, there are many different types of spoken pushed-output activities that may be implemented, making it logical to ask which are the most effective.

Study 2 explored the effects of three different spoken pushed-output activities on learning polysemous single-word verbs and phrasal verbs: sentence reconstruction, listen-and-retell meaning, and picture description. The results indicated that all three activities resulted in similar recall scores but differed in their effectiveness for meaning recognition. The sentence reconstruction activity was found to be the most effective activity at the recognition level (as shown by the scores of the receptive test). The results also indicated that under similar instruction conditions, phrasal verbs are likely to be learned receptively and productively at a similar rate to single-word verbs. The results also showed that the first meaning sense was more easily recognised; however, no differences emerged neither in the recall scores nor in the mean length of utterances scores.

Overall, the findings presented in the thesis support the use of spoken pushed-output instruction in the classroom for teaching single words and formulaic sequences. Further, the findings support the idea that, if the type and amount of instruction are controlled to be the same for single-word verbs and phrasal verbs, the learnability of these two types of items may be the same. While the findings cannot be easily generalised to other types of formulaic sequences, they do encourage further research on the teaching of formulaic sequences.

Impact Statement

The findings of the present thesis provide useful insights into pedagogical approaches to vocabulary learning. This thesis supports the spoken pushed-output instruction for teaching polysemous words. The evidence presented in the present thesis also lends further support to the hypothesis that the learnability of both formulaic sequences and individual words are likely to be similar if other factors were controlled for, such as the amount and type of instruction. Further, the findings also provide insights into the feasibility of teaching the various meaning senses of polysemous words simultaneously.

Together these findings are also useful for language practitioners. This thesis provides evidence that teaching vocabulary with the means of spoken production is valuable for improving vocabulary knowledge of learners. It furthers supports the incorporation of formulaic sequences to the curriculum as learners were able to learn them as much as they learned single-word items.

Acknowledgments

First and foremost, praises and thanks to Allah, the Almighty, for all the countless gifts and showers of blessings throughout my PhD journey. The four years of my PhD were blessed and accompanied by truly remarkable people, without whom this thesis would have not been possible. My thanks go first to King Abdul-Aziz University and the Saudi Cultural Bureau for giving me the opportunity to pursue my studies and for all the support throughout my PhD.

I give my deepest and sincerest gratitude to my supervisor, Dr Ana Pellicer-Sánchez. She is truly an amazing person, who has made this experience unforgettable and given me invaluable guidance throughout my research. Her vision, dynamism, motivation, sincerity and immense knowledge constantly inspire me. She has taught me a great deal about the methodology I used to carry out this research, as well as how to present this research as clearly as possible. It has been, and hopefully will be, a great privilege and honour to work with her and carry out research under her supervision. I would like to thank her for her encouragement and for all I have learnt from her.

I would also like to thank Dr Michael Rodgers, my former and very first supervisor. Starting from my MA at the University of Nottingham, he has provided me with invaluable guidance. It was he who sparked my interest in vocabulary and was involved in the first phase of this thesis.

I am also grateful to all members of the Vocabulary Research Group at the University of Nottingham: Prof Norbert Schmitt, Samuel Barclay, Dr Mélodie Garnier, Dr Beatriz González Fernández, Dr Benjamin Kremmel, Dr Duyen Le Thi, Dr Marijana Macis, Dr Pawel Szudarski and Dr Laura Vilkaitė, for their encouragement, insightful

comments and hard questions. They all have inspired me to be a better and more critical researcher.

Very special thanks also go to my family in Nottingham: Aunt Salma Alruhaili, Abeer Al-Johani and Nojoud Aljuhani, for supporting me throughout my PhD journey and for making me bring some great memories back home with me.

My special thanks go to my dearest friends: Afaf AlThumali, Ayan Hersi, Khaled AlHamed, Nourah AlSuriakh, Reem AlAhmadi, and Wesam Almehmadi, for being there for me whenever I needed them, both in person and virtually.

My sincere appreciation also goes to all my participants at King Abdul-Aziz University in Saudi Arabia, whose tireless participation made this work possible. Also, I am grateful to Dr Mohammed Alburakati and Dr Hanady Alfuliag, for making it possible to collect my data at the university.

Last but not least, I am indebted to my family in Saudi: my mother Khadija, my aunts Salha and Sadiga, my sisters Doaa, Hanady, Rasha, and Nada, my cousins Hani, Naif and Mohannad, my nephews Ahmad, Atef, and Yousef and my niece Joury. Special thanks for their patience, encouragement and belief in me. This thesis is dedicated to them. شكراً لوجودكم في حياتي دائماً وابداً.

Contents

List of Figures

List of Tables

List of Abbreviations

BNC	British National Corpus
COCA	Corpus of Contemporary American English
DCT	Discourse Completion Tasks
EFL	English as a foreign language
ESL	English as a second language
FL	Formulaic language
FSs	Formulaic sequences
IL	Involvement load
ILH	Involvement load Hypothesis
L1	First language
L2	Second language
LD	Lexical density
LDV	Lexical diversity
LFP	Lexical frequency profiles
LS	Lexical sophistication
MC	Multiple choice
MLU	Mean length of utterances
NNS	Non-native speaker
NS	Native speaker
PO	Pushed-output
PVs	Phrasal verbs
SLA	Second language acquisition
SWs	Single words
SWVs	Single-word verb
TFA	Techniques Feature Analysis
VKS	Vocabulary Knowledge Scale
VLT	Vocabulary Levels Test

Chapter 1

Introduction

Vocabulary is an integral aspect of learning and using a language and is essential to all language skills: listening, reading, writing and speaking (Nation, 1990, 2013; Schmitt, 2010; Webb & Nation, 2017). It is also a central predictor of general language proficiency in both first (L1) and second language (L2) learning (Alderson, 2007; Laufer & Goldstein, 2004; Milton & Fitzpatrick, 2014). Vocabulary knowledge consists of not only single words (SWs) but also formulaic language (FL). Research has shown that a high percentage of language is formulaic (Conklin & Schmitt, 2012; Erman & Warren, 2000; Wray, 2002). Therefore, the ability to learn and use FL is considered a good predictor of language proficiency (Crossley, Salsbury & Mcnamara, 2014; Durrant & Schmitt, 2009; Kremmel, Brunfaut & Alderson, 2015; Stengers, Boers, Housen & Eyckmans, 2011) and highly valuable for improving fluency (Schmitt, Jiang & Grabe, 2011; Wood, 2006), facilitating language processing (Conklin & Schmitt, 2008), and improving reading comprehension (Kremmel et al., 2015).

Research on vocabulary pedagogy has traditionally centred on learning SWs (Pellicer-Sánchez & Boers, 2019), and in the last decade, there has been a substantial increase in the number of studies examining the learning of formulaic sequences (FSs) (Pellicer-Sánchez, 2020). Various studies have shown that FSs can be learned through both incidental approaches (e.g., Laufer & Girsai, 2008; Pellicer-Sánchez, 2015; Sonbul & Schmitt, 2013; Szudarski, 2012; Szudarski & Carter, 2016; Webb, Newton & Chang, 2013) and deliberate approaches (e.g., Boers, Dang & Strong, 2017; Boers, Demecheleer, Coxhead & Webb, 2014; Chan & Liou, 2005; Jones & Haywood, 2004; Le-Thi, Rodgers & Pellicer-Sánchez, 2018; Nassaji & Tian, 2010; Peters & Pauwels, 2015; Sun & Wang,

2003). However, comparing the learning of SWs and FSs is an under-researched area. The direct comparison of SWs and FSs learning is highly important for understanding whether the approaches typically used to teach SWs are also appropriate for teaching FSs. Research has shown that EFL learners often struggle with the appropriate use of FSs; for instance, Laufer and Waldman (2011) have shown that advanced EFL learners tend to overuse or misuse FSs. Further, it has been shown that learners' knowledge of FSs often lags behind their knowledge of SWs (e.g., Bahns & Eldaw, 1993; Granger, 1998; Nesselhauf, 2003, 2005). These studies have indeed shown that learning FSs was more challenging for learners. In addition, the formal features of single words and FSs are also different. FSs are by definition longer and have a more complex form and often have more figurative meanings which have been claimed to add to their learning burden (Webb & Nation, 2017). These features of FSs could support the claim that FSs and intrinsically more difficult than single words. If FSs are indeed more difficult to acquire than SWs, one would expect that the learnability of FSs might require different type and amount of exposure. However, enough empirical evidence to support this claim is yet to be provided. Very few studies have experimentally compared the acquisition of FSs and single words. The few available studies suggest mixed results (Alali & Schmitt, 2012; Kasahara, 2011, 2010; Laufer & Girsai, 2008; Peters, 2012, 2014). Therefore, one of the central aims of the present thesis is to investigate differences in the acquisition of SWs and FSs, represented in the present studies by phrasal verbs.

Learning vocabulary, either SWs or FSs, is a complex process requiring consistent and sustained practice. Many words have multiple meaning senses (Gyori, 2002; Murphy, 2004), which further adds up to the difficulty and complexity of learning vocabulary. Polysemous words – i.e., words with multiple meaning senses – are considered challenging by both learners and teachers for multiple reasons. Firstly, the meaning senses of these words are sometimes unsystematic (Csábi, 2004); that is, one meaning sense

cannot be inferred from the other. For instance, the word *bank* can be said to have two meaning senses: (1) *the land alongside or sloping down to a river* and (2) *an organisation where people and businesses invest or borrow money*. These two meaning senses are not systematically related and thus cannot be inferred from one another. Secondly, learners may be inclined not to abandon the meaning sense they already know, even if it does not fit the given context (Laufer, 1997). Thirdly, it is very common for the different meaning senses to be taught in an unorganised list of meanings represented through one phonological form, which might give the impression that these meaning senses are arbitrary (Evans & Tyler, 2004).

Research on the learning of polysemous words is scarce, and most research has focused on testing learners' knowledge of polysemous SWs (e.g., Bensoussan & Laufer, 1984; Crossley, Salsbury & McNamara, 2010; Kellerman, 1986; Schmitt, 1998) and FSs (Garnier & Schmitt, 2016). Pedagogical studies on polysemous words have relied heavily on cognitive linguistics, examining the roles of cognitive mechanisms in building up the network of meaning senses (e.g., Boers, 2000, 2013; Boers & Demecheleer, 1998; Boers, Eyckmans & Stengers, 2007; Boers & Lindstromberg, 2008; Lindstromberg, 1996; MacLennan, 1994). These studies have considered two main approaches: the all-at-once approach, in which all the meaning senses are taught at one exposure (e.g., Brodzinski, 2009; Csábi, 2004; Verspoor & Lowie, 2003), and the piecemeal fashion, in which learners are introduced to one meaning sense per exposure (Shortall, 2002). Findings from the few available studies support the all-at-once approach for teaching polysemous words, as it has the potential to promote successful linking of the meaning senses in the mental lexicon, which is highly recommended for stronger recall (Verspoor & Lowie, 2003). Given the limited empirical evidence on the teaching of polysemous vocabulary, it remains unclear which pedagogical approaches are more beneficial for learning it. Hence, examining the learning of multiple meaning senses of vocabulary, both individual words

and multiword units, is the second central aim of the present thesis.

In the relevant research, two main approaches to vocabulary learning have been identified: incidental and deliberate approaches (e.g., Newton & Nation, 1997; Schmitt, 2000; Siyanova-Chanturia & Pellicer-Sánchez, 2019; Webb & Nation, 2017; Wode, 1999). Incidental learning refers to the learning of vocabulary while reading and listening when vocabulary acquisition is not the focus of attention or the target of teaching (Wode, 1999), whereas deliberate learning requires purposefully directing learners' attention toward vocabulary (Schmitt, 2000). One deliberate approach that has received considerably less attention in research is pushed-output instruction. Pushed-output instruction has been shown to be beneficial for developing productive knowledge of vocabulary (e.g., De la Fuente, 2006; De la Fuente, 2002; Ellis & He, 1999; Nassaji & Tian, 2010). The third central aim of the present thesis is to address the scarcity of studies investigating this approach, which is reviewed in more detail in the next section.

1.1. Vocabulary learning through spoken pushed-output instruction

There has been constant debate surrounding the role played by spoken and written output in second language acquisition (SLA). Within SLA theories, particularly the Input Hypothesis (S. D. Krashen, 1981, 1984), the notion that input is the only factor significant for SLA appeared to neglect or minimise the role of output in interlanguage development. Krashen's Input hypothesis became the prominent paradigm for SLA research, and more attention was devoted to input over output, which was given a relatively inactive role in language acquisition. However, Merrill Swain, the originator of the Comprehensible Output Hypothesis and the first advocate for the significant role of output in SLA, demonstrated that language learning requires both comprehensible input and output processes. Swain argued that learners may have limited comprehensible output and that

this limitation leads to a lack of native-speaker (NS) productive competence. Learners' output is limited in two ways. First, they have little opportunity to productively use the language in the classroom. Second, they are not "pushed" in their output: in order to improve, learners must make efforts to stretch their interlanguage resources and move beyond their current level of language development (Swain, 1985, 1995). Swain (1985) contended that in order to learn how to speak, learners must actually speak. Swain also posited the idea that factors other than input affect language performance; particularly, in her Comprehensible Output Hypothesis, she suggested that when learners are forced to speak more often, they attend to linguistic elements such as grammar and vocabulary more closely when listening.

Vocabulary acquisition through output can help learners in three ways: by encouraging the use of new vocabulary, by helping learners negotiate the meaning of unknown vocabulary and by strengthening learners' knowledge of partially known items by giving them opportunities to use them in production (Nation & Meara, 2010). The theoretical basis behind this type of vocabulary learning draws from the Depth of Processing Hypothesis (Craik & Lockhart, 1972), which essentially states that the more cognitively involved learners are with an item, the better they will remember it. Researchers have suggested that this cognitive involvement and depth of processing can be encouraged when learners encounter items in contexts different from those in which they have previously encountered the words – known as "generative use" (Nation & Webb, 2011). Overall, studies have shown that the design features of language acquisition activities may have an effect on vocabulary learning (Webb & Nation, 2017). However, while researchers have looked at how task design may affect grammar acquisition, the effect of different output activities on vocabulary learning is an under-researched area (Nation & Webb, 2011). This is surprising, given that many language-learning materials involve learners producing vocabulary orally from an early stage (Shintani, 2011).

There is empirical evidence supporting the pushed-output instruction regarding different components of language competence, such as the acquisition of grammatical structures (e.g., Izumi, 2002; Izumi & Bigelow, 2000; Izumi, Bigelow, Fujiwara & Fearnow, 1999; Morgan-Short & Bowden, 2006; Nobuyoshi & Ellis, 1993; Shehadeh, 2002) and the acquisition of vocabulary (e.g., De la Fuente, 2006; De la Fuente, 2002; Ellis & He, 1999; Ellis, Tanaka & Yamazaki, 1994; Loschky, 1994). Of particular relevance to the present thesis, studies on vocabulary acquisition through output instruction have demonstrated that output activities are more beneficial for language learning than input-only activities (e.g., De la Fuente, 2002; Ellis & He, 1999; Ellis, Tanaka & Yamazaki, 1994; Loschky, 1994). However, previous research in this area is insufficient for several reasons. Firstly, most of the studies focused on investigating written output activities for vocabulary learning (e.g., Folse, 2006b; Holster & DeLint, 2012; Keating, 2008; Laufer & Hulstijn, 2001; Nassaji & Tian, 2010; Rassaei, 2017; Tahmasbi & Farvardin, 2017; Webb, 2005; Zou, 2017). Only a few studies examined vocabulary learning through spoken output activities (e.g., De la Fuente, 2006; De la Fuente, 2002; Ellis & He, 1999). Secondly, the target items in most of these previous studies were nouns. Thirdly, most studies examined output only in comparison with input instructional styles, leaving a gap in empirical evidence regarding which type of output activities promote vocabulary learning. Furthermore, with the exception of Nassaji and Tian's (2010) study, research investigating pushed-output activities for the learning of FSs has been neglected. Finally, no previous study has examined the acquisition of polysemous words, SWs or FSs through spoken pushed-output instruction. These neglected areas are, therefore, the central aims of the present thesis. It focuses on filling these gaps by examining the role of pushed-output in vocabulary acquisition, specifically high frequency single-word verbs (SWVs) and phrasal verbs (PVs). The components of vocabulary knowledge targeted in the present thesis are the breadth of knowledge

(acquisition of new vocabulary items, focusing on receptive and productive knowledge of form and meaning) and depth (conceptualised as polysemy). This thesis explores how pedagogical spoken pushed-output activities can be effectively designed and implemented. The context in which the studies of the present thesis were conducted is reviewed in the following section.

1.2. The Context: English teaching at Saudi universities

In Saudi Arabia, students begin learning English in the fourth year of elementary school, as a standalone course of a maximum of three hours a week. After secondary school, they enter the intensive English Preparatory Year programme in the first year of university. This programme aims to further advance the English proficiency of Saudi students moving into the university system and improve their English comprehension. Learners receive 15 hours of English teaching per week for a full academic year. The textbooks used are the New Headway series, from beginner to upper-intermediate level. Apart from these classes, whether students are further exposed to English as the medium of instruction depends on their area of study: students in medical fields are taught in English, while engineering, arts and humanities students are taught in Arabic, using English textbooks (Al-Seghayer, 2011).

Saudi universities also offer English-related bachelor's degree programmes, such as English literature, English translation, and linguistics. These degree programmes last for four years, beginning after the preparatory year. The syllabus in the first year is the same in all Saudi universities, teaching the four language skills – reading, writing, listening and speaking – separately, as well as grammar and vocabulary. However, grammar, reading and writing are allotted three to four hours of instruction per week, while listening and speaking are allotted only one class. After the first year, students begin

studying different subjects within their chosen courses. For instance, students on a translation course begin taking classes in different types of translation, such as medical and political, and students of English literature begin studying novels and plays. The assessment scale of student achievement is the same across Saudi universities. The full mark is 100: 10 marks for attendance and participation, 30 marks for the midterm exam and 60 marks for the final exam. In order to pass the course, students need to earn 60 marks in total. Considering all of these factors, it is clear that Saudi universities perceive the learning of English as crucial to the education of their students (Al-Seghayer, 2011).

Despite this commitment to the study of English, several studies have reported that Saudi students often end up lacking the skills they need to communicate in academic settings (e.g., Al-Akloby, 2001; Al-Motairi, 2005; Al-Nafisah, 2000; Al-Nujaidi, 2003; Al-Seghayer, 2011). Secondary school graduates appear to have low proficiency in English, despite having spent an average of nine years formally studying it. For instance, Al-Nujaidi (2003) reported that Saudi students graduated secondary school with a vocabulary of only 500–700 words. Saudi Arabia ranked 39th of 40 nations that took part in English academic and general training tests conducted by the Cambridge Examination Centre in 2009 (Al-Seghayer, 2011). There appear to be several reasons for this insufficient level of proficiency among Saudi learners. First, the method of teaching is highly dependent on textbooks and particular methods such as the grammar-translation and audio-lingual methods (Al-Motairi, 2005; Al-Nafisah, 2000). Secondly, the activities and topics used in the textbooks do not appear to engage students' interest (Al-Akloby, 2001). Further, regarding vocabulary, learners' failure to learn English words may be related to the fact that only two aspects of a word are given most of the attention in the Saudi classroom: pronunciation and meaning (Al-Akloby, 2001). Even learners graduating with an English bachelor's degree were found to have achieved an unsatisfactory proficiency in English (Al-Seghayer, 2011). One unpublished study by the

Ministry of Education in Saudi Arabia revealed that the average TOFEL score of Saudi English-major students after graduating was 430 (Al-Seghayer, 2011). It is evident that Saudi learners lack sufficient knowledge of English to effectively communicate in contexts beyond the classroom and that this lack may be related to either the materials or instruction style provided in formal education. There are plenty of instructional approaches that may help improve the vocabulary knowledge of learners. One of these is pushed-output instruction, which has been shown to effectively improve vocabulary learning, as well as writing and speaking skills. Thus, the present thesis focuses on examining pushed-output instruction in the Saudi context, but only in spoken mode.

1.3. Structure of the thesis

This thesis reports findings of two classroom intervention studies looking at teaching polysemous SWVs and PVs through spoken pushed-output instruction. Chapter 2 provides an extensive literature review and general theoretical background for the two studies, beginning with a discussion of the different conceptualisations of vocabulary knowledge and identifying the one followed in the present thesis. Then, I summarise the research on polysemy and the findings of previous studies on the pedagogy of polysemous words. After presenting the different definitions of FSs, I propose the definition that will be used in the present thesis. I also summarise the identification criteria for FSs and define four categories of FSs. Following that, I present the different approaches for learning FSs and explore studies examining differences between SWs and FSs. Next, I present a comprehensive overview of the pushed-output instruction in vocabulary learning, starting with the characteristic and working definition of spoken pushed-output vocabulary instruction used in the present thesis. Further, the theoretical grounds for pushed-output vocabulary instruction in the present thesis are reviewed within several frameworks: the

Output Hypothesis, the Interaction Hypothesis, and the Transfer Appropriate processing theory. Then, research into the pushed-output instruction in grammar and vocabulary studies are discussed. The chapter concludes with an overview of the frameworks that can potentially be implemented to design pushed-output activities.

Each chapter is written as a stand-alone study with a detailed description of preparation, analysis, results and interpretations of the findings. Chapters 3 and 4 report Study 1, which examined the effectiveness of spoken pushed-output instruction in comparison to traditional (i.e., vocabulary-focused) instruction for polysemous SWVs and PVs. Chapter 3 presents the analysis of the receptive and productive (spoken) learning gains by learners in the treatment groups, and Chapter 4 presents the analysis of the lexical profiling of learners' production. Study 2, presented in Chapters 5 and 6, investigated the differences between three spoken pushed-output activities for learning polysemous SWVs and PVs. Chapter 5 reports the differences in receptive and productive learning gains, and Chapter 6 explores differences in lexical profiling of utterances produced by learners within the three treatment groups.

Finally, Chapter 7 summarises the findings of previous chapters and discusses them within the broader context of research on the teaching of vocabulary. This final chapter also presents several limitations of the studies presented in the thesis and provides potential directions for future research.

Chapter 2

Pushed-output instruction in vocabulary learning: A review of the Literature

This chapter provides the background for the empirical work presented in this thesis. It provides a broad overview of the research on pushed-output instruction in L2 vocabulary studies. It is divided into four main sections. The first one presents the different components and dimensions of lexical knowledge, including the notions of receptive vs. productive knowledge of vocabulary and depth vs. size of vocabulary knowledge. It then provides an overview of polysemy in L2 vocabulary learning. The second part presents definitions and types of FSs along with an overview of the learning approaches investigated in the literature. The third section introduces the pushed-output instruction in vocabulary learning and presents the conceptualisation of spoken pushed-output instruction for the present thesis, including an overview of the theoretical frameworks that lend support for this instructional method: the Output Hypothesis, the Interaction Hypothesis and Transfer-Appropriate Processing theory. It then reviews the body of research evaluating its effectiveness in two linguistic dimensions: the acquisition of grammatical structures, as well as the acquisition of vocabulary. Finally, the last section focuses on some key frameworks that can be used to design pushed-output vocabulary activities, followed by a summary and conclusion of significant points.

2.1. Dimensions of lexical knowledge

Many vocabulary scholars acknowledge the complex and multifaceted nature of vocabulary knowledge (e.g., Chapelle, 1994; Freebody & Anderson, 1981; Nation, 2001, 2005; Read, 2000, 2004). This multifaceted nature of vocabulary knowledge is often described through multi-dimensions frameworks. One such multifaceted approach to the conceptualisation of vocabulary is Chapelle's (1994) framework, in which vocabulary ability is described as having three aspects: (1) context of vocabulary use, (2) vocabulary knowledge and its fundamental processes and (3) lexicon organisation (see Figure 2.1). The context of vocabulary use includes both the linguistic context and the pragmatic context. The linguistic context describes the sentences in which a word occurs, while the pragmatic context includes variances between formal and colloquial uses of words across different generations (Read, 2000). The second aspect is vocabulary knowledge and its fundamental processes, which formally consist of four main components. The first of these is vocabulary size, i.e., how many words a person knows within a particular context. The second component is knowledge of word characteristics, which includes graphemic, phonemic, semantic, syntactic, collocational and phraseological properties. The third component is lexicon organisation, which refers to how words are related to one another. The fourth component is fundamental vocabulary processes, which include the automaticity with which one accesses and uses words. The third aspect described by Chapelle (1994) is metacognitive strategies for vocabulary use. This refers to how to implement language components in contextualised communication, i.e., using words to execute communicative goals and intentions. Chapelle’s (1994) framework is considered to be a broad vocabulary ability framework that includes both knowledge about language and knowledge about how to use language. Moreover, it distinguishes between the two constructs of vocabulary ability and vocabulary knowledge.

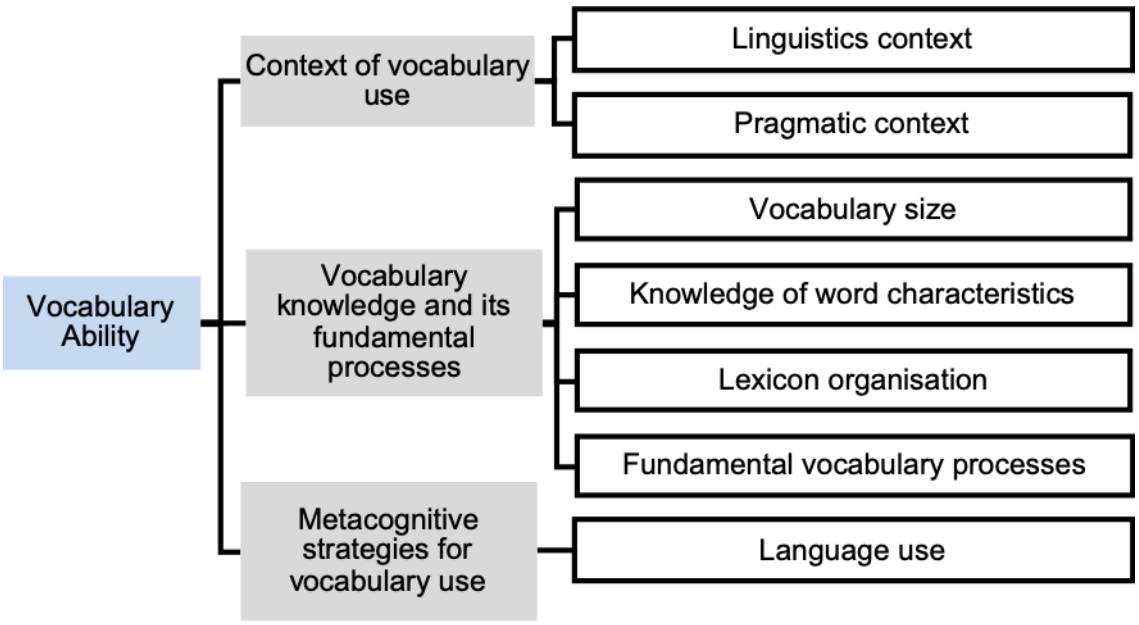


Figure 2.1 Chapelle's framework of vocabulary ability

Another vocabulary knowledge framework was proposed by Qian (2002). It consists of four dimensions: (1) vocabulary size, (2) depth of vocabulary knowledge, (3) lexical organisation, and (4) automaticity of receptive and productive knowledge. All the components in Qian's framework are linked to the second aspect in Chapelle's framework (1994), i.e., vocabulary knowledge and its fundamental processes. Depth of word knowledge is equivalent to knowledge of word characteristics in Chapelle's (1994) framework, such as phonemic, graphemic, morphemic, syntactic, semantic, collocational, and phraseological properties (Qian, 2002). Lexical organisation refers to "the storage, connection, and representation of words in the mental lexicon", and automaticity of receptive productive knowledge includes "all fundamental processes through which access to word knowledge is achieved for both receptive and productive purposes" (Qian, 2002, p. 516).

Henriksen (1999) suggested three dimensions for lexical competence: (1) partial-precise knowledge, (2) depth of knowledge, and (3) receptive to productive knowledge. Partial-precise knowledge refers to degrees of vocabulary knowledge, and receptive to productive knowledge is the ability to use vocabulary in comprehension and production.

A more influential framework of vocabulary knowledge was provided by Nation (1990, 2001, 2013), who described multiple aspects of what is involved in knowing a word. At a general level, vocabulary knowledge involves three aspects: form, meaning and use. In each aspect of these, Nation (1990, 2001, 2013) includes receptive and productive dimensions. For example, the first aspect in the framework, form, includes three dimensions of knowledge: spoken, written and word parts. Knowledge of the spoken form incorporates the word's sound (receptive) and pronunciation (productive). Knowledge of the written form involves knowing what a word looks like (receptive) and knowing how to write or spell a word (productive). Knowledge of word parts integrates the recognition of affixes and stems and relating those parts to a word's meaning (receptive), as well as the ability to construct a word using the right word parts (productive). Table 2.1 includes all the different lexical elements included in Nation's (1990, 2001, 2013) framework.

Table 2.1 *What is involved in knowing a word, by Nation (2013, p. 49)*

Form	Spoken	[R]	What does the word sound like?
		[P]	How is the word pronounced?
	Written	[R]	What does the word look like?
		[P]	How is the word written and spelled?
	Word Parts	[R]	What parts are recognizable in this word?
		[P]	What word parts are needed to express this meaning?
Meaning	Form and Meaning	[R]	What meaning does this word form signal?
		[P]	What word form 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 does this make us think of?
		[P]	What other words could we use instead of this one?
Use	Grammatical Functions	[R]	In what patterns does the word occur?
		[P]	In what patterns must we use this word?
	Collocations	[R]	What words or types of words occur with this one?
		[P]	What words or types of words must we use with this one?
	Constraints on Use	[R]	Where, when and how often would we expect to meet this word?
		[P]	Where, when and how often can we use this word?

Nation's (1990, 2001, 2013) framework is considered to be the most comprehensive and highly influential framework. This taxonomy can be seen as the most accurate one for vocabulary dimensions to date; however, it does not specify the relationships between these dimensions (González-Fernández & Schmitt, 2019). The contributions of the different components to the knowledge construct are not clear, for instance, whether knowledge of the form-meaning link explains the variance in vocabulary and whether some components will be acquired before others (González-Fernández & Schmitt, 2019). González-Fernández and Schmitt (2019) explored the relationships and order of acquisition of multiple components of Nation's framework, namely: the form-meaning link, derivatives, multiple meanings, and collocations. They

reported that all the components were strongly intercorrelated and a high level of parallel learning of the components occurred in the process. This suggests that knowledge of one component facilitates the learning of other aspects. They also found that recognition knowledge tended to be acquired before recall knowledge across the four components, which suggests that the distinction between receptive and productive knowledge in Nation's framework is fundamental to vocabulary knowledge development. They concluded that "productive knowledge of all components comes later in the learning process and would seem a sensible target for pedagogy" (González-Fernández & Schmitt, 2019, p. 21). The distinctions between receptive vs. productive knowledge as well as size vs. depth of vocabulary knowledge are examined in turn.

2.1.1. Receptive and productive knowledge of vocabulary

One common distinction when discussing vocabulary knowledge is the distinction between receptive and productive knowledge (Nation, 2001, 2005; Read, 2000; Schmitt, 2000). It refers to different degrees of knowing a word, that is, understanding a word when one listens or reads is receptive knowledge, while being able to produce a word, either in spoken or written form, is productive knowledge, which requires deeper knowledge of a word (Nation, 2005; Schmitt, 2000). This distinction between receptive and productive knowledge of vocabulary is widely accepted among vocabulary scholars of both L1 and L2 vocabulary development. Laufer (1998) has used alternative terms for these same concepts, 'passive' to mean receptive and 'active' to mean productive vocabulary knowledge. Receptive (passive) vocabulary knowledge refers to understanding the most frequent core meaning sense of a word, compared to productive (active) vocabulary knowledge that can be divided into two types: free and controlled (Laufer, 1998). Free productive vocabulary knowledge entails the use of words at one's free will, without any specific prompts for particular words, whereas controlled

productive vocabulary knowledge has to do with producing words when prompted by a task (Laufer, 1998).

Some studies, such as Melka (1997), viewed receptive and productive knowledge of vocabulary as two opposite ends of a continuum, representing different degrees of familiarity with a word. In other words, learners will develop limited knowledge when they encounter a word for the first time in such a way that they may not remember it until they come across it again. But when they start developing knowledge of a word's spelling, pronunciation, meaning, grammar, range of use and so on, they will be able to use it themselves. This means that receptive knowledge gradually transforms into productive knowledge; however, Melka (1997) acknowledged that if there is a continuum, it is not easy to define the cut-off point between receptive and productive knowledge. Other scholars, such as Meara (1990), have considered productive (active) and receptive (passive) vocabulary knowledge as distinct entities (see Figures 2.2 & 2.3). Meara (1990) used Graph Theory (Wilson & Beinke, 1979), which is primarily used in vocabulary research for exploring the organisation of vocabulary in the lexicon, to show different association patterns in English vocabulary.

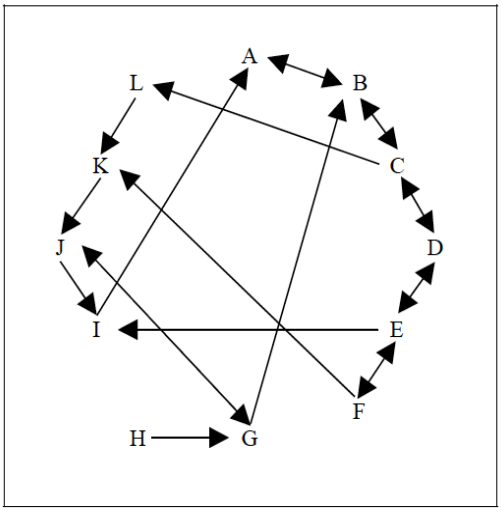


Figure 2.2 A hypothetical association network, by Meara (1990, p. 153)

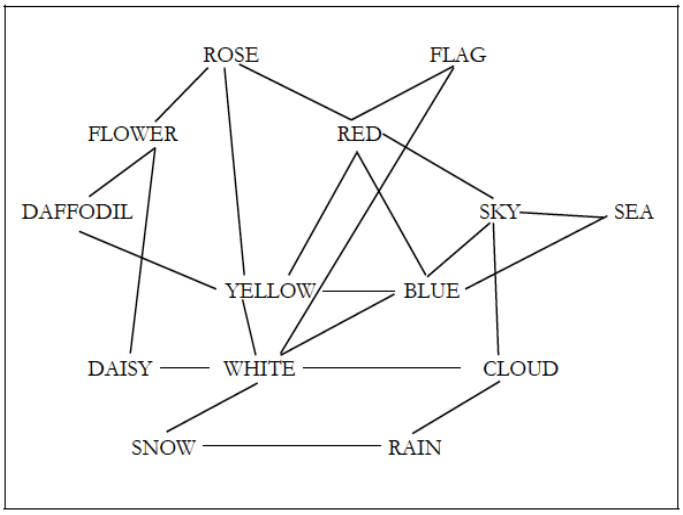


Figure 2.3 A simple word association network, by Meara (1990, p. 152)

Figure 2.2 presents a group of words (nodes) connected together by directional associations (arcs). Node H corresponds to Meara's (1990) notion of passive vocabulary. This node is part of the overall network; yet, it is different from the other nodes, in that all the other nodes have two directional arcs either leading to them or leading away from them. Node H has nodes that lead from it, which means that if you start at node H, you may reach all the other nodes, but if you start anywhere else in the network, then node H is unreachable. This implies that active (productive) vocabulary is accessible from anywhere in the vocabulary network and allows access to other parts of the network, too. Moreover, active vocabulary can be activated by word association, i.e., it can be activated by other words without requiring external support to recall it. On the other hand, passive (receptive) vocabulary is part of the overall network but is not accessible from other parts of the network. It can be activated by different external stimuli, such as spelling and sound. In other words, you cannot recall a passive vocabulary without external stimuli, but you can recognise it when you see it or hear it.

In an attempt to conceptualise receptive and productive vocabulary knowledge,

Read (2000) presented four distinctive types of receptive and productive vocabulary knowledge. He classified receptive and productive knowledge as context-independent and context-dependent. Receptive vocabulary was divided into: recognition (i.e., the ability to recognise a word when it is shown) and comprehension (i.e., the ability to understand a word when encountering it in a context); and productive vocabulary into: recall (i.e., recovery of a word from memory, activated by a stimulus word) and use (i.e., using the word productively in writing or speaking). Recognition and recall are context-independent processes, whereas comprehension and use are context-dependent ones.

While Read (2000) divided productive vocabulary on the ground of having a context or not, Laufer, Elder, Hill, and Congdon (2004) divided recognition and recall components according to two perspectives: what to retrieve (i.e., whether to use form to retrieve meaning or meaning to retrieve form) and how to retrieve vocabulary knowledge (i.e., recognition vs. recall), which resulted in a four-type classification of vocabulary knowledge (see Table 2.2).

Table 2.2 *Types of Vocabulary Knowledge, by Laufer, Elder, Hill, and Congdon (2004, p. 206)*

	Recall	Recognition
Active (Productive) (retrieval of form)	Active Recall	Active Recognition
Passive (Receptive) (retrieval of meaning)	Passive Recall	Passive Recognition

Schmitt (2010) agreed with the importance of Laufer et al.'s (2004) classification; however, he argued that their terminology was too confusing for both researchers and teachers. Hence, Schmitt (2010) relabelled Laufer et al.'s (2004) classification in a more

transparent way (see Table 2.3).

Table 2.3 *Level of mastery of the form-meaning link, by Schmitt (2010, p. 86)*

Word Knowledge	**Tested Word Knowledge**	
Given	*Recall*	*Recognition*
Meaning	Form recall (supply the L2 item)	Form recognition (select the L2 item)
Form	Meaning recall (supply definition/ L1 translation etc.)	Meaning recognition (select definition/L1 translation etc.)

2.1.2. Breadth and depth of vocabulary knowledge

Another common distinction when discussing vocabulary knowledge lies in its breadth (or size) and depth (or quality), as first introduced by Anderson and Freebody (1981). Vocabulary breadth is a quantitative aspect that refers to how many words are known by their primary meaning, whereas vocabulary depth is a qualitative aspect that refers to the degree, or depth, of word knowledge (Anderson & Freebody, 1981, p. 93). Further, Read (1993) defined depth as the quality of word knowledge, which includes knowledge of collocations and synonyms.

Most vocabulary research has focused on the examination of vocabulary size. Vocabulary size is generally accepted as the form-meaning link of words, though it may also sometimes include knowledge of the written and spoken forms of vocabulary (Milton, 2009). Research on depth of vocabulary knowledge, on the other hand, has been overlooked (Henriksen, 1999; Nation & Webb, 2011; Schmitt, 2010), mainly due to the multifaceted nature of depth of vocabulary knowledge. This has resulted in having

multiple conceptualisations of depth with no agreement on the best conceptualisation of depth of vocabulary knowledge. For instance, depth of vocabulary knowledge can be conceptualised by the components first identified by Richards (1976) and Nation (1990), including word characteristics such as word association and frequency, syntactic characteristics and affix knowledge. In Nation's (2001) framework (see Table 2.1), depth of vocabulary knowledge involves aspects that go beyond basic form-meaning links, such as: concept and referents, associations, collocations, and constraints on use. Depth of vocabulary knowledge reflects knowledge of the semantic links between individual words in the lexicon, as asserted by Haastrup and Henriksen (2000) and Meara (1996).

Read (2004) introduced another framework for conceptualising the components of vocabulary depth, which includes three aspects; first, precision of meaning, reflecting the degree of knowledge about a word's meaning, which can be either partial or deep. The second component is comprehensive word knowledge, which includes all the word's characteristics (i.e., its orthographic, morphological, phonological, syntactic, semantic, collocational, and pragmatic features). The third component is network knowledge, which describes the integration of a word into a lexical network in the mental lexicon, as well as the ability to either associate it or distinguish it from related words. Li and Kirby (2014) described depth as richness of vocabulary knowledge, which includes knowledge of the collocations, morphological components and polysemy of a word. Depth of vocabulary knowledge may indicate that a word is well learned, with its different characteristics and multiple meanings, and this may be reflected in using it productively in either writing or speaking. Nation (2013) stated that depth of vocabulary knowledge refers to having some degree of proficiency in multiple word-knowledge dimensions. Overall, the intrinsic difficulty in defining and measuring depth has made its conceptualisation more problematic and resulted in some variations in what depth of knowledge involves among researchers (Milton, 2009).

At first glance, these two constructs are distinct. From a pedagogic point of view, it seems reasonable to think of size and depth as discrete constructs that may justify different teaching strategies (e.g., intensive for a deeper knowledge of words or extensive to teach a wider range of words) (Schmitt, 2014). However, there have been some reservations about the usefulness of this distinction; Vermeer (2001) asserted that there is no clear distinction between depth and breadth of vocabulary knowledge, due to the strong correlation between the two (correlations between size and depth in his study were .85 for monolingual and 0.76 for bilingual Dutch Kindergarteners, n=50). Nurweni and Read (1999) proposed that the level of language proficiency may affect the relationship between the two dimensions. That is, if learners are at lower levels, depth and breadth of vocabulary do seem to be more distinct, whereas the distinction between the two may not apply to learners of higher proficiency, which suggests that the two dimensions facilitate each other. Meara (1996) contended that in a case where a learner's vocabulary is small (under 5,000–6,000 words, or English lemmas), size is the only significant aspect, and once this level is reached, vocabulary size becomes less significant (Meara, 1996).

Examinations of size and depth have been conducted for both NSs of English (e.g., D'Anna, Zechmeister, & Hall, 1991; Dupuy, 1974; Seashore & Eckerson, 1940) and non-native speakers (NNSs) (e.g., Hazenberg & Hulstun, 1996; Meara & Buxton, 1987; Nurweni & Read, 1999). In fact, vocabulary size research, for both NS and L2 learners, is the most prominent topic in the vocabulary literature (e.g., Laufer et al., 2004; Laufer & Goldstein, 2004; Webb, 2005). However, one can argue that vocabulary size takes a shallow view of how well any word is known (Read, 2000). Vocabulary depth, on the other hand, has received less attention in the literature than vocabulary size, although the latter "can give a more representative picture of the overall state of the learners' vocabulary than an in-depth probe of a limited number of words" (Read, 2000, p. 115). Examinations of vocabulary depth have been mostly limited to psycholinguistic studies

concerning the development of children's word knowledge (Anderson & Freebody, 1981); studies involving either NSs or adult NNSs of English are scarce (Boyle, 2009; Stahl & Bravo, 2010). Regardless of the limited number of studies concerning vocabulary depth, the prevailing findings advocate that ascertaining the depth of learners' vocabulary knowledge is relevant to pedagogic practices (Read, 2000). The reason why most research has consistently focused on the size of vocabulary is the lack of a clear understanding of vocabulary depth. Although it is accepted that word learning is incremental and may start with vague knowledge of the form-meaning connection, word learning must continue in order to become more precise and have deeper knowledge.

To conclude, Nation's (1990, 2001, 2013) framework is highly relevant to the current study for several reasons. It is the one referred to by most vocabulary scholars and the most comprehensive framework, consisting of nine components, each with two levels: receptive and productive mastery. Further, it is the only framework that differentiates between spoken and written forms of knowledge. The present study mainly focuses on the spoken production of form in an appropriate context. Moreover, as the aforementioned frameworks indicate, depth of vocabulary knowledge is reflected through mastery of different aspects of word knowledge, such as word association, word frequency, syntactic characteristics, affix knowledge, and so on. This study looks at both the size of vocabulary knowledge as well as the depth of vocabulary knowledge represented in the knowledge of multiple meaning senses of a word. The study also explores receptive and productive knowledge of vocabulary, and so Schmitt's (2010) conceptualisation of the level of mastery of the form-meaning link is highly relevant to the present study, as receptive vocabulary knowledge in speaking research empowers the learner to retrieve the meaning of a word using external stimuli, as suggested by Melka (1997), i.e., the sound or spelling of a word, while productive vocabulary knowledge enables the learner to retrieve the form of a word in order to be able to use it. Thus, this distinction can help to identify the

direction of the connection between form and meaning. The following section explores polysemy as one component of depth of vocabulary knowledge and summarises studies examining learning polysemous in EFL context.

2.1.3. Polysemy

2.1.3.1. Knowledge of Polysemy

Polysemy concerns the multiplicity of word-related senses due to contextualised variation (Fellbaum, 2000). The word *paper,* for example, may refer to a substance made from wood pulp, a daily publication, a blank sheet, or an article. A polysemous word has one core meaning with multiple related senses. Some authors distinguish between polysemy and homonymy (e.g., Gries, 2006; Lewandowska-Tomaszczyk, 2007). In homonymy, words have identical spellings and pronunciations but unrelated meanings (e.g., *bank* as in by a river and *bank* as in a financial institution), while in polysemy, one word has meanings that are often conceptually-related (e.g., *chest* as the upper front part of the human body, a case or box with a lid, and treasury of a public institution) (Lewandowska-Tomaszczyk, 2007). Other researchers distinguish between polysemy and contextual variation or vagueness (e.g., Cruse, Cruse, & Cruse, 1986; Kilgarriff, 1997; Lewandowska-Tomaszczyk, 2007). A vague word has only one meaning sense (Geeraerts, 2009; Tuggy, 1993). For example, the noun *student* does not specify gender, which means that this noun is unmarked or vague for gender, but that does not necessarily mean that it has multiple meanings (Lewandowska-Tomaszczyk, 2007). Crossley, Salsbury, and McNamara (2010) and Gries (2006) pointed out that homonymy and vagueness are two extremes on one continuum, and polysemy is found somewhere in the middle. Thus, there is no strict dividing line, and there will always be some degree of overlap between homonymy and polysemy, or between polysemy and vagueness. Further, polysemous words are more common than homonyms and vague words, and more often

the rule than the exception (Gyori, 2002; Murphy, 2004).

Weinreich (1964) classified polysemy into two types: contrastive polysemy and complementary polysemy. Contrastive polysemy includes cases where a word has multiple unrelated and distinct meanings (e.g., *bank*), whereas complementary polysemy includes instances where the meanings of the word are dependent, shared or overlap. Here, what Weinreich (1964) and Pustejovsky (1991) called contrastive polysemy is in line with what the researchers above called homonymy, which highlights different conceptualisations of polysemy. Moreover, Pustejovsky (1991) extended complementary polysemy to include two types — first, logical polysemy, where the meanings have preserved their categories or parts of speech (e.g., *door* is a noun that has the two meanings of *a physical object* or *an opening*); second, cross-categorical polysemy, where the meanings are category-changing (e.g., *hammer* can be used as a noun that refers to an object or a verb that describes an action).

Meaning variation studies often face the problematic issue of lumping vs. splitting (Gries, 2006). For lexicographers, it is particularly important to decide whether the meaning of a word is relevant, different or frequent enough to earn its own entry in a definition (Gries, 2006). It is difficult to draw a clear-cut line between the different meanings of a word – if a word is split too much, every combination of verb and object suggests different meanings, and if meanings are lumped together too much, the outcome might be one general meaning, with no variation (Yarowsky, 1993). In this thesis, in the selection of target items for the two empirical studies, it was found that splitting is indeed often the case for several very frequent spoken verbs, such as *break*, which makes it harder to decide on the most frequent senses of the verb (see Chapter 3, Section 3.3.4.1).

Corpora studies have established that context plays a significant role in determining the intended meaning of a lexical item. Polysemous lexical items are often ambiguous in isolation because they have different meanings; yet, when used contextually, they are not

at all ambiguous (Miller, 1991). For instance, Yarowsky (1993) concluded that polysemous words only have one meaning per collocation, with over 90% precision. Polysemous words tend to exhibit different meanings in different environments, with varied collocates and syntactic discourse (Gilquin, 2010); thus, different lexical items around a word help in determining the meaning the word adopts.

Polysemy in language is not rare. Nagy (1995) stated that dictionary entries show that over 40% of English words are polysemous, with an average of two or three meanings each. However, knowing one facet of a word does not require being familiar with its other meanings. Bensoussan and Laufer (1984) studied lexical guessing and stated that students face some challenges with polysemous words, which cause the highest numbers of errors. Students who already have knowledge of one meaning of a target word are often inclined not to look for its other meanings, even though that particular meaning may not be appropriate in a given context.

The present study is concerned with teaching contextual, polysemous, high-frequency spoken verbs and phrasal verbs for two reasons: variation in meaning is more frequent in high-frequency words, and verbs often differ in the sense they adopt according to the words they are combined with (Goldberg, 1995; Pustejovsky, 1991).

2.1.3.2. Polysemous words in L2 learning studies

Polysemous words are often problematic for teachers and learners because their meanings are unsystematic (Csábi, 2004), which may disincline both learners and teachers to deal with them. Thornbury (2002) stated that words with different meanings are often perceived as "a complete headache for learners" (p. 8). Laufer (1997) explained that learning polysemous words is challenging for learners because they belong to the category of "words you think you know", and learners are more reluctant to abandon a meaning they know even if does not fit into a given context (p. 26). According to Evans

and Tyler (2004), a teaching strategy is often an "unorganised list of unrelated meanings that are accidentally coded by the same phonological form" (p. 152), which may give learners the impression that those different meanings are arbitrary. They also contend that lexical classes are represented in textbooks in a piecemeal way, and when learners encounter their various uses, systematic associations remain unexplained. They exemplified this with the use of the preposition *over,* in the four sentences below, and concluded that modern teaching materials fail to explain the different meanings associated with the form *over*:

a. The picture is over the mantle.

b. The teller at the central bank switched the account over to a local branch.

c. The film is over.

d. Arlington is over the river from Georgetown. (p. 258)

Some studies have examined knowledge of polysemous words. Bensoussan and Laufer (1984), for instance, examined lexical guessing from context with a 574-word text containing 70 target words, among them some polysemous words. They tested the comprehension of words through guessing the meaning from the sentence's context. Bensoussan and Laufer (1984) reported that polysemous words were guessed considerably worse by learners than monosemous words. Learners usually opted for the meaning of a word they knew, even if it was contextually inappropriate.

Kellerman (1986) investigated the likelihood of learners' L1 transferring their L2 knowledge of polysemous words. More precisely, he asked participants to make subjective decisions about how acceptable it felt to transfer lists of meaning senses from L1 to L2. Kellerman investigated factors, such as age, context, and familiarity, that might affect the transferability of an L1 meaning sense to an equivalent L2 form. He reported

that learners were more reluctant to transfer metaphorical senses or senses unrelated to the core meaning of the form.

Schmitt (1998) tracked the use of English polysemous words used by three advanced L2 learners over the course of a year. Receptive and primed knowledge of word senses were the main aspects of vocabulary knowledge that Schmitt (1998) examined through a survey instrument. The survey required L2 learners to explicitly produce their lexical knowledge of 11 target items and this was administered three times over the year. Schmitt found that learners increased their knowledge of polysemous words; however, they were unlikely to show complete knowledge of all their senses. L2 learners had only partial knowledge of all the possible word senses available for a word, and only two of the students showed steady progress during the study (Schmitt, 1998). Schmitt argued that learning a new meaning sense is a slow process, but once learned it is rarely forgotten. Similarly, Crossley et al. (2010) conducted a longitudinal study investigating the development of L2 polysemous word use in English. They examined whether the use of L2 learners of polysemous words increased as proficiency in English improved. The researchers reported an initial increase in the number of senses produced in the first two months as the learners were developing their English proficiency; however, the use of polysemous words levelled out afterwards.

Garnier and Schmitt (2016) investigated L2 learners' knowledge of high frequency phrasal verbs in English. The study used phrasal verbs included in the PHaVE List, which was developed in an earlier study by Garnier and Schmitt (2015), presenting the most frequent meaning senses of the most frequent phrasal verbs in English. The learners were tested on their knowledge of these PVs using a gap-fill productive test. Further, several factors were investigated in relation to their knowledge of PVs, such as semantic opacity, corpus frequency, previous L2 instruction, immersion in an L2 environment, year of BA study, time spent reading and social networking per week, and hours spent listening to

music and watching films in English. Garnier and Schmitt (2016) reported that, on average, 40% of meaning senses of phrasal verbs were known, but all the different meaning senses of each phrasal verb tested were quite low, only 20%. Further, only corpus frequency seemed to predict knowledge of meaning senses along with the time spent reading and social networking. Garnier and Schmitt (2016) suggest that PVs deserve more attention in instructed contexts in the classroom or L2 engagement outside it.

Researchers working on the pedagogy of polysemy have relied heavily on cognitive linguistics, and several cognitive linguists have stated that it has the potential to inform the pedagogy of polysemy (e.g., Boers, 2000, 2013; Boers & Demecheleer, 1998; Boers, Eyckmans, & Stengers, 2007; Boers & Lindstromberg, 2008; Lindstromberg, 1996; MacLennan, 1994). Cognitive linguistics can offer insights into the semantic networking that underlies aspects of language learning, such as polysemy; hence, both researchers and teachers can incorporate these insights to develop the most appropriate teaching strategies for polysemy (Boers, 2000). Few studies have examined the efficacy of teaching strategies informed by cognitive linguistics (e.g., Csábi, 2004; Khodadady & Khaghaninizhad, 2012; Morimoto & Loewen, 2007; Verspoor & Lowie, 2003). Verspoor and Lowie (2003), for instance, examined the ability of L2 learners to guess, from the context, the peripheral meanings of a polysemous word when only given its core meaning, i.e., its most literal and concrete sense. Their findings showed that having the core meaning helps in guessing other peripheral meanings and improved the retention of meanings in long-term memory.

Csábi (2004), Khodadady and Khaghaninizhad (2012) and Morimoto and Loewen (2007) examined teaching methods relying on cognitive schemata that underline the meanings of polysemous words (Csábi, 2004; Khodadady & Khaghaninizhad, 2012; Morimoto & Loewen, 2007). Schema-based instruction is defined as “a form of

vocabulary instruction in which the process of learning a word is mediated by the use of schema" (Morimoto & Loewen, 2007, p. 351), this mainly focuses on providing learners with a basis that assists them in effectively processing the various meaning senses in a given input. Both Morimoto and Loewen (2007) and Khodadady and Khaghaninizhad (2012), a replication study of Morimoto and Loewen (2007), examined image-schema-based instruction and translation-based instruction, whereas Csábi (2004) examined force dynamics, as explained by Talmy (1988). Image-schema-based and force dynamics instruction describes language according to an experiential basis; however, the difference is that the former works better with nouns and prepositions, whereas the latter is more suited to verb phrases. Morimoto and Loewen (2007) examined the learning of two English polysemous words, *break* and *over*, by Japanese students; Csábi (2004) investigated the learning of two English verbs, *hold* and *keep*, by Hungarian students; Khodadady and Khaghaninizhad (2012) examined the learning of two French words, *arriver* and *sur*, by Iranian students. In terms of their study procedures, both Morimoto and Loewen (2007) and Khodadady and Khaghaninizhad (2012) included a control group besides their experimental groups (image-schema-based and translation-based instruction), whereas Csábi (2004) only had experimental groups but did also employ two proficiency levels: high and low proficiency groups.

Two tests were administered in Morimoto and Loewen's (2007) and Khodadady and Khaghaninizhad's (2012) studies: an acceptability judgement test to assess the receptive knowledge of target words (*over* and *break*), and a written production test three times: as a pre-test before instruction, in a two-days-delayed post-test, in a two-weeks-delayed post-test. Morimoto and Loewen (2007) reported that only in the acceptability judgement test for the preposition *over* was image-schema-based instruction significantly better than translation-based instruction. However, in all other cases, both forms of instruction were similarly effective for learning L2 polysemous words. Khodadady and

Khaghaninizhad (2012), on the other hand, reported that image-schema-based instruction tended to be as effective as translation-based instruction in an acceptability judgement test (receptive knowledge); however, it was considerably more influential in a production test than in translation-based instruction. Both studies suggest that image-schema-based instruction, from a cognitive semantics point of view, could serve as "a pedagogical device" for teaching L2 polysemous words (Morimoto & Loewen, 2007, p. 347).

Csábi (2004) used a gap-fill completion test as a one-day delayed post-test for low proficiency groups and a two-day delayed post-test for high proficiency groups. The results showed that there were significant differences between the treatments for low and high proficiency learners. The translation-based instruction group outperformed the schema-based instruction group on the second post-test for high proficiency learners. Otherwise, learners performed better in the schema-based group than the translation-based group in the other three tests.

A more recent study was carried out by Macis (2018), in which the learning of literal and figurative meanings of collocations through extensive reading was examined. Three case studies of relatively advanced L2 learners were examined in the study. The reading materials consisted of a semi-authentic novel in which the target items (n-38) were embedded and learners were assessed through one-to-one interviews on their meaning-recall level, as well as how repetition could affect this knowledge. The results showed that learners knew only about 33% of the target figurative collocation meanings. The results also showed that the effect of repetition was not always significant, but overall it had a positive effect on learning.

Several studies have suggested an all-at-once approach for the teaching of polysemous items (e.g., Brodzinski, 2009; Csábi, 2004; Verspoor & Lowie, 2003). All-at-once instruction means that all the literal and peripheral senses are presented together, as doing this allows learners to "to incorporate the figurative sense into a semantic

network more effectively and recall it later more easily" (Verspoor & Lowie, 2003, p. 569). Brodzinski (2009) argued that it might be better to approach multiple meaning senses as core meanings and derive extensions, especially for phrasal verbs. Such an approach "constitutes a neat way of presenting the different senses of a polysemous word in relation to each other" (Gilquin, 2008, p. 36).

On the other hand, other studies have propounded that polysemous words should be addressed in a piecemeal fashion, one meaning sense per exposure (Shortall, 2002; Webb & Nation, 2017). Webb and Nation (2017) suggested that the learning burden of polysemous words might be reduced by approaching one meaning sense per exposure and teaching other meaning senses as a matter of practice, i.e., only when encountered in a particular context. Perdek (2010) argued that, in the case of polysemous phrasal verbs, several phrasal verbs might hold several core meanings, and cognitive connections between the core and the figurative might be too complicated and perhaps challenging for language learners to grasp (Shepherd, 2009). It might also be challenging for language learners to perceive the links from the core meaning to figurative or peripheral ones (Shepherd, 2009); however, as Webb and Nation (2017) suggested, raising learners' awareness of the idea that one word can have multiple meaning senses, through different methods such as checking dictionaries, might help in reducing the learning burden for polysemous words.

To conclude, very few studies have examined teaching approaches for learning polysemous words, and the findings of these studies are mixed about the effectiveness of the instructional approaches employed. Hence, the present thesis examines the learning of polysemous lexical items, both SWs and FSs, within two instructional approaches: traditional and pushed-output.

2.2. Formulaic sequences: definitions, types, and learning approaches

2.2.1. Definition of formulaic sequences

A variety of terms have been used to refer to the phenomenon of formulaicity in language. Wray (2002) identified around 50 terms, including *holophrases*, *conventionalised forms*, *collocations*, *formulas*, *formulaic speech*, *prefabricated routines*, *ready-made utterances*, *multi-word units*, and *chunks*, to express FL in the literature. There is no consensus as to what the best umbrella term is to use. Siyanova-Chanturia and Pellicer-Sánchez (2019) suggested that FL is a broad, inclusive term, as it includes both SWs, such as expletives, and multiword units as well. They, therefore, opted to use FL as an umbrella term, and FSs as only one instance of it.

Defining FSs is a challenging task for applied linguists. Sinclair (1996), based on a corpus investigation, proposed that we should start conceptualising the main linguistic element not in SWs but rather in multiword units. This claim is based on the fact that SWs carry little meaning by themselves and the meaning of an utterance emerges from the co-selection of SWs (Sinclair, 1996). Since then, FL has become a central area in language research. Sinclair (1996) proposed two principles involved in the co-selection process: the open choice and idiom principles. As the terms imply, open choice refers to using SWs to deliver a meaning whereas the idiom principle implies a phraseological tendency. Basically, what a phraseological tendency suggests is that writers and speakers do not select their words in splendid isolation from each other but rather co-select the words they use to produce meaningful units of language. There might be a few exceptions to which this phraseological tendency is not applicable, and words are chosen according to the open choice principle.

Another widely cited definition is offered by Wray (2002). She considered FSs as “a sequence, continuous or discontinuous, of words or other elements, which is, or appears

to be, prefabricated: that is, stored and retrieved as a whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar" (Wray, 2002, p. 9). This definition incorporates both the meaning of the sequence and the retrieval process for it. That is, FSs are processed, memorised, and retrieved as units rather than individual words. Siyanova-Chanturia and Pellicer-Sánchez (2019) argued that this definition implies the concept of "holistic storage"; however, one unanswered question is what holistic storage is and whether it exists (Siyanova-Chanturia & Pellicer-Sánchez, 2019; Siyanova-Chanturia & Sidtis, 2019). Further, Siyanova-Chanturia and Pellicer-Sánchez (2019) stated that FL should be defined and identified in relation to common features, such as frequency, familiarity, predictability, and fixedness, and their pragmatic function. They offered a more inclusive definition of FL that does not refer to the holistic storage of sequences:

> FL, as conceived in this book, may comprise strings of letters, words, sounds, or other elements, contiguous or non-contiguous, of any length, size, frequency, degree of compositionality, literality/ figurativeness, abstractness and complexity, not necessarily assumed to be stored, retrieved or processed whole, but that necessarily enjoy a degree of conventionality or familiarity among (typical) speakers of a language community or group, and that hold a strong relationship in communicating meaning. (p. 5)

This definition is inclusive as it includes elements at both the word and phrase levels. Siyanova-Chanturia and Pellicer-Sánchez (2019) considered a FS to be a "string that is perceived by the agent (i.e., learner, researcher, etc.) to have an identity or usefulness as a single lexical unit", based on Wray's (2019) definition (p. 267). They also suggested that "this identity of usefulness as a single unit" differed in its construction depending on

the specific needs of teachers, learners, or researchers. Overall, there is no universally accepted definition of FL or FSs.

For the purposes of this thesis, following Wray's (2002) definition, FSs are defined as a string of two or more words that co-occur in near proximity to each other at the sentence level, and which are commonly processed, memorised, and retrieved as a single unit.

2.2.2. Selection and identification of FSs

Two approaches have been used to identify FSs (Barfield & Gyllstad, 2009; Paquot & Granger, 2012):

- The distributional approach: This approach, as the name implies, relies on the strength of co-occurrence of the lexical items (Halliday, 1961; Sinclair, 1996).
- The phraseological approach: This approach is a meaning-based approach in which FSs are conceptualised based on their transparency and substitutability (Cowie, 1994; Howarth, 1998)

2.2.3. Types of FSs

As indicated in the previous section, FSs are a multi-faceted concept and many different types of FSs have been identified in the literature, for instance: binomials, collocations, formulas, idioms, lexical bundles, lexical phrases, phrasal expressions, phrasal verbs, proverbs, and so on. These different categories of FSs have received different amounts of attention by researchers in different disciplines, i.e., corpus analysis studies, pedagogical research, and psycholinguistic studies, with some types (such as collocations and phrasal verbs) benefiting from more research than others (such as proverbs). While a comprehensive review of all types of FSs is beyond the scope of the

present chapter, this section reviews four types that have received the most attention in applied linguistics and SLA research: collocations, idioms, lexical bundles, and phrasal verbs. All of these are defined in the following sections. The section on phrasal verbs is the most detailed, as they are the focus of the two experimental studies presented in the thesis.

2.2.3.1. Collocations

Collocations are defined in various ways in the literature; for instance, Sinclair (1996) defined a collocation as any co-occurrence of words within a certain span. Other researchers, such as Nation (2001), considered knowledge of collocations to be a key component involved in depth of word knowledge. As Nation put it, "knowing a word involves knowing what words it typically occurs with" (p. 56) (see Section 2.1 for more details on Nation's framework on what is involved in knowing a word). Henriksen (2013) defined collocations as "frequently recurring two-to-three-word syntagmatic units which can include both lexical and grammatical words, e.g., verb + noun (*pay tribute*), adjective + noun (*hot spice*), preposition + noun (*on guard*) and adjective + preposition (*immune to*)" (p. 30). In all of these definitions, a collocation is commonly defined as the frequent co-occurrence of two or more words.

2.2.3.2. Idioms

A phrase that has a holistic figurative meaning in itself, not retrieved from the meanings of its parts, such as *have an axe to grind* or *spill the beans*, is called an idiom (Čermák, 2001; Grant & Nation, 2006). Idioms are considered one of the most prominent types of FSs due to their metaphorical, non-compositional nature, as well as their fixedness. Several researchers have argued that idioms can be modified and used in more creative ways by changing or omitting some words (Omazic, 2008; Philip, 2008; Simpson

& Mendis, 2003). For instance, Simpson and Mendis (2003) reported that the idiom, *put the heat on,* is used in at least six creative ways*: under some heat, puts some heat on, put heat on, putting heat on themselves, heat put on them* and *put more heat on* (p. 430). However, as Biber and Biber (1999) claimed, even with such lexical variations, idioms retain the figurative meaning they hold through certain content words. Idioms are considered the most prototypical type of FSs for language learners (Sidtis, 2012; Wray, 2002). McCarthy (1990) suggested that all languages have idioms, and, in English, some words are more likely to be "idiom-prone" than others, such as *bite*, e.g., *bite the bullet*, *bite the dust* (p. 6). Idioms seem to be the most prominent type of FSs in certain genres, such as informal conversation and journalism (Moon, 1998). Idioms have received a lot of attention in FL research (Kuiper, Columbus, & Schmitt, 2009), yet they constitute only a small proportion of the phrasal lexicon of NS and L2 learners.

2.2.3.3. Lexical bundles

Lexical bundles are defined as "extended collocations: bundles of words that show [a] statistical tendency to co-occur" (Biber et al., 1999, p. 989). They are very frequently used and might be considered structural chunks accompanying some communicative functions, such as expressing one's stance or organising ideas (Biber et al., 1999). Some classic examples of these phrases are *I don't know*, *on the other hand*, *in the case of the*, *it is likely to*, etc. The identification of lexical bundles relies mostly on frequency (Cortes, 2004); they are not considered FSs from a phraseological approach. Lexical bundles are in many ways the opposite of idioms as they are considered highly fixed and non-idiomatic, in the sense that their meanings are driven by the individual components of the phrase (Cortes, 2004).

2.2.3.4. Phrasal verbs

A phrasal verb is a combination of verbs and inflexible prepositional particles or adverbials, such as *take off*, *bring out*, or *put down*. Granger and Paquot (2008) distinguished between phrasal verbs (i.e., verb + particle) and grammatical collocations (i.e., verb + preposition). Grammatical collocations are made up of a lexical word (noun, verb, adjective, etc.) and a grammatical word (particle) such as aim at, afraid that, etc. (Granger & Paquot, 2008). The difference between the two is that the verbs in grammatical collocations hold the original meaning of the verb and do not have a meaning for the combination, as it is the case with phrasal verbs. Darwin and Gray (1999) define a phrasal verb as consisting of "a verb proper and a morphologically invariable particle that function together as a single unit both lexically and syntactically" (p. 76). As the definition implies, phrasal verbs can be conceptualised in two dimensions: syntactic and lexical. In the syntactic dimension, a phrasal verb functions as a single grammatical unit, which is different from prepositional verbs (e.g., *look at*). In the example *he looks at his watch*, the verb *looks* acts in isolation from the particle *at*, which is part of the prepositional phrase *at his watch*. The second dimension is the lexical one in which the phrasal verb holds its own meaning, which is different from the meaning of the verb proper by itself (*call off* vs. *call*). This is evident from the possibility to substitute PVs with SWVs (*call off* vs. *cancel*), as well as the possibility to be accompanied by a different particle and hold a different meaning (*call off* vs. *call around*).

A phrasal verb is considered a FS due to the fact that it is treated as a single lexical unit with its own holistic meaning, even though it consists of two orthographic units (Schmitt & McCarthy, 1997). Phrasal verbs are seen as notoriously problematic for L2 teaching and learning for many reasons (Boers, 2000; Garnier & Schmitt, 2015, 2016; Hulstijn & Marchena, 1989; Kao, 2001; Kurtyka, 2001; Littlemore & Low, 2006; Moon, 1998; Siyanova & Schmitt, 2007). One source difficulty is the structural property of

phrasal verbs, which consist of two or more orthographic units, each holding a different meaning. This can lead to misinterpretation of their holistic meaning and an inclination to decode the meanings of single units. Further, their syntactic peculiarity, which allows for particle movement in some phrasal verbs, could be another reason. The particle can be relocated to the other side of the object (e.g., *pass on the paper* to me vs. *pass the paper on to me).* This possibility of particle movement can be particularly challenging for learners, especially since it is optional in some phrasal verbs but forbidden in others. Third, even though phrasal verbs are considered very common in English, and other languages such as German, some languages lack such structures (such as Arabic, the participants' L1 in the present thesis). That might affect how L2 learners perceive phrasal verbs and might make them think they are unnatural constructions. Another problematic issue relates to the semantic nature of phrasal verbs (having a meaning distinct from their components) and their semantic complexity, that is, some phrasal verbs have multiple meanings, which may include idiomatic meanings. This polysemic nature of phrasal verbs is another factor adding to the learning burden of phrasal verbs.

2.2.4. Learning approaches to FSs

FSs constitute 58.6% of spoken English discourse and 52.3% of written discourse, as demonstrated by McCaig, Manser, and Warren (1994). If FSs are so frequent in English discourse, it means that proficient English speakers must have mastery of them. Knowledge of FSs is valuable for L2 learners for various reasons: fluency in communication (Schmitt, Jiang, & Grabe, 2011; Wood, 2006); facilitation effects of language processing (Conklin & Schmitt, 2008); and improving reading comprehension (Kremmel, Brunfaut, & Alderson, 2015). Thus, an important question in SLA and vocabulary learning research is how FSs can be learned and taught.

Nation (1990) states that the manner in which vocabulary is learned affects the learning burden, which is applicable to both individual words and FSs. Newton and Nation (1997) state that vocabulary can be taught through two main approaches: direct and indirect, which are not necessarily alternatives, though they do complement each other. A direct teaching approach implies explicitly directing learners' attention towards vocabulary. This explicit attention can be imparted through vocabulary lessons, which certainly includes explicit vocabulary activities such as word-building exercises, form-meaning(s) matching, split-information activities, semantic mapping, and pushed-output activities (Newton & Nation, 1997). An indirect approach incorporates vocabulary learning into meaning-focused communicative activities, with no direct focus on vocabulary, such as listening to stories, group work, and graded reading. The main difference between the two approaches lies in the manner in which lexical items are learned; that is, in a direct approach, it is probable that deliberate learning will occur, whereas in an indirect approach, incidental learning may occur. Deliberate learning refers to actively directing learners' attention towards lexical items (Schmitt, 2000), whereas defined incidental learning is "language learning as a by-product of language use by the teacher or anybody else in the classroom, without the linguistic structure itself being the focus of attention or the target of teaching manoeuvres" (Wode, 1999, p. 345). These two approaches may be pertinent to both individual words and FSs. Despite the fact that most vocabulary learning studies have focused on investigating the effectiveness of both approaches for SWs, the last decade has witnessed a considerable increase in the number of studies investigating the teaching and learning of FSs. Empirical evidence suggests that either incidental learning (e.g., Laufer & Girsai, 2008; Pellicer-Sánchez, 2015; Sonbul & Schmitt, 2013; Szudarski, 2012; Szudarski & Carter, 2016; Webb, Newton, & Chang, 2013) or deliberate learning (e.g., Boers, Dang, & Strong, 2017; Boers, Demecheleer, Coxhead & Webb, 2014; Chan & Liou, 2005; Jones & Haywood, 2004;

Le-Thi, Rodgers, & Pellicer-Sánchez, 2018; Nassaji & Tian, 2010; Peters & Pauwels, 2015; Sun & Wang, 2003) could be effective for learning FSs. Thus, more research is needed to have a better picture of the effectiveness of different approaches to the learning of FSs, and how this effectiveness compares to the learning of SWs. The following sections provide an overview of available studies examining the learning of FSs via these two approaches.

2.2.4.1. Incidental learning of FSs

A few studies have investigated the effectiveness of the incidental learning of FSs, with a focus on collocations: adjective-pseudoword collocations (Pellicer-Sánchez, 2015); verb-noun collocations (Szudarski, 2012; Webb et al., 2013); and adjective-noun collocations (Sonbul & Schmitt, 2013; Szudarski & Carter, 2016). Most of the available empirical studies have looked at incidental learning in different reading conditions: reading with repeated exposure to target items (Pellicer-Sánchez, 2015; Webb et al., 2013); reading while listening (Webb et al., 2013); reading only vs. reading plus explicit vocabulary activities (Szudarski, 2012); reading only vs. input-enhanced reading (Szudarski & Carter, 2016); enriched-input reading, enhanced-input reading, and decontextualized exercises (Sonbul & Schmitt, 2013).

Pellicer-Sánchez (2015), for instance, found that collocational knowledge can be learned by English as foreign language (EFL) intermediate learners incidentally from reading. The researcher asked learners to read a story containing six adjective-pseudoword collocations (e.g., *old holter* 'old workhouse', *small berrow* 'small bowl'). Two versions of the story were created, differing in the frequency of exposure to the target items (four vs. eight), and participants were assigned to one of these conditions. Learners' (n=41) collocational knowledge (i.e., form recall and recognition) was assessed through a combination of paper-and-pencil and interview tests. Although the results showed that

the learners' knowledge of pseudowords and their collocates was acquired incidentally from reading, no significant differences were found between the gains of four and eight occurrences, which contrasts with the findings of previous studies reporting that more repetition led to higher gains in terms of collocational knowledge (e.g., Peters, 2014; Webb et al., 2013).

Webb et al. (2013) studied the incidental learning of 18 verb-noun collocations through reading while listening. The materials they used varied in terms of the frequency of exposure to the target collocations, i.e., one, five, ten and 15 exposures. The authors used a pre-test to assess the receptive knowledge of form and four post-tests to assess: productive knowledge of form, receptive knowledge of form, productive knowledge of form and meaning, receptive knowledge of form and meaning. The results showed that the receptive gains in form of the target collocations were improved by 27%, 33%, 55%, and 76%, respectively. Learning not only improved their receptive skills but also their productive knowledge of form and meaning. As noted by Webb et al. (2013), the results need to be interpreted with caution due to the absence of a pre-test measuring three components: receptive knowledge of meaning, productive knowledge of form, and productive knowledge of meaning. Overall, this study showed that collocational knowledge can be learned incidentally from reading while listening.

Both Szudarski (2012) and Sonbul and Schmitt (2013) compared the effectiveness of reading-only and reading-plus treatment conditions on learning two types of collocations: delexical verb-noun collocations, e.g., *take a photo* (Szudarski, 2012), and medical adjective-noun collocations, e.g., *frosted heart* (Sonbul & Schmitt, 2013). The learners, in three groups (i.e., reading-only, reading with meaning-focused instruction, reading with form-focused instruction) in Szudarski's (2012) study, were exposed to the target items through reading texts (the frequency of exposure per collocation was six times) in a 45-minute session every week for three consecutive weeks. The reading-plus group

completed explicit vocabulary activities on collocational patterns: a cloze activity (week 1), a form-meaning activity (week 2), and a table-filling activity (delexical verbs were given, week 3). Three instruments were used: form recall of collocations, form recall of verbs, form recognition of verbs. Results showed that there was a significant improvement in reading for the form-focused instruction group on all three tests, whereas the improvement was significantly smaller in the reading with meaning-focused instruction-only group. The reading-only group did not seem to be different from the control group. In a later study, Szudarski and Carter (2016) examined two conditions: reading only and an input-enhanced reading condition for learning verb-noun and adjective-noun collocations; they found that the reading-only condition did not significantly improve learners' collocational knowledge.

In the case of Sonbul and Schmitt's (2013) study, three learning conditions were examined: an enriched condition (i.e., reading with the target items embedded three times); an enhanced condition (i.e., reading with embedded target items made salient in red, bold font); a decontextualized condition (i.e., explicit teaching of target items through a PowerPoint presentation). Three measures were used to assess collocational knowledge: an explicit form-recall test in a summary-cloze format (fill in blanks in a summary with target items and definitions glossed), an explicit form-recognition test in a summary-matching format (form-meaning multiple-choice activity in the margin), and an implicit priming paradigm test (the first word of the collocation was given as the prime and the second as the target, and learners had to decide whether the second word was a word or not). The results of the explicit instruments suggested that the enhanced condition was more effective than the enriched condition but not more effective than the decontextualized condition, whereas the results of the implicit instrument showed no improvement. More recently, Toomer and Elgort (2019) examined the development of implicit and explicit knowledge of medical collocations within three treatment conditions:

reading only (no typographic enhancement), bolding, and bolding-plus-glossing. The participants were sixty-two advanced English as a second language (ESL) speakers. The treatments included reading a text containing repeated occurrences of low-frequency medical collocations. Learners were assessed through two explicit knowledge tests: cued recall and form-recognition, and a primed lexical decision test to assess the implicit knowledge. The findings showed that repeated exposure to bolded collocations produced greater explicit knowledge than repeated exposure to typographically unenhanced collocations. Further, only the unenhanced (reading only) treatment provided evidence for the development of implicit knowledge but not the other two conditions.

Montero Perez, Peters, and Desmet (2018) investigated the effects of an enhancing video on learning vocabulary and examined exposure to YouTube videos on L2 vocabulary acquisition. The target items in their study included both single-words and FSs; however, no separate analysis was conducted to differentiate between the two (i.e., single-words and FSs). They investigated different types of subtitling: no captioning, full captioning, keyword captioning, and glossed keyword captioning which provides access to meaning, in two conditions: informing vs. not informing learners that a vocabulary test would follow the video. Participants were 227 Dutch-speaking university students. The results showed that the glossed keyword captions with access to meaning group outperformed the other groups on form recognition and meaning recall, and also revealed that looking up a given word was positively related to learning that word.

Research on the incidental acquisition of FSs has been scant but recent studies have shown that it is possible to learn FSs through reading, reading while listening, and viewing. The factors affecting this seem to be frequency of exposure and input enhancement, and most studies have looked at the acquisition of collocations.

2.2.4.2. Deliberate learning of FSs

As mentioned earlier, the majority of studies exploring the learning of FSs have focused on examining the effectiveness of deliberate approaches. Many studies have investigated the explicit learning of FSs (specifically collocations) through different approaches and activities: the use of corpus tools (Chan & Liou, 2005; Y.-C. Sun & Wang, 2003); inductive vs. deductive approaches (Y.-C. Sun & Wang, 2003); the use of dictionaries (Laufer, 2011); fill-in the blanks (Boers et al., 2017, 2014); reading and writing activities (Jones & Haywood, 2004); a contextualised vs. decontextualized manner (Le et al., 2018); receptive (presenting FSs with translation) and productive (cloze) tasks (Webb & Kagimoto, 2009); sentence writing (Zhang, 2017); recognition activities vs. cued output activities (Peters & Pauwels, 2015); and pushed-output instruction (Nassaji & Tian, 2010). A selection of these studies, mainly receptive approaches, will be reviewed in turn. Studies looking at productive approaches to FSs will be reviewed in Section 2.3.3.4.

The knowledge of verb-noun collocations was investigated through the use of a Web-based Chinese-English bilingual *concordancer* (keyword retrieval program) in Chan and Liou's (2005) study. The subjects were 32 EFL college students, and they were taught collocations using five Web-based units: three units incorporating a Chinese-English concordancer and two units without the use of a concordancer. In the three units, learners could read various examples in concordances and were exposed to different language patterns in verb-noun collocations. In the other two units, collocations were taught using pattern explanations, examples and exercises, such as a semantic field grid for synonymous verbs. All five units included some interactive exercises such as multiple-choice (MC), Chinese-English sentence translation, and gap-filling sentences. Learners were assessed through the use of a pre-test and two post-tests (immediate and delayed). The results showed that collocation knowledge of learners significantly improved after

the treatment but decayed later. Similarly, Sun and Wang (2003) reported the superior effectiveness of a *concordancer* when used with an inductive approach, i.e., use first then explain the rules, rather than a deductive approach, i.e., explain the rules first and then do examples, as it promoted the possibility of learning through a discovery process.

Explicit learning of FSs (verb-noun collocation) through dictionary use was investigated in Laufer's (2011) study. The participants (n=95) were intermediate high school EFL learners who used dictionaries to complete an assigned fill-in-the-blanks task (the missing item was a verb in each collocation). Learners were assessed through a pre-test and a post-test of their recall knowledge of target collocations. The results showed that learners' collocational knowledge improved significantly with dictionary assistance. However, Laufer (2011) recommended supplementing it with form-focused instruction to raise learners' awareness of collocations, as some learners in the study were not aware of collocations (unfamiliar structure) or were unable to find the target collocations in dictionaries.

Another type of explicit activity extensively examined in previous studies is fill-in-the-blanks activities. Boers et al. (2014), for instance, examined learners' knowledge of verb-noun collocations in four trials through the use of a noun-verb matching exercise, in the form of gap-filling, which mimicked the format of exercises found in their course books. Knowledge of collocations was assessed prior to the treatment through a form (verb) recall test. After the treatment, learners received teacher-guided corrective feedback and completed a delayed post-test two or three weeks later. As the authors acknowledged, only marginal learning gains were observed in the four trials, which suggests that this type of exercises leads to more confusion than learning. In a later study, Boers et al. (2017) examined fill-in-the-blanks exercises in three formats: choose the appropriate verb, complete the verb (a first-letter clue was given), and choose the whole phrase, and the results revealed that the third format, where learners worked with intact

phrases, performed better than the other two. The authors concluded that teachers should give preference to holistic-unit exercises for teaching target collocations (or multiword units in general) in order to enhance learners' awareness of the makeup of collocations, and also to minimise the risk of any cross-associations.

The receptive and productive knowledge of verb-noun collocations was compared in both Webb and Kagimoto's (2009) study and Zhang's (2017) study. With regard to Webb and Kagimoto's (2009) study, the subjects were 145 EFL Japanese learners, and they were assessed through four tests: productive knowledge of collocations, receptive knowledge of collocations, productive knowledge of meaning, receptive knowledge of meaning. The treatment included two conditions: receptive and productive treatments. In the receptive treatment, learners encountered each collocation with its L1 meaning and three sentences and they were required to read and understand the given sentences. In the productive treatment, learners encountered each collocation in sentences identical to the ones in the receptive treatment, but the collocations were replaced by blanks and learners were asked to write down the collocations. The results revealed that both the receptive and productive treatments were similar in terms of their effectiveness for learning the target collocations; however, when examining the results by learners' level of proficiency, the productive treatment was more effective for learners with a high level of proficiency, while receptive treatments worked better for low-level learners.

Similarly, Zhang (2017) investigated three tasks: a receptive task, a productive task, and a receptive-productive integration task for learning verb-noun collocations in terms of form, form and meaning, and grammar. In the receptive task, learners encountered each target collocation highlighted in bold with its L1 translation, followed by a glossed sentence. In the productive task, learners were provided with each collocation and its L1 translation, and they were required to write each collocation in four different sentences in English. In the receptive-productive integration task, learners were required to first read

each collocation with its L1 translation and a glossed sentence, and then write a sentence using it. Zhang (2017) reported that an integration approach using a receptive-productive task performed better than only receptive or only productive approaches in terms of both the immediate and long-term gains of verb-noun collocational knowledge (form, form and meaning, grammar). Zhang (2017) concluded that some productive activities could be beneficial for improving the receptive and productive knowledge of collocations (and FSs in general), either by themselves or incorporated into receptive tasks, which is relevant to the present study. This conclusion supports the use of more output-based activities in the classroom; yet, this is only concerned with written output.

The aforementioned studies have shown that several types of explicit teaching activities can improve learners' knowledge of FSs; however, this explicit learning could be affected by several intralexical and interlexical factors (Nesselhauf, 2003; Peters, 2016; Webb & Kagimoto, 2009). Peters (2016), for instance, examined the effect of congruency, type of collocation (i.e., adjective-noun, verb-noun, and phrasal verb-noun), and length of collocations on the learning burden of collocations. EFL learners (N=41) were asked to read a word list containing the target collocations (N=18), their translation and a sample sentence. Then, they were asked to complete four online form-focused activities in which the 18 collocations were presented twice: two fill in the blanks activities and the other two activities asking learners to provide a synonym, an antonym, a superordinate or a hyponym for the collocations. Learners were assessed using a form-recall test, a form-recall test with cues, and a form recognition test. The results showed that the length of collocations significantly affected their learning, whereas congruency was only important at the form-recall level, but not in the form-recognition task. The study also showed that adjective-noun collocations were learned better than the other types of collocations. To sum up, the study showed that apart from repetition, there are

other factors that also influence the learning of collocations such as congruency, length, and type of collocation.

Overall, these studies seem to suggest that FSs can be substantially learnt through multiple explicit activities, such as the use of corpus tools, inductive vs. deductive approaches, cloze tasks, and so on. These few available studies reviewed above, examining the deliberate learning of FSs through receptive approaches, focused on collocations as a type of FSs. None of these studies examined the learning of phrasal verbs through deliberate approaches.

2.2.4.3. SWs and FSs: same or different?

Research has shown that learners' knowledge of FSs often lags behind their knowledge of SWs (e.g., Bahns & Eldaw, 1993; Granger, 1998; Nesselhauf, 2003, 2005). Further, studies have shown that learning FSs was more challenging for learners due to the fact that FSs are intrinsically different from SWs on multiple aspects: complexity, holding figurative meaning, etc. If FSs are indeed more difficult to acquire than SWs, one would expect that the learnability of FSs might require different type and amount of exposure. Further, one important question in the learning of FSs would be whether the teaching approaches reviewed above would have a similar effect on the acquisition of SWs and FSs. Very few studies have experimentally compared the acquisition of FSs and single words. The few available studies suggest mixed results (Alali & Schmitt, 2012; Kasahara, 2011, 2010; Laufer & Girsai, 2008; Peters, 2012, 2014). Therefore, one of the central aims of the present thesis is to investigate differences in the acquisition of SWs and FSs, represented in the present studies by phrasal verbs.

One of the few studies that have empirically examined the acquisition of SWs vs. FSs is Alali and Schmitt's (2012) study. It explored the differences between SWs and idioms using deliberate focused instruction. The explicit teaching approaches examined

here were: encountering target idioms with translations, encountering target idioms with translations and oral repetition, encountering target idioms with translation and fill-in-the-gaps activities. Learners were 35 EFL Arabic speakers and their knowledge of four components was assessed: form recall, meaning recall, form recognition, and meaning recognition. The results showed no differences between SWs and idioms at the recognition level, but significant differences emerged at the level of recall. Learning gains for idioms were significantly lower than those for SWs at the recall level of mastery.

Laufer and Girsai (2008) examined the learning of verb-noun collocations and SWs through the use of meaning-focused instruction and two forms of focused instruction. The meaning-focused instruction included a reading-plus condition with two communicative activities: reading comprehension and pair/group discussion. The form-focused instruction incorporated reading-plus text-based vocabulary activities in one condition, and reading-plus text-based translation activities in the other. Two instruments were used: a passive recall test (meaning-recall test), and an active recall test (form-translation test). They reported that, although the gains in the meaning-focused group were quite low, indicating no learning, the pattern of these gains indicated that collocations were better learnt than SWs in both tests. The results also showed that the two form-focused groups outperformed the meaning-focused group in tests for both SWs and collocations.

Kasahara (2010, 2011) examined the acquisition of SWs and collocations, and the treatment involved memorising a list of 20 Japanese meanings, the equivalents of low-frequency unknown SWs or low-high frequency collocations. The subjects were 39 high-school EFL learners. Two recall tests were administered at two different times (immediately and one-week later): the meaning-recall of only SWs (removing the cues from the collocations), and a second one that included all the items as they were presented in the list, that is, SWs and collocations. Kasahara (2010,2011) found that the written meaning recall, and retention of collocations was significantly higher than for SWs, which

is in line with Laufer and Girsai (2008). The results of both the immediate and delayed tests showed that the retention and retrieval of meanings of collocations was better than for SWs.

Peters (2014) also compared the acquisition of SWs and collocations. The subjects were 35 EFL learners studying business at a Flemish university. They were taught 24 items: 12 SWs and 12 collocations, which were divided into three groups based on their repetition in the materials: one repetition, three repetitions, and five repetitions. The treatment included presenting a word list of the target items with their L2 definitions, followed by eight written decontextualized vocabulary exercises: providing synonyms, antonyms, hyponyms, or superordinates; matching collocations; filling in the gaps in sentences; retrieving the target item when given a definition. The instruments used for assessment were: a form recall pre-test and two form recall post-tests, administered either immediately or one to two weeks after the treatment. Peters (2014) reported that the written form recall of SWs two weeks after the treatment was higher than the recall of collocations, suggesting that collocations were more difficult to learn than SWs, which contradicts the findings in Alali and Schmitt's (2012) study. The results concerning frequency of occurrence revealed that more repetitions resulted in higher recall scores for both SWs and collocations.

Overall, previous studies comparing SWs with FSs have suggested mixed results. This is not surprising, considering that different types of FSs were examined, and different approaches of explicit teaching were used. Pellicer-Sánchez (2020) has suggested that the learnability of SWs and FSs may be similar if both receive the same amount and type of exposure. There has been a limited amount of pedagogical advice for FSs and there is a clear need to examine the effectiveness of the different approaches that we use for SWs for teaching FSs.

The literature reviewed so far has shown the importance of FL knowledge (and of different types of FSs) as a key component of vocabulary knowledge, and the different approaches that have been used to teach and learn different types. Studies have shown that FSs can be learned incidentally from reading, reading while listening and viewing, as well as from a range of explicit activities. Factors such as frequency of exposure and input enhancement have been shown to play an important role in this learning. The available evidence examining whether FSs are similarly learned from the same amount of exposure as SWs is still limited. The studies reviewed above mainly focused on written instruction approaches. However, learning through spoken output has received less attention from researchers and it is clearly not a common component in many classroom contexts, like the one investigated in this thesis. The following sections presents the conceptualisations of pushed-output instruction for the present thesis, including an overview of the theoretical frameworks that have lend support for this instructional method: the Output Hypothesis, the Interaction Hypothesis and Transfer-Appropriate Processing theory, and reviews how pushed-output approaches have been applied to the learning of vocabulary and FSs (and phrasal verbs) in particular.

2.3. Learning vocabulary through output

This section focuses on the acquisition of language features, and particularly vocabulary, through output. I will first introduce the construct of pushed-output instruction. I will then review the theoretical frameworks that have lend support for the benefits of instruction that is based on pushed-output. Finally, I will review empirical research that has evaluated the effectiveness of output for vocabulary learning.

2.3.1. The construct of pushed-output instruction

As Nation and Newton (2008) argue, knowledge of the L2 does not automatically transfer from reception to production and learners might need to be 'pushed' to produce the language. One reason for this "push", as Robinson (2011) asserted, is the noticing function, as proposed by Swain (see Section 2.4.2.1). Being pushed to produce language may mean forceful speech, yet it can also be related to L2 learners being engaged in a stimulating environment encouraging speech. This means the role of the instructor is not to force the learner, but to create a peer encouraging environment where the L2 learner sees language learning as a joint problem-solving process. When they are "pushed" to produce speech, orally or written, they begin to notice gaps in their language knowledge. Prabhu (1987) noticed that a competent L2 learner in progress, with the right motivation skills, would consequently be encouraged to upgrade their interlanguage system using internalised recourses from the dialoguing process. Additionally, the "push" would support an upgrade from a low-level system of speech production to a higher rank of internalised structures, which in turn would contribute to fluency development and coherence (Swain, 1995). It is imperative to understand how teachers can enhance the extra "push" in an L2 learning environment.

Nation and Newton (2008) illustrated several factors, particularly in relation to vocabulary learning, that help in developing the "pushing" element of an activity. These factors include covering a range of topics, a range of text types and a range of performance conditions. Learners should be pushed to speak about a range of topics, or at least one unfamiliar topic. The topic is most likely to affect vocabulary learning as each topic has its own particular technical, topic-related vocabulary (Nation & Newton, 2008). Covering a good range of topics in a course ensures a wider range of vocabulary being learnt. Topic is also related to the amount of background information learners have, as some learners may be more familiar with the content of one topic than another. Further, text type is also

considered a significant factor in developing a pushed-output activity. Biber (1989) distinguished eight major text types on the basis of clustering largely grammatical features. These text types included interpersonal interaction, scientific exposition, imaginative narrative, and involved persuasion. Even though most of these types are merely related to written output, they have some spoken equivalents. To ensure learners are pushed to cope with a range of text types, the most useful distinctions proposed by Nation and Newton (2008) are: involved interaction versus monologue, colloquial speech versus formal speech, short turns versus long turns, interactional versus transactional speech, and narrative versus non-narrative speech. The third factor to consider in developing the "pushing" element is the use of a variety of conditions of performance, such as planning, time pressure, amount and type of support and feedback, and standards of performance expected to be produced. Such conditions could help in strengthening the "pushing" element of an activity, and perhaps result in learners being pushed to produce language beyond their level.

Further, beside the 'pushing' element, Swain (2005) also emphasised the role of learners' talk while completing the activity in language learning, the metalinguistic function. This highlights the fact that interaction is one important factor for output to be a facilitative factor for language learning. A pushed-output activity should be interactive, and communicative. Interaction refers to 'the conversations that learners participate in' while learning (VanPatten & Williams, 2014). These interactions are important for learning in the classroom as learners receive information about the accuracy, or more importantly, the inaccurateness of their production (VanPatten & Williams, 2014).

Laufer (2005) argued that the pushed output hypothesis (Swain 1985, Swain and Lapkin 1995) can be regarded as part of the theoretical underpinning of Focus on Form (FonF). FonF was first defined by Long (1991) as incidentally drawing learners' attention to the linguistic elements as they arise in lessons. Advocates of FonF believe that L2

learners need to attend to linguistic elements during a communicative activity to achieve high levels of grammatical competence (Long 1991, De Keyser 1998, Norris and Ortega 2000, Ellis 2001). Laufer (2005) argued that this idea of FonF is not applicable for only grammatical instruction but can also be applied to lexical instruction. She claimed that "attention to form must be motivated by and carried out within a communicative task environment" (Laufer, 2005, p. 223). Another related pedagogical approach is Focus on Forms (FonFs) in which linguistic elements are taught in separate lessons in a sequence determined by syllabus writers (Laufer & Girsai, 2008). Ellis (2001) better illustrated the differences between FonF and FonFs in which he viewed students in FonF as language users and the language as just tool for communication, whereas in FonFs learners view themselves as learners and the language is the subject they study. FonF can be related to three hypotheses: noticing (Schmidt, 1990, 1994), limited processing capacity (VanPatten, 1990), and 'pushed output' (Swain, 1985), as stated by Laufer (2006). Learners notice the forms and their meanings; attend to the meanings while communicating; and are 'pushed' to use more syntactic processing mode than they would in comprehension. On the other hand, FonFs could be theoretically justified in terms of the skill acquisition theory in three stages: declarative knowledge, proceduralised knowledge (i.e., what to do with the language), and automatization of procedural knowledge (i.e., using language according to certain rules without thinking about them) (Laufer, 2006). In relation to the present thesis, pushed output activities resembles FonFs in the way that direct attention will be paid to the taught vocabulary and learners were aware that they need to learn specific set of items within the activity. However, it could be also considered as an example of FonF since one of the main elements of the activities is to be communicative, particularly in activities such as picture description, sentence reconstruction activities. However, it must be acknowledged that the amount of communication involved in the pushed-output activities can be considered limited

compared to the one within FonF instruction, as described by Laufer (2005). Learners can engage in different forms of collaborative behaviours while completing the pushed-output activities, ranging from simple repetition to testing their own hypothesis about the appropriateness of words and different meanings.

While pushed-output includes both written and spoken pushed-output activities, in this thesis I am focusing on spoken modality. Spoken pushed-output vocabulary instruction in the present thesis is defined as activities that encourage learners to use vocabulary orally while completing the activity, and that allows exchanging information that learners received with their peers, either in a very simple form such as repetition, or more advanced form such as rephrasing or negotiating the meaning. Further, certain performance conditions were implemented in the designed activities to further develop the 'push' element such as time pressure, nature of feedback (pair-work), and information distribution, as advocated by Nation and Newton (2008). Thus, the following features are characteristics of the pushed-output instruction examined in the present thesis: interactive (through the need to interact with peers in each activity), communicative (through the need to exchange information and give feedback), and of spoken modality that are implemented under time pressure.

2.3.2. Theoretical support for pushed-output instruction

In this section, I review the relevant theoretical frameworks that lend support for the benefits of pushed-output instruction including the Output Hypothesis, the Interaction Hypothesis, and the Transfer Appropriate processing theory. Particular emphasis is placed on the Output hypothesis since it is the first hypothesis to emphasize the role of production in language learning, and also because most vocabulary studies researching output instruction are built around this hypothesis.

2.3.2.1. The Comprehensible Output Hypothesis

The need to provide opportunities for speaking was reflected in Swain's (1985) Comprehensible Output Hypothesis. To begin with, output can be demonstrated in the form of speaking or writing in the language classroom. The word output was synonymous with the product of a learning process in the 1980s and was used in SLA research to demonstrate what learners had learnt (Swain, 2005). However, following Swain's (1985) paper on the role of output in the SLA process and proposal of the Output Hypothesis, the term shifted in meaning. Output as a product shifted to output as a process and action, playing an important role in L2 learning. In this section, the origins of the Output Hypothesis is discussed, including the potential functions output has in the learning process. Research evaluating the Output Hypothesis is briefly discussed and some criticisms of this viewpoint are also explored.

Origins of the Output Hypothesis

Numerous SLA theories have sought to clarify, challenge, or develop beliefs concerning the acquisition process. For instance, in the 1980s, the dominant theoretical paradigm for SLA was information processing theory, where input was given a predominant role. This paradigm was characterised by Krashen's thoughts on the Input Hypothesis (1981, 1984), which stated that the only necessary condition for SLA was the presence of comprehensible input. More specifically, Krashen maintained that, if learners were at stage i in their learning development, they advanced to stage i + 1 through the comprehension of input that contains i + 1. According to Krashen’s Input Hypothesis, comprehensible input guides the learner towards more natural acquisition of the language, rather than consciously learning it. Hence, several authors have suggested ways in which input could be made comprehensible for learners. For example, Long (1985) proposed clarification requests and comprehension checks. Pica (1994) discussed the importance

of negotiation for achieving comprehensibility, such as repetition and rephrasing, as well as modifying and restructuring interactions to increase input comprehensibility.

In line with the input theory of SLA, several French immersion programmes for school children were developed in Canada. English-speaking children were taught all or some of the curriculum in French in these programmes, in early immersion (i.e., the start of schooling), mid-immersion (i.e., after four years of schooling), or late immersion (i.e., after six years of schooling) (Swain, 2005). These programmes were input-rich and therefore served as a test of Krashen's hypothesis that comprehensible input is "the only true cause of second language acquisition" (Krashen, 1981, p. 61). The results of tests measuring children's skills provided a number of insights. First, the immersion children scored higher on French tests than children who studied French as a foreign language for 20-30 minutes per day. Furthermore, some of the immersion children scored as high as native French-speaking children on some of the French listening and reading tests, providing evidence of the benefits of immersion programmes and comprehensible input (Swain & Lapkin, 1982). However, the speaking and writing skills of the immersion students differed from native French-speaking children. The test results indicated that there was a lack of speaking and writing skills for children who had spent many years in immersion programmes (Gass, 2003).

Swain compared the results of a number of different grammatical, discourse, and sociolinguistic measures from sixth-grade children in the French immersion programme and sixth-grade native French-speaking children. The results showed a general lack of proficiency in all the measures in the immersion children. Swain questioned the idea of input being the only necessary element for acquisition and began to search for alternative explanations of SLA. She postulated that input alone could not account for SLA and that output also has a role in SLA. This conclusion was based on her formal and informal observations of the French immersion classrooms in Canada. She noticed that children in

the immersion context did not have the same opportunities to productively use the French language as they did in English, and they were merely using this language for comprehension. Furthermore, Swain noticed that immersion students spoke more in the English portion of the day than in the French portion. More importantly, the teachers did not require the students to use accurate grammar or socio-linguistically appropriate language (Swain, 1985). Swain (1985) developed the Output Hypothesis and, in her first paper, referred to the need for "comprehensible output" in language learning. It is important to note that Swain did not claim that output is the only source of SLA but rather that output promoted language acquisition under certain conditions. Swain called it comprehensible output, drawing a parallel with Krashen's i + 1 comprehensible input. Comprehensible output refers to an extension of the usual meaning of getting one's message across, whereby messages are conveyed using grammatically deviant and socio-linguistically inappropriate language. Comprehensible output has been also been labelled by some researchers as pushed output. This might be because Swain argued that effective L2 learning depended on language instructors and interlocutors "pushing" the learner to achieve a higher level than their comfort zone. Push, in this sense, portrays coercion towards competence enhancement (Swain, 1985). Swain suggested that learners need to be 'pushed' to produce language that conveys messages in a precise, coherent and appropriate manner.

As Swain (1985) proposed, the process of "pushed output" occurs during the negotiation of meaning in interactions with others. Mackey (2002) provided empirical evidence for its existence by asking adult English as a second language (ESL) students to watch videos of themselves interacting with others and to recall their thoughts when the original interaction occurred. In the following example from Mackey (2002), an interaction occurred between NS and NNS. The NNS is pushed to modify their output to make it clearer:

NNS: And in hand in hand have a bigger glass to see.

NS: It's err. You mean, something in his hand?

NNS: Like spectacle. For older person.

NS: Mmmm, sorry I don't follow, it's what?

NNS: In hand have he have has a glass for looking through for make the print bigger to see, to see the print, for magnify.

NS: He has some glasses?

NNS: Magnify glasses he has magnifying glass.

NS: Oh aha I see a magnifying glass, right that's a good one, ok.

Recall by NNS: "In this example I see I have to manage my err expression because he does not understand me and I cannot think of exact word right then I am thinking it is nearly in my mind, thinking bigger and magnificate and eventually magnify. I know I see this word before but so I am sort of talking around around this word but he is forcing me to think harder, think harder for the correct word to give him so he can understand and so I was trying, I carry on talking until finally I get it, and when I say it, then he understand it, me."

Mackey (2002, pp. 389- 390)

As can be seen in the recall comments, the learner was able to recognise that their comments were not clear and that the NS was pushing them to produce language that was clearer, more articulate, and more appropriate.

In later papers on the Output Hypothesis, Swain stopped using the term comprehensible output because Swain considered it misleading: it obstructs the idea of output as a process rather than a product (Swain, 2005). Swain claimed that this

misconception has led researchers to focus on the product of learning in two main ways. First, it has been taken literally and out of context, so that the term comprehensible is understood to mean the ability to be understood rather than a process of improving output in terms of its informational content. Second, Swain (2005) believed that the term comprehensible output, as a noun, emphasised the product of learning rather than the process of how learning occurs. As a result, Swain has recently re-emphasised how the Output Hypothesis views output as part of the learning process, rather than a product that is synonymous with what the learner has learnt.

To conclude, Swain's proposal of output as a factitive factor for SLA to occur does not neglect the significance of input as one source for learning to occur but rather imply that output could be incorporated along with input for learning to be maximised.

How does output play a role in the learning process?

Swain (1995) proposed three main ways in which output may play a role in the L2 learning process: (1) a noticing/ triggering function, (2) a hypothesis testing function, and (3) a metalinguistic (reflective) function. These are examined in turn. It is worth mentioning that all the studies used in this section are related to language learning in general (and not to vocabulary in particular).

The noticing/triggering function. One important focus of SLA research has been on examining cognitive processes in L2 learning, particularly noticing, which refers to learners' attention and awareness of language (e.g., Gass & Varonis, 1994; Robinson, 2011; Richard Schmidt & Frota, 1986). According to the Noticing Hypothesis (Schmidt, 2001), learners must consciously notice input for it to become intake. That is, if output is shown to promote the noticing process, it follows that output may be helpful in the acquisition process. But what do learners notice that can help in the learning process? There are several levels of noticing.

The first level where noticing may take place is if something in the target language is frequent or salient (Gass, 2017). Learners may notice this language form and use it in their own output. The second level was proposed in the "notice the gap principle" by Schmidt and Frota (1986). This principle suggests that learners may notice form in the target language and that it is different from their own interlanguage. This can focus learners' attention on parts of discourse that are problematic from a productive point of view. In other words, learners may notice gaps between what they produce and what is produced by ESL (Gass & Varonis, 1994). The third level where noticing may occur was suggested by Swain (1995), when learners notice a hole in their interlanguage that needs rectifying. That is, learners may notice that they do not know how to express a meaning at the moment when they are attempting to produce it. In other words, the activity of producing output may prompt learners to consciously recognise that they have a linguistic problem. This may in turn prompt learners to find a solution to this problem, possibly by directing their attention to relevant input, such as a dictionary or another person (Swain, 2005). This awareness of a linguistic problem may also trigger cognitive processes that have been suggested to contribute to SLA, such as those allowing learners to generate linguistic knowledge that is new or that consolidates their existing knowledge (Swain & Lapkin, 1995).

Previous studies have reported a positive effect of noticing on language learning (e.g., Abdalla, 2014; Izumi, 2013; Mackey, 2006; Mennim, 2007; Mirzaee & Ranjbar, 2012; Swain & Lapkin, 1995). For instance, Swain and Lapkin (1995) reported that output not only motivates language acquisition through activating various internal processes but also promotes conscious noticing of linguistic problems in the learner's interlanguage knowledge. They argued that learners were able to notice gaps in their existing interlanguage capacity, which then pushed them to reprocess their performance consciously to produce modified output instances. Research has shown that noticing

while producing language can be encouraged through either external feedback (e.g., clarification requests) or internal feedback (e.g., internal noticing) (e.g., Swain & Lapkin, 2001; Shehadeh, 2002). Not only these but also a number of contextual factors, such as gender differences, task type, and type of interlocutor, promote learners' reformulating of their output towards comprehensibility in an attempt to fill their interlanguage gaps (e.g., Pica, 1988; Pica, Holliday, Lewis, Berducci, & Newman, 1991; Pica, Lincoln-Porter, Paninos, & Linnell, 1996; Shehadeh, 2002, 2003; Van den Branden, 1997). It might be argued that noticing should not be considered as involving a deep level of quality of attention; yet, it is still relatively important as it determines what will be learnt (Barcroft, 2006).

The hypothesis testing function. The hypothesis testing function of output is based on the claim that output, from the learner's perspective, may be a trial run for producing language, reflecting their hypotheses about how language is used (Swain, 2005). Learners test out new language that they have encountered and determine whether these new forms are comprehended by their interlocutors. During an interaction, learners may notice a problem with their speech, either through self-introspection or because an interlocutor communicates their misunderstanding. Then, the learner will pay closer attention to this language form. They may, as a result, test out a new hypothesis about how the language is used by reformulating their utterance to make themselves better understood (Choi & Kilpatrick, 2014).

Research has shown that learners often expect to receive feedback on the language they produce, and they make changes to their output based on this feedback (e.g., Chaudron, 1988; Gass, Behney, & Plonsky, 2013; Nobuyoshi & Ellis, 1993; Skehan, 1998). For instance, Gass et al. (2013) pointed out that the feedback learners receive from their interlocutors played an important role in helping them to determine whether their hypotheses were correct. Pica, Holliday, Lewis, and Morgenthaler (1989),

for example, found that, in a laboratory setting, over one third of learners' utterances were modified in response to feedback. Furthermore, Loewen (2002) found that, in a classroom setting, almost three-quarters of learners' utterances were modified as a result of teachers' feedback. According to Swain (2005), the differences in the levels of modified output in these two studies can be attributed to their different settings. Learners would probably feel more comfortable testing hypotheses in their own classroom setting rather than with a stranger in a laboratory setting. If they were not testing hypotheses, then they would arguably not make changes to their output following the feedback they receive (Swain, 2005). Nobuyoshi and Ellis (1993) argued that interlocutor feedback is a crucial part of language learning because learners either confirm or disconfirm their hypotheses about the target language based on this feedback. Similarly, Chaudron (1988) argued that learners use feedback to readjust their interlanguage. However, some studies have shown that interlocutors often ignore hypothesis-testing episodes when they are ungrammatical but comprehensible, which may lead learners to believe that their hypotheses are correct and their utterances grammatically correct when they are not (Choi & Kilpatrick, 2014; Shehadeh, 2003). Thus, learners may require a greater push towards modifying their output when it is incorrect so that successful acquisition may take place.

For hypothesis testing to take place, a number of conditions are necessary, including alertness, attention, and awareness (Choi & Kilpatrick, 2014). Alertness is characterised by a learner's receptiveness to the input they receive (Tomlin & Villa, 1994). It means that the learner is ready for input and is in a state where they are cognitively capable of processing incoming information. A learner's alertness is the first step towards attention, whereby learners focus on a specific part of the input they need as feedback about their hypothesis (Tomlin & Villa, 1994). The notion of awareness is also important

for hypothesis testing, as awareness is required so that learners may detect problems with their utterances and subsequently try out different forms of modified output.

There are a number of different explanations as to why hypothesis testing and modified output may contribute to SLA. In an examination of learners' think-aloud protocols when producing a written text, Swain and Lapkin (1995) found that the communicative need of a task pushed learners to think about their linguistic output and forced them to transition from semantic to grammatical language processing. Swain and Lapkin (1995) suggested that, as a result, those learners were able to apply their existing knowledge to new situations and internalise new language rules. Conversely, De Bot (1996) argued that the benefit of output and hypothesis-testing for language learning stemmed from learners' increased control over language forms and their increased automaticity in processing as a result of producing output. This in turn frees up cognitive space for higher-level processing. Izumi (2002) further suggested that hypothesis testing in output may sensitise the learner to the possibilities of what can and cannot be expressed in the target language. While the explanation of how output contributes to the L2 learning process remains unclear, there is a general agreement that hypothesis testing, feedback as a result of this output, and subsequent modified output all contribute to the learning process in some way.

The metalinguistic (reflective) function. The metalinguistic function of output involves using output to talk about language as a means of mediating language learning (De Bot, 1996). The idea originates from Vygotsky's sociocultural theory of the mind, wherein people operate with mediating tools, such as language (Swain, 2005). Through speaking, dialogue and cooperation with others, learners are able to engage in psychological processes that subsequently become internalised. That is, learning takes place through dialogue with others (Donato & Lantolf, 1990). Using this theory, Swain (1995) and Swain and Lapkin (1995) developed a number of classroom activities that

asked learners to work together in pairs to solve language-based problems in a target language. Through discussions of problems, learners arguably become more aware of them and provide each other with further relevant input based on their shared knowledge or lack of knowledge (De Bot, 1996). This can be seen in the following extract from Swain and Lapkin (1998), where two learners are participating in the shared task of writing a story based on a set of picture prompts:

LL1: et brosse les cheveux.
(and brushes her hair)
LL2: et les dents.
(and her teeth)
LL1: Non, non, pendant qu'elle brosse les dents et …
(No, no, while she brushes her teeth and…)
LL2: Elle se brosse … elle SE brosse
(She brushes … she brushes)
LL1: Pendant qu'elle se brosse les dents et peigne les cheveux.
(While she brushes her teeth and combs her hair)
LL2: Ya!

(Swain & Lapkin, 1998, p. 331)

As can be seen in this example, the dialogue focuses the learners' attention on the particular language points of the task and provides opportunities for each member to offer alternatives and provide input for their interlocutor. L1 shows L2 that the verb *brosse* should not be used with hair, and L2 draws L1's attention towards the reflexive nature of the verbs. Through the dialogue, they each regulate the other's activity and they are both provided with opportunities to reflect on their own language use (Swain, 2005). According to Swain and Lapkin (1998), their jointly constructed performance surpassed

their individual competencies; thus, the dialogue represents "collective cognitive activity which serves as a transitional mechanism from the social to internal planes of psychological functioning" (Donato, 1988, p. 8). In other words, the dialogue mediated the learners' learning and provided an opportunity for this learning to become internalised. Swain (2005) referred to this type of learning as learning through collaborative dialogue.

These three main functions of output, i.e. the noticing, hypothesis testing, and metalinguistic functions, were extensively examined for the learning of grammar features (e.g., Izumi, 2002; Izumi & Bigelow, 2000; Izumi, Bigelow, Fujiwara, & Fearnow, 1999; Morgan-Short & Bowden, 2006; Nobuyoshi & Ellis, 1993; Shehadeh, 2002, Swain, 1985, 1995: Swain & Lapkin, 1991, 2001). One might question if the same functions and theory behind it could be compatible with vocabulary learning. While these functions were originally tested in relation to the acquisition of grammar, empirical studies have also applied the output hypothesis and the different functions of output to the learning of vocabulary. Several studies have examined the noticing function in relation to vocabulary learning (e.g., Philp & Iwashita, 2013; Rassaei, 2017; Tahmasbi & Farvardin, 2017) and reported that noticing in a pushed-output vocabulary activity mainly involves the 'noticing the gap principle' in their interlanguage. That is, when completing the activity, learners may notice that they do not know how to express a meaning at the moment when they are attempting to produce it. Thus, learners notice a hole in their interlanguage that needs rectifying. The second function is the hypothesis testing, in which Swain (1995) advocated that learners have the opportunity to practise implementing a hypothesis about how a language works when pushed to use that language. This function has been examined in vocabulary studies (Mackey, 2002; Shehadeh, 2002) and they concluded that learners output included series of hypotheses representing their guesses about L2 forms/ meaning. Furthermore, evidence has also been obtained that metalinguistic talk occurring

within collaborative dialogue, learners' talking while completing the activities (i.e., language-related episodes), are mostly 'language learning in progress' (Swain & Lapkin, 1998, p. 321). Learners tend to retain the knowledge co-constructed through collaborative dialogue either about vocabulary or grammatical structures in classroom environments (e.g. Lapkin, Swain & Smith, 2002; Shekary & Tahririan, 2006; Storch, 2008; Swain & Lapkin, 2002; Watanabe & Swain, 2007; Williams, 2001; Zeng & Takatsuka, 2009).

Other functions of output. In addition to the aforementioned functions of output proposed by Swain (1995), other researchers have suggested additional functions. The first function is the output role in creating greater automaticity in language use (De Bot, 1996; Gass, 2003; Skehan, 1998). Automaticity refers to the ability to retrieve items from long-term memory during language processing with only minimal or no delay (Loewen & Reinders, 2011). Automatic processes develop as a result of "consistent mapping of the same input to the same pattern of activation over many trials" (McLaughlin, 1987, p. 134). In terms of L2 development, this suggests that a certain amount of practice is needed for language use to become routinised. Therefore, output contributes to automaticity by providing opportunities for practice.

Another essential function of output in relation to SLA process was proposed by Skehan (1998). Skehan demonstrated that output provides opportunities for learners to develop discourse skills, such as by producing "long turns". Moreover, Skehan asserted that output is important for helping learners to develop a personal voice by steering conversations towards topics that they are interested in contributing to. Ellis (2003) further added that output provides learners with "auto-input", whereby learners attend to the input provided by their own language production. Therefore, output performs a number of significant functions in the language learning process.

Swain's comprehensible output hypothesis was based in her research on immersion students who, despite the fact they were receiving considerable amount of

comprehensible input, their productive skill did not seem to develop to establish a native-like ones. Swain concluded that:

> The argument, then, is that immersion students do not demonstrate native-speaker productive competence, not because their comprehensible input is limited but because their comprehensible output is limited. It is limited in two ways. First, the students are simply not given—especially in the later grades—adequate opportunities to use the target language in the classroom context. Second, they are not being "pushed" in their output. (Swain 1985, p. 249)

Though Swain's main point of interest in her research was the grammatical accuracy of learners' production, these points included above, i.e., not having opportunities to use language and not being pushed in their output, could also be applicable to vocabulary learning (e.g., Folse, 2006b; Holster & DeLint, 2012; Hulstijn & Laufer, 2001; Keating, 2008; Nassaji & Tian, 2010; Rassaei, 2017; Tahmasbi & Farvardin, 2017; Webb, 2005; Zou, 2017). In the Saudi EFL context, the most commonly production-based activities used are those representing the grammar-translation (i.e., L1 -L2 translation exercises), or the audiolingual method (simple oral pattern drills and performance of memorized dialogues). So, the spoken pushed-output vocabulary activities as defined in Section 2.3.1, the main area of interest in the present thesis, are rarely included in the Saudi EFL vocabulary classroom.

2.3.2.2. The Interaction hypothesis

As I have explained in Section 2.3.1, interaction is an important element in the conceptualisation of spoken pushed-output instruction in the present thesis, as important functions of output are realised through interaction. The important role of interaction for language learning was highlighted in the Interaction hypothesis (IH). The IH was first

introduced by Long (1983) who claimed that input serves as the 'the linguistic forms (morphemes, words, utterances)—the streams of speech in the air—directed at the non-native speaker' (Long, 1983, p.127). According to Long (1996), the IH stated that second language acquisition occurred due to one's interactions with the external environment. Long (1996) argued that the IH assumed that for the productivity and performance of L2 acquisition to be enhanced, the learner had to interact with the language in the external environment such as through face to face interactions, verbal expressions, and others. Yee, Ning, and Hua (2017) stated that the IH began with Krashen's (1981) work on the Input Hypothesis. Krashen (1981) stated that crucial to any language learning process was the ability to understand the input being received by the student. Krashen (1981) also developed the interactional restructuring concept, which occurs when learners using L2 fail to understand the input. As a result, those L2 learners have to make repetition, clarification, and comprehension reviews amongst themselves to understand the information, which relates to the concept of feedback in the Interaction Hypothesis. These instances of feedback and modified input which learners are exposed to during interaction or negotiation of meaning is what makes IH different from Input Hypothesis (Lightbown & Spada, 1993).

> Negotiation for meaning, and especially negotiation work that triggers interactional adjustments by the NS or more competent interlocutor, facilitates acquisition because it connects input, internal learner capacities, particularly selective attention, and output in productive ways. (Long, 1996, pp. 451–452).

Gass (2017) argued that the IH focuses more on the environment and its influence on the L2 acquisition process. The IH also posited that through external interactions, L2 acquisition occurred efficiently rather than the individual internal traits of the students (Long, 1996). The advocates of IH also believe that interaction provides learners with

opportunities to receive input and feedback (Gass, 1997; Long, 1996; Pica, 1994) and also opportunities to amend their output (Swain, 1995).

Gass (2017) stated that the IH focuses on four crucial areas that include Input, Interaction, Feedback, and Output. Gass (2017) also argued that input referred to the information received by the student, and in this case, it is the L2. The output, as stated by IH, refers to the linguistic approaches used by the students to complete the input instruction they were given or respond to a discussion concerning the instruction given in a grouped setting. Relating to the spoken pushed-output instruction in the present thesis, the input would be input communicated by their peers. Further, the role of output and input is incorporated with negotiation of meaning to promote acquisition. Thus, this implies that negotiation of meaning should be incorporated in the language pedagogic tasks. Since Long's hypothesis was formulated, a body of research has examined the nature of activities which promote negotiation of meaning. For instance, Crookes and Rulon (1988) contended that closed activities (e.g., deciding on a candidate) require more negotiation of meaning than open activities (e.g., opinion sharing). Another study by Long (1990) established that two-way tasks and group work generate more negotiation of meaning. Skehan (1998) examined the role of planning in task performance and concluded that planning improves the occurrence of a more accurate and fluent production of language, as well as more genuine negotiation of meaning. Polio and Gass (1998) found a positive effect for negotiated interaction on SL production and comprehension.

> …. there is still no direct evidence to link interaction to acquisition and precious little to demonstrate that it promotes comprehension, …[however] the absence of supportive research does not warrant the abandonment of a hypothesis that, in many ways, has contributed substantially to our current understanding of how leaning takes place in the classroom context. (p. 19)

Al Khateeb (2015), examining the IH features in the Saudi context, argued that the interaction element of the IH referred to the way information is exchanged from one person to the next. Al Khateeb (2015) found that interaction occurred in a group setting where the information that was received via the input was negotiated amongst the students. During the occurrence of the interaction process, L2 acquisition can occur with improved productivity and performance as the students put the lessons into practice allowing their memories to recall and interpret the L2 leading to learning.

Gass (2017) examined the feedback aspect of the IH and argued that it had significant correlations to the input process addressed previously. However, Gass (2017) argued that feedback was received after output was reached during SLA process. There are two types of feedback: explicit and implicit (Al Khateeb, 2015). For the explicit feedback, it referred to the corrections made based on the output of L2 acquisition among the learners for those that made mistakes or can include further explanations of the other language use. On the other hand, implicit feedback comprised of negotiating approaches like confirmation, clarification, and others. Yee, Ning, and Hua (2017) agreed with Gass (2017) that there were also positive and negative feedback approaches that were proposed for the IH. For the positive feedback, the educator concurs with the students' outputs about L2. On the other hand, negative feedback refers to the process of teachers rejecting the student's understanding of L2.

To summarize, the fundamental elements of IH (i.e., input, output, interaction, and feedback) highly relate to the spoken pushed-output instruction in the present thesis (i.e., interactive, communicative, and of spoken modality), as discussed in Section 2.3.1. First, learners receive the input within the activities, and they have to interact with peers in each activity delivering that original input, either via repetition or rephrasing. The use of peers ensures that the students can easily relate with each other and will be more open.

In addition, the peer pressure environment would motivate them towards completing the assigned activities, which would be a high stimulating factor. Also, the communication approach among learners creates a highly stimulating environment that would see them efficiently encode the L2 information. In relation to output, communicating the given input with their peer can be considered one instance of output and not only that but also learners are 'pushed' to make the input they received more comprehensible. Further, spoken pushed-output instruction in the present thesis could serve as generator of better modified input, as listeners' feedback regarding the received input pushes the speaker to modify their production to be more comprehensible to their peers. In other words, learners need to exchange the information they received with their peers and give feedback about their peers production in order to complete the activities. The spoken pushed-output instruction in the present thesis also relates to the feedback aspect as it is through the process of rephrasing or negotiating the meaning that feedback is constructed. Further, learners are more encouraged to use vocabulary orally while interacting with their peers and completing the activities.

2.3.2.3. Transfer appropriate processing theory

Another theory that provides theoretical support for the potential benefits of spoken pushed-output instruction, and output instructions in general, in the language classroom is the Transfer-Appropriate processing theory (TAP). TAP is one of the long-standing theories in cognitive psychology (Morris, Bransford, & Franks, 1977). According to Morris, Bransford, and Franks (1977), TAP postulated that for one to effectively recall a memory, the person had first to achieve efficient encoding of that memory. Morris, Bransford, and Franks (1977) examined tasks in which learners had to encode information with rhyme processing or semantic processing. Later, in the retrieval stage, learners had to complete either a semantic recognition test or a rhyme recognition

test. The results showed that the initial semantic processing group outperformed the rhyme processing in the semantic recognition test, but that initial rhyme processing group was superior to the semantic processing group in the rhyme recognition test. This showed that the resemblance between the initial encoding and the later retrieval influenced performance.

The fundamental principle in TAP is that the ability of memory is also dependent on the process of encoding the information and not just the level of memory processing. In other words, what has been learned is better retrieved if the cognitive processes that are active while learning resemble those that are active when retrieving information (Blaxton, 1989; Morris et al., 1977). This suggests that learning language in communicative contexts might improve the retrieval of it in such contexts (Segalowitz & Lightbown, 1999). Learning is further optimised when the encoding and retrieval processes are similar, due to the fact that, when learning something, memories also record aspects of the setting in which it was learned as well as the cognitive processes involved in the learning (Lee, 1988). TAP emphasises the importance of encoding specificity principle which states that memory is better enhanced when information processed at the encoding stage is also processed the same way at the retrieval stage (Tulving & Thomson, 1973). One good example of TAP effects is Godden and Baddeley's (1975) study. Participants in this study learned lists of words in two settings: on land and under the water. They were later asked to recall the words either in the same setting or the opposite one and they found that participants performed better when they recalled the words in the same setting where they had initially learned those items.

The hypothesis seems to offer a plausible description of a widely accepted phenomenon in SLA: knowledge is easily accessible when both learning activities and testing activities are similar. Mismatch between the learning and actual retrieval contexts of language learning is one of the big criticism of the audiolingual language teaching

method (Lightbown, 2007). Learning vocabulary through drill activities, for example, may be easier to access if the retrieval stage resembles these drill activities, more than in communicative situations. In other words, when language is learned isolated from communicative language, such as in drill activities, it is not readily available to be used in communicative situations. Instruction that integrates language learning within communicative interaction are consistent with TAP as a framework for classroom practice. One such approach was developed by Gatbonton and Segalowitz (1988, 2005). They proposed creating opportunities for extended form–meaning mapping practice. Learners use limited expressions repeatedly to complete the classroom task. Later learners engage in a consolidation activity in which they focus explicitly on these forms and their meanings, with no need to communicate as in the first activity. Finally, learners engage in a more open-ended exchange in which the learned phrases will probably be used. This integrated instructional model can be considered as communicative drills (DeKeyser, 1998) in which learners make repeated use of limited number of expressions.

There have been studies that have provided empirical support for the tenets of TAP for language learning, especially in relation to individual differences such as working memory capacity. These studies reported that there was a link between the stimuli and the memory process, in which higher results were obtained when there is a match between the encoding and testing stages. It is not one of the main goals of the present thesis to examine differences in such factors. However, TAP can be used to explain results of the studies presented in this thesis. In relation to language classroom pedagogy, if the mode of learning is to some extent congruent with the mode of retrieval, language learning could be maximised.

To conclude, if the goal of vocabulary classroom is to improve learners' vocabulary production, the learning stage should involve vocabulary production as well, and one such approach is the spoken pushed-output vocabulary instruction. The studies

in the present thesis involve oral production and later in the testing stage learners will be also completing an oral productive test. So, TAP can be used to explain if the activities resulted in better performance due to this resemblance between the learning and testing phases.

2.3.3. Empirical research on the role of output in language learning

Having reviewed the theoretical frameworks that support the potential benefits of pushed-output instruction, I move on now to review empirical research on the role of output in language learning. Several studies have empirically evaluated the role of output for language learning and the main tenets of the Output Hypothesis. Many of these have focused on the benefits of output with regard to the acquisition of grammatical structures, and a small number have looked at the acquisition of vocabulary. This section is divided into four parts: studies investigating output instruction for grammar acquisition, studies exploring the learning of SWs through spoken output, studies exploring the learning of SWs through written output, and studies exploring output instruction for learning FSs.

2.3.3.1. Research into the acquisition of grammar through output

Many studies have investigated the role of output in the acquisition of grammar, and a selection of these will be reviewed in more detail in this section (e.g., Izumi, 2002; Izumi & Bigelow, 2000; Izumi, Bigelow, Fujiwara, & Fearnow, 1999; Morgan-Short & Bowden, 2006; Nobuyoshi & Ellis, 1993; Shehadeh, 2002). For instance, Nobuyoshi and Ellis (1993) conducted a small-scale study on the role of output in the development of learners' use of the past tense in English. They compared three experimental participants with three control participants. The experimental participants took part in a "focused meaning negotiation" where they received a clarification request every time they made an

error in using the past tense and were encouraged to modify their output. The control participants took part in an "unfocused meaning negotiation" where they received a clarification request only when there was a genuine communication problem. A week later, both groups took part in an unfocused meaning negotiation interaction. The results showed that, of the three learners from the experimental group, two demonstrated increased accuracy in using the past tense, whereas none from the control group improved their accuracy. These results seem to support the Output Hypothesis: the pushed output of the experimental group that focused on one particular linguistic item led to sustained improvements in production. However, the small sample size of the study means that wider conclusions cannot be drawn.

Larger scale studies were carried out by Izumi et al. (1999) and Izumi and Bigelow (2000), and these explored the potential benefits of pushed output in learning counterfactual conditionals (i.e., past hypothetical conditional). Experimental groups in both studies were given texts that included rich examples of the structure and were asked to generate a similar text themselves. Control groups received the same texts but were also asked to complete other tasks, such as answering comprehension questions. All participants completed pre- and post-tests. The results indicated that while the experimental groups showed significant improvement in their use of the structure in writing tasks, the control groups performed just as well as the experimental groups in a post-test. This indicates that a combination of rich input and comprehension-checking activities may be just as effective for learning grammar as activities that ask learners to produce output. Morgan-Short and Bowden (2006) examined the effects of meaningful input- and output-based practices on the acquisition of preverbal direct object pronouns in Spanish. Experimental groups received the same input but differed in their practice activities, which were either input- or output-based. The post-test results indicated that both the input- and output-based practice groups showed significant gains in immediate

and delayed interpretation, as well as production tasks, indicating that both input- and output-based instruction led to linguistic development. The findings of the studies by Izumi et al. (1999), Izumi and Bigelow (2000), and Morgan-Short and Bowden (2006) not only contradict the findings of Nobuyoshi and Ellis (1993) but also reject the contentions of the Output Hypothesis. However, in a later study concerning the acquisition of relativization in English through pushed output, Izumi (2002) reported that learners who were involved in a text re-construction activity outperformed those who were engaged in input comprehension activities, which also confirms the positive effect of pushed output on acquiring some grammatical forms of language.

These mixed results indicate that the benefits of pushed output for grammar development in L2 learning are somewhat elusive and difficult to demonstrate. Thus, Shehadeh (2002) claimed that "there is still a severe lack of data showing that learner output or output modification have any effect on second language learning" (p. 597). This conclusion is debatable in many ways: first, in an effective L2 learning process, comprehensible input plus comprehensible output leads to a successful learning experience in a new language. This means that the two theories complement each other and lead to strong language acquisition. Second, within a classroom context, the positive environment the teacher creates during lessons makes a huge difference in boosting the motivation as well as interest of learners in the process of SLA. Lastly, feedback serves an immense role in shaping learners' experiences. It is relevant to understand that the push theorem does not necessarily mean an extrinsic drive or force in enhancing the language skills of novices, it may also mean the provision of materials, the right motivation, and encouraging using interesting and appealing materials that challenge and motivate learners to open and share their experiences with themselves, their peers, and their instructors.

2.3.3.2. Research into the acquisition of SWs through spoken output

Most studies on the effectiveness of output for vocabulary learning have focused on the examination of written output activities (will be reviewed in the following section). Fewer studies have examined the effectiveness of spoken output for vocabulary learning. Two studies were conducted to explore the effects of non-negotiated input, negotiation without output, and negotiation with output on the acquisition of vocabulary. Ellis and He (1999), for instance, studied the effects of spoken output, input, and modified input on the acquisition of unfamiliar furniture vocabulary (n= 10, overall level of non-recognition was 88%). The subjects were 50 low-level English learners. All groups carried out a similar activity: they were asked to place pictures of furniture on a plan of an apartment. The input group (n=18) received pre-prepared instructions that could not be modified, the modified input group (n=16) received the same instructions but were able to negotiate and ask for clarification, and the output group (n=16) was required to give instructions to an interlocutor. The study included a form-recognition pre-test given one week before the treatment; a picture-matching meaning-recognition post-test given in one-week, three-week, and four-week intervals of time; and an oral picture-labelling post-test in which learners worked in pairs and produced instructions on where to place a given picture of the target item (piece of furniture) on a map. This post-test was given at different time intervals, i.e., two-week and four-week delayed post-tests. The results showed that the output group outperformed the other two groups in their ability to recognise meaning in the picture-matching tests and to produce lexical items in the picture-labelling post-tests at all time intervals up to four weeks later. They explained that the advantage that the output group had over the other groups was related to the fact that the learners, when producing new words, were engaged in a deeper level of processing than only hearing them, which resulted in higher gains in both comprehension and production.

In a similar study, De La Fuente (2002) investigated the effects of non-negotiated input, negotiation without output, and negotiation with spoken output on the acquisition of Spanish vocabulary. The subjects were 32 English speakers, intermediate learners of Spanish in a basic language programme at Georgetown University, receiving 90 hours of formal exposure to the L2, and they were randomly assigned to one of the treatment groups. The target words in this study were nouns that refer to people or everyday objects (those whose referents evoke images). The design of the study included receptive and productive vocabulary knowledge scale post-tests, given at two intervals of time: immediately after the treatment and three weeks later (based on Wesche & Paribakht, 1996). Unlike Ellis and He (1999), the results showed that in terms of receptive acquisition of L2 words, the negotiated groups, both with and without output, outperformed learners in the non-negotiated groups. However, in terms of the productive acquisition of L2 words, the negotiated group that incorporated output was the only one that promoted productive acquisition of the target words. The findings of this study provide evidence for the importance of negotiation in facilitating recognition of L2 vocabulary and the importance of the role of output in the acquisition of productive knowledge of lexical items. The findings of studies Ellis and He (1999) and De La Fuente (2002) suggest that the negotiation of specific lexical items may aid in vocabulary acquisition, "provided that the students have the opportunity to use the items they have begun to acquire and to receive feedback from other speakers" (Ellis, Tanaka, & Yamazaki, 1994, p. 483).

To sum up, both Ellis and He' (1999) and De La Fuente' (2002) provided evidence for the benefits of the spoken pushed-output instruction for SWs (mainly nouns); however, more research is needed to examine different activities of spoken pushed-output activities for learning other parts of speech (e.g., verbs) or FSs.

2.3.3.3. Research into the acquisition of SWs through written output

Exploring the effects of written output-based activities on vocabulary learning has received a lot of attention in research (e.g., Folse, 2006b; Holster & DeLint, 2012; Hulstijn & Laufer, 2001; Keating, 2008; Nassaji & Tian, 2010; Rassaei, 2017; Tahmasbi & Farvardin, 2017; Webb, 2005; Zou, 2017). Webb (2005), for instance, examined vocabulary learning in receptive (reading) and productive (writing) conditions and demonstrated that both reading and writing tasks had a positive effect on five dimensions of vocabulary knowledge: orthography, association, syntax, grammatical functions, and form-meaning. Learners were assessed on these five components by means of five receptive tests and five productive tests. The results also showed that when learners were involved in writing novel sentences using the target items, the retention of vocabulary was better than when they were simply asked to read sentences with glossed target items. It is important to mention that, to the best of my knowledge, Webb's (2005) study is the only one to have examined the acquisition of multiple components of vocabulary knowledge through output-based tasks. Further, Webb's (2005) study can be considered as the first attempt to examine learning FSs through output-based tasks, as he tested the knowledge of syntagmatic associates (e.g., collocations) in the syntax component, and knowledge of paradigmatic associates (e.g., coordinates, superordinates, subordinates, antonyms, and synonyms) in the association component, both receptively and productively.

Several studies have used the Involvement Load Hypothesis (ILH) framework to examine the impact of written output activities on L2 vocabulary learning (see Section 2.4.2 for more details on the ILH). Hulstijn and Laufer (2001) explored vocabulary learning in three conditions that differed in the involvement load indices they induced: reading with marginal glosses (no involvement), reading plus a gap-filling activity using target items (moderate involvement), and composition-writing using target items (strong

involvement). The rate of vocabulary learning was the highest in the composition-writing condition, followed by reading and gap-filling, and then the reading with glosses conditions, which confirms the contention of the ILH; that is, the stronger the involvement load induced by an activity, the more beneficial it is for vocabulary learning. Keating (2008) obtained similar results in an investigation of three conditions: reading only, reading plus gap-filling, and reading plus sentence-writing. The reading plus sentence-writing, the highest in terms of involvement load, resulted in greater vocabulary learning than the other two conditions.

Rassaei (2017) and Sun (2017) both investigated different post-reading output conditions of L2 vocabulary knowledge. Rassaei (2017) compared the effectiveness of three post-reading activities: summarising, asking and answering questions about the text, and a prediction task. The subjects were 88 intermediate level EFL learners enrolled on an EFL programme. A list of 14 words were identified as the target items, based on the learners' initial recognition scores prior to treatment. All participants read two texts twice and completed one of the three written output activities (except for the control group, who only read the texts). The results showed that the output activities had a marked effect on word learning, as reflected in the recognition and recall of the form of target items, with the prediction activity being the most effective, then questions and answers, finally summarising. This superiority of the prediction activity was related by the author to the higher involvement load induced, as learners were required to use the target items in a novel context, while the other two tasks involved more familiar contexts in production.

Sun (2017) examined three picture-book reading conditions: reading only, reading with vocabulary instruction by the teacher, reading with a collaborative written output activity. The knowledge of 20 unfamiliar target items (as shown by the learners' scores in a pilot study) was measured using a Vocabulary Knowledge Scale (VKS) in immediate and one-month delayed post-tests, combining both receptive and productive knowledge.

The results showed that in the immediate post-test, the condition of reading with vocabulary instruction resulted in the best results, followed by the reading with output and then the reading only conditions. The reading with output condition, however, was the most effective one in the delayed post-test, while only a slight difference was found between the other two conditions. Similar to how Rassaei (2017) explained the results, Sun explained that the output condition was considered to be a higher involvement load condition, involving 'search', 'strong need', and 'strong evaluation'. However, one of the limitations of Sun's (2017) study is that the task-test-task-test effect was not controlled for in the study. Participants received three different treatment sessions on different weeks, and the testing instruments were administered separately at different times. It is possible that learner could have anticipate that a vocabulary test would come at the end of each lesson, and directed their attention to vocabulary, which could have affected the test scores for the last two sessions.

While all the aforementioned studies examined post-reading output activities, Zou (2017) investigated the learning of ten target items through three output activities: a cloze exercise, sentence-writing, and short essay-writing. Receptive and productive knowledge of the target items were measured through a modified version of VKS (Paribakht & Wesche, 1997). This study was unique in that it is the only one that included analysis of think-aloud and retrospective interviews about how they were handling words when completing the assigned tasks. Composition-writing and sentence-writing, both with an identical involvement load (+need, –search, ++evaluation), resulted in better performance than the cloze exercise. Zou suggested introducing an additional level of evaluation (++ strong), driven by the data from think-aloud protocols and interviews. In the composition task, learners were using both chunking and the hierarchical structure of their information, while such actions were not noted in either in the sentence-writing or cloze exercises. Chunking refers to processing information in meaningful grouped units rather than

individual units. The use of chunking helps in memorising words and improves the ability to successfully retrieve them (Zou, 2017). Hierarchical structure refers to how to make coherent connections between several chunks of information to build up a coherent piece of writing or speaking (Zou, 2017). However, one of limitations of Zou's (2017) study is that time on tasks was not properly controlled (30 minutes in the cloze-exercises compared to 35 minutes in the two writing tasks, and that could have affected the results.

It is important to note that the type of output involved also has an effect on the acquisition of L2 vocabulary. VanPatten (2003) distinguished between output with or without access to meaning. "Output with access" refers to "activating the lexical items and grammatical forms necessary to express particular meanings" (p. 63), whereas "output without access" refers to the production of language that does not require this type of activation, such as the repetition of a lexical item without the intention to convey any meaning. Current SLA processing theories emphasise the greater cognitive demands of developing productive knowledge of a L2 as opposed to receptive knowledge, as discussed in Section 2.1.1, describing the benefits that depth of processing has on language acquisition (Skehan, 1998). Depth of processing refers to the degree of analysis and manipulation carried out in the target language. Processing that requires greater depth of processing (e.g., paraphrasing rather than repeating) results in longer-term memory traces (Swain, 2005). Therefore, output that does not require access or that involves very little depth of processing may not aid in vocabulary acquisition. In fact, Barcroft (2006) found that writing target vocabulary words as a form of output had a negative effect on acquisition. For Spanish ESL learners, he compared the effects of copying target vocabulary with no required output. Post-tests revealed that productive learning was greater when learners did not write down the words, which suggests that forced output without access to meaning may detract from learning, as it exhausts the processing resources that are needed for encoding novel lexical forms. Similarly, De la Fuente (2006)

found that when learners were asked to perform shallow output tasks with vocabulary, such as repeating words in delayed tests, their retrieval of lexical items decreased compared to learners who were asked to perform deeper tasks, such as online retrieval in information-gap activities. However, she also noted that the increased frequency of retrieval in the latter tasks may have contributed to the increase in learning. Therefore, output of lexical items that promotes deeper processing may lead to more elaborate processing of form, and therefore a more durable memory trace (Izumi, 2002).

Another study that compared two modes of output: mechanical and creative output for learning vocabulary, was conducted by Holster and de Lint (2012). Mechanical output is based on bilingual example sentences in which learners compare a gapped L2 example sentence with a complete L1 translated sentence and choose a target word to complete the gap, whereas creative output requires the creation of authentic meanings, simply writing original sentences. The research sample was entirely composed of first-year students with TOEIC Bridge Test scores that were below 100. The curriculum included two mandatory English lessons weekly; using a MC completion test, mechanical output students and creative output students were provided with identical assignments that included diary-writing, homework, and vocabulary assignments. The common expectation was that the mechanical output students would record significant vocabulary gains, the contention being whether they were significant enough to justify the amount of classroom time spent on mechanical tasks (Holster & de Lint, 2012). The pre- and post-test outcomes of both groups showed significant improvements in vocabulary knowledge. This meant that an individual who had a 50% likelihood of getting an item correct on the pre-test would have a 59% and 62% likelihood of getting similar items correct on the post-test. Mechanical output students had no greater apparent vocabulary gains compared to creative output students. Therefore, the study did not support the notion that mechanical output tasks offered significant long-term vocabulary gains compared to creative output (Holster & de

Lint, 2012). In fact, students who did mechanical output tasks showed lower gains than creative output students, which may indicate there was no significant difference, either in substance or statistics, between the two groups. The unexciting nature of the mechanical output tasks also led to a greater attrition rate of over 60% compared to the attrition rate of 31% for students undertaking creative output tasks (Holster & de Lint, 2012). Thus, it was postulated that creative output tasks resulted in positive objective appraisals, which were seen as facilitating learning depth and sustainability, whereas students perceived mechanical output tasks as intensive and dull, resulting in higher levels of demotivation and attrition.

To conclude, the findings of the different studies investigating written pushed-output instruction supports such instruction for vocabulary learning, mainly in relation to the ILH or the amount of negotiation involved in the written pushed-output activities.

2.3.3.4. Research into the acquisition of FSs through written output

Of particular relevance to the present study, a few studies have examined output instruction for the acquisition of FSs. Peters and Pauwels (2015), for instance, explored the effect of vocabulary-focused instruction on learners' recognition, cued output, and spontaneous use of academic FSs. Participants were 29 Dutch-speaking EFL learners, and they were engaged in vocabulary-focused activities on FSs. Three instructional conditions for learning academic FSs were examined: recognition activities, written cued output activities, and a combination of the two. The recognition activities focused on raising awareness of target FSs in academic writing and three types of activities were included: underlining FSs relevant to academic writing in sentences, recognising the more academic of two sentences, and indicating which section in their paper would use given sentences, e.g., Literature, Aim, and Method. Written cued output exercises included fill-in-the-blanks activities, rephrasing activities with or without a clue, and use-in-sentences

activities. Three tests at different levels of sensitivity were administered: a recognition test, a written cued output test, and a writing test. The recognition test required learners to underline academic FSs in several excerpts from scholarly articles. The written cued output test asked learners to supply the FSs in given sentences, with L1 translations provided in brackets. The writing test involved reading an interview (spoken register) in which researchers explain a study they have recently conducted (rationale, research questions, method, results), and then writing a 250-350-word summary of that interview. The findings suggested the vocabulary-focused approach to academic FSs did indeed result in significant learning gains on the three tests: recognition, cued output, and use. The findings regarding the type of instruction were not conclusive, but it seems that activities involving cued output exercises might be more beneficial not only at the level of production but also at the level of recognition.

Nassaji and Tian (2010) examined output for the acquisition of phrasal verbs in two conditions: collaborative and individual conditions. The activities they used were: a reconstruction cloze and a reconstruction editing activity, and both were operationalised in both conditions. The target items of this study were 16 English phrasal verbs. Participants were chosen from an adult ESL programme conducted in Canada. They were divided into two groups, with 12 members in one group and 14 in the other. The study comprised a pre-test, treatment, and a post-test after four days of treatment. Four activities were prepared: two cloze activities and two editing activities. In both treatments, one activity was operationalised individually and the other collaboratively. Each task targeted four of the 16 target words. The collaborative sessions were recorded, transcribed, and analysed to understand the thought processes involved in collaborative learning. The results showed that, although the rate of task completion was better in a collaborative setting, there was no significant difference in learning between the two activities completed either collaboratively or individually. The editing activity invoked a

significantly higher gain of knowledge compared to the cloze activity. The success of editing activities over cloze activities could be attributed to the fact that editing activities are easier and provide cues to learners for picking up target words (Nassaji & Tian, 2010). Hence, all in all, a collaborative reconstruction editing activity was found to be effective for learning phrasal verbs. It is possible that the collaborative activities did not show an improved level of learning over individual activities due to the absence of a training session and familiarity with how to collaborate and learn effectively in pairs. The backgrounds of the participants and their collaborative effectiveness should also be considered an important factor in these kinds of studies — post-treatment activities could be conducted to enable learners to assess their individual improvements from the pre-treatment stage.

Overall, the studies on vocabulary learning by means of pushed-output instruction, written or spoken, seem to suggest that vocabulary can be learned better using such instruction than other instructions, such as input-only or modified/negotiated input, etc. However, the consequence of research having consistently focused on examining written pushed-output instruction points to the need to explore the effectiveness of spoken pushed-output instruction for vocabulary learning. Further, no clear conclusions about the acquisition of FSs by means of spoken pushed-output instruction can be drawn. Most of the studies reviewed so far only looked at the acquisition of SWs and overlooked the fact that a high percentage of language is formulaic and hence examining learning FSs is needed as well.

2.4. Designing pushed-output activities

An important question in research on pushed-output instruction is how to design effective pushed-output activities. Different frameworks have been used in the literature

to design and categorise pushed-output activities. One such framework was used by Nation and Newton (2008) in their conceptualisation of pushed output; that is, using a different range of topics, a different range of text types, and various performance conditions (planning, time pressure, amount/ type of feedback, information distribution). Further, as discussed in the previous section, several studies in the pushed-output literature used the ILH in either the design or discussion of their activities. While not particularly designed for the discussion of pushed-output instruction, frameworks such as Levels of Processing theory (Craik & Lockhart, 1972), the ILH (Hulstijn & Laufer, 2001), and Technique Feature Analysis (TFA) (Nation & Webb, 2011), are relevant for the design and discussion of pushed-output activities. In this section, these three main frameworks are reviewed in turn.

2.4.1. Levels of Processing Theory

The L1 acquisition process entails a gradual mastery of linguistic elements already present in child cognitive knowledge to express ideas (Lightbown & Spada, 2013). More precisely, learning a word in a L1 involves the gradual acquisition of the properties of a lexical item through exposure to those lexical items in various authentic contexts (Groot, 2000). In an instructed L2 environment, creating such a setting could be considered problematic, not only because there is not enough time to guarantee enough exposure to new words with the same intensity as in L1 acquisition, but also because the exposure provided is commonly considered superficial (i.e., it does not entail presenting or learning different components of word knowledge beyond the form-meaning link). Such superficial exposure leads to shallow processing of L2 words, which might fail to establish "enough associations and links with other words for solid storage and efficient retrieval" (Groot, 2000, p. 61). Further, such exposure cannot be contextualised as it is in L1 (Jiang, 2004) and might not incorporate the different components of vocabulary

knowledge (Schmidt, 2001). Hulstijn (2001) suggested that the retention of new words is further facilitated through the way they are processed, which suggests that a deeper, more elaborate level of processing is needed for better word retention.

The Levels of Processing theory was proposed by Craik and Lockhart (1972), arguing that there are different levels of depth in which an item can be processed in the brain. Craik and Tulving (1975) proposed that for better retention of the meanings of lexical items, a rich encoding of materials was needed. Craik and Lockhart (1972) suggested three levels of processing: 'shallow' involving how the word look like, 'phonetic' involving how the word sound, and 'deep' related to what does the word mean (Craik & Lockhart, 1972). Craik and Lockhart (1972) suggested that the new word that is stored in long-term memory is not the outcome of long-time storage in short-term memory, but rather the depth to which that piece of information was initially processed. Several studies reported that semantic processing (or deep processing) led to better recall of vocabulary than the other two levels of processing (e.g., Craik & Lockhart, 1972; Craik & Tulving, 1975; Lockhart, 2002). Baddeley (1978), Eysenck (1978), and Nelson (1977) criticised the depth of processing theory on the ground that the conceptualisation of levels of processing was vague; therefore, it was difficult to differentiate between deeper and shallower levels of processing. However, it has been confirmed that when learners process information in more depth, their retrieval of that information was found to be enhanced (e.g., Craik & Lockhart, 1972; Craik, 2002; Loaiza, McCabe, Youngblood, Rose, & Myerson, 2011; Rose & Craik, 2012). The Levels of Processing theory provided a platform for research examining vocabulary acquisition and retention and the development of both the ILH as well as the TFA framework, which are examined in turn in the following sections.

2.4.2. Involvement Load Hypothesis (ILH)

The ILH was first introduced by Hulstijn and Laufer (2001) in an attempt to operationalise the general cognitive notions of the Depth of Processing theory (Craik & Lockhart, 1972) and the richness of encoding and elaboration (Craik & Tulving, 1975) into L2 vocabulary teaching. ILH is considered a motivational-cognitive construct of involvement, which consists of three main constructs: need, search, and evaluation. The *need* construct is a motivational component reflected in the extent of need induced by the task of processing a word. A task can induce a 'moderate' need if it is imposed by an external agent such as a teacher, or a 'strong' need if it is intrinsically motivated or self-imposed by learners. For example, a task that asks learners to use a word in a sentence induces a moderate need as the teacher is the one requesting it, whereas deciding to look up a word in a dictionary when writing induces a strong need as it is imposed by learners themselves.

The *search* and *evaluation* components are cognitive elements of ILH hinging on drawing attention to form-meaning relationships (Schmidt, 2001). Search refers to trying to ascertain the meaning of unknown lexical items. It can be either present if the task requires the learners to seek the meanings of unknown words in order to complete the task (e.g., look-up in a dictionary tasks), or absent when such an effort is not required (reading comprehension with marginal glosses tasks). Evaluation involves comparisons and making decisions about a word's suitability in a given context; for instance, comparing a word with other words, comparing the meaning of a word with its other meanings to decide which better fits the context. Evaluation can be moderate if it requires decisions such as choosing the best meaning to fit a given context, or strong when the task involves deciding how to combine a word with additional words in an original sentence or text. Laufer and Hulstijn (2001) suggested that tasks with a higher involvement load (IL) should be considered more effective for better vocabulary learning

and retention than those with lower IL. They also suggested that tasks that involve drawing more attention to components of vocabulary knowledge (e.g., pronunciation, orthography, meaning, collocational knowledge) would support greater retention than those tasks drawing attention to only one or two of those components.

Tasks are assigned an IL index on the basis of the presence or the absence of the three components, as well as the strength of their presence (indices of 0, 1, or 2 to reflect the absence, moderate presence, or strong presence of a factor). Hulstijn and Laufer (2001) compared two tasks with different ILs: task 1 is sentence-writing with target words with glossed translations (and explained by the teacher), and task 2 is reading a text with glossed words with meanings and answering comprehension questions that require knowledge of the meaning of some unknown vocabulary in the text. Task 1, as they illustrated, induced an index of 3, a moderate need, no search and strong evaluation (i.e., '1' need, '0' search, and '2' evaluation), whereas task 2 induced an index of 1, a moderate need, no search or evaluation (i.e., '1' need, '0' search, and '0' evaluation). The results supported the contentions of the ILH in which the first task, a higher IL task, was more effective for vocabulary learning than the second task inducing lower IL (p. 546).

Fan and Nyikos (2007) further hypothesised that the essence of teaching materials is what leads to more effective learning than the strategic approach of the learner. Some empirical studies have adopted the ILH in an attempt to examine the effectiveness of tasks (or activities) for better learning and retention of vocabulary either in relation to individual words (e.g., Eckerth & Tavakoli, 2012; Hulstijn & Laufer, 2001; Keating, 2008; Kim, 2008; Laufer & Girsai, 2008; Rassaei, 2017; C.-H. Sun, 2017; Zou, 2017) or collocations (e.g., Fan, 2009; Laufer & Girsai, 2008). Several studies, such as Beal (2007), Eckerth and Tavakoli (2012), Huang, Willson, and Eslami (2012), Hulstijn and Laufer (2001), Keating (2008), and Kim (2008), have all supported the ILH as they found writing tasks (IL=3: need =1, search=0, evaluation=2) to be more effective than cloze

exercises (IL=2: need =1, search=0, evaluation=1) and reading comprehension tasks (IL=1: need =1, search=0, evaluation=0). Pichette, De Serres, and Lafontaine (2011) further reported that reading tasks with no or moderate evaluation were less effective than writing tasks with strong evaluation. The value of writing tasks with higher ILs was also noted in Laufer and Rozovski-Roitblat's (2011) as well as Moonen, De Graaff, Westhoff, & Brekelmans's (2014) studies. Niu and Helms-Park (2014) highlighted that both oral and written output inducing high ILs worked better for vocabulary learning than reading comprehension tasks, which induce lower ILs. Additionally, Nassaji and Hu (2012) observed three types of meaning inferential strategies: regular inferring of meaning with no options (IL=3: need=1, search=1, evaluation=1); inferring meaning from the options provided (IL=2: need=1, search=0, evaluation=1), inferring meaning through a derivational transformation exercise (IL=5: need=1, search=2, evaluation=2). They concluded that inferring meaning through a derivational transformation exercise was the most effective strategy, compared to inferring meaning without options and inferring meaning with options, respectively. Kim (2011) examined two tasks that were assumed to represent the same IL: composition- and sentence-writing and reported that the two tasks resulted in equal learning and retention of vocabulary. These studies were all in relation to the learning and retention of SWs; with regard to FSs, Laufer and Girsai (2008) investigated the effectiveness of two tasks for learning collocations: a reading comprehension task and a reading with translation task, and they reported a similar pattern: the higher the IL task (i.e., reading with translation), the better it was for the learning and retention of vocabulary.

However, other studies have provided evidence contradicting the ILH’s predictions. Laufer (2003), for instance, found that sentence-writing (IL=3: need=1, search=0, evaluation=2) was less effective than sentence-completing with dictionary consultation (IL=3: need=1, search=1, evaluation=1), even though the two tasks were assumed to

induce the same ILs with different degrees of prominence of the components. Folse (2006b) and Liu (2013) reported that some cloze exercises (IL=2: need=1, search=0, evaluation=1) were better than a sentence-writing activity (IL=3: need=1, search=0, evaluation=2), which contradicts the ILH. Zou (2017) suggested that the differences between composition-writing and sentence-writing should be clearly discussed, or perhaps another level of degree of prominence should be incorporated, as even though the two activities induced the same amount of IL, the specific functions of the two activities were different. Previous researchers, such as Folse (2006b), Keating (2008), and Webb (2005), questioned the contentions of the ILH as the effect of tasks with higher ILs, i.e., leading to better retention, disappeared when the time spent on a task was controlled. Nation and Webb (2011) stated that the quality of any analytic measure hinged on the reliability of the analysis and, since the ILH consists of only three components, it might add more ambiguity and difficulty to being consistent in the analysis. They further argued that the time-on-task and the contributory effect of each of the components of need, search, and evaluation (Kim, 2008) put this hypothesis on shaky ground. Nation and Webb (2011) made three suggestions to overcome these issues: "increasing the number of components, increasing the number of response categories for each component, and providing explicit anchoring labels for each category" (p. 5). Thus, they introduced TFA, reviewed in the next section, as a more elaborate framework for analysing vocabulary activities.

2.4.3. Technique Feature Analysis framework (TFA)

TFA was first introduced by Nation and Webb (2011) as a more elaborate and extended version of ILH. It combined the three components of vocabulary knowledge proposed by Nation (2000): noticing, retrieval, and generation, along with two new components: motivation and retention, to the components of the ILH as a way to

compensate for the inadequacies of the ILH and make the framework more quantitative and comprehensive for both researchers and teachers. Table 2.4 presents the five components of TFA and specific questions for each component, which implies a maximum score of 18 for vocabulary teaching activities.

Motivation relates to the purpose of the task; noticing deals with the attention to target words in the task; retrieval deals with whether tasks involve receptive or productive retrieval; recognition vs. recall retrieval; single vs. multiple retrieval and spaced vs. massed retrieval; generation can be receptive when encountering new words in a new context or productive when using words in an original context; and retention deals with ensuring a successful form-meaning link and whether forming this link requires instantiation or imaging with no interference. Webb and Nation (2017) used another term to describe the generation aspect of the same framework (varied encounters and varied use). Nakata and Webb (2016) analysed three vocabulary activities according to TFA: flashcards, crossword puzzles, and cloze exercises, and found that though flashcards induced higher TFA scores, though they have their own weakness with regard to the ability to promote generation when compared to crossword puzzles and cloze exercises. They further recommended that teachers and materials developers combine several activities in order to complement each other's weaknesses.

Table 2.4 *A checklist for TFA (Nation & Webb, 2011, p. 7; Webb & Nation, 2017, p. 236)*

Criteria	Score	
Motivation		
Is there a clear vocabulary learning goal?	0	1
Does the activity motivate learning?	0	1
Do the learners select the words?	0	1
Noticing		
Does the activity focus attention on the target words?	0	1
Does the activity raise awareness of new vocabulary learning?	0	1
Does the activity involve negotiation?	0	1
Retrieval		
Does the activity involve retrieval of the word?	0	1
Is it productive retrieval?	0	1
Is it recall?	0	1
Are there multiple retrievals of each word?	0	1
Is there spacing between retrievals?	0	1
Generation (varied encounters and varied use)		
Does the activity involve generative use (varied encounters and use)?	0	1
Is it productive?	0	1
Is there a marked change that involves the use of other words?	0	1
Retention		
Does the activity ensure successful linking of form and meaning?	0	1
Does the activity involve instantiation?	0	1
Does the activity involve imaging?	0	1
Does the activity avoid interference?	0	1

Only two empirical studies have compared the efficacy of the scoring of both ILH and TFA. The first was an empirical study conducted by Nation and Webb (2011), in which they compared the scoring of ILH and TFA on several vocabulary activities. They reported some dissimilarities in scores between the two frameworks. For instance, some

activities were considered high IL activities whereas the same activities were given a lower index according to TFA, such as word cards and fill-in-the-blanks. Word cards involved splitting information about a word on the two sides of a card and learners reviewed information on one side and tried to retrieve the information on the other side. This activity was given an index of three out of six by Nation and Webb (2011), according to the guidelines of the ILH, whereas according to the TFA framework, this activity would induce a higher index of 11 out of 18. The fill-in-the-blanks activity, on the other hand, was assumed to induce an index of 4 out of 6 according to the ILH, but a score of 8 out of 18 for the TFA (Nation & Webb, 2011). The differences between the two frameworks can be linked to the way in which the two operationalise depth of processing and the varying degrees of each component. Nation and Webb (2010) encouraged conducting empirical studies to further examine the two frameworks, stating that "there is certainly scope for experimentally comparing ILH and TFA" (p. 26).

Second, Hu and Nassaji (2016) compared four tasks that differed in their scores according to the ILH and TFA: reading with MC items, reading with choosing definitions, reading with fill-in-the-blanks, and reading with rewording sentences. Ninety-six EFL learners were assigned to four groups completing one of these four activities and required to learn the meanings of 14 words. Table 2.5 presents an analysis of the four activities according to both TFA and the ILH (Hu & Nassaji, 2016, p. 33). Concerning motivation, as the points suggested in the frameworks, all the activities motivated the learning of the target words to some level, even the MC activity as it was only concerned with sections in which the target items occurred and it was necessary for learners to comprehend the meaning to complete the activity. With regards to the generation aspect, all the activities involved reading first, which offered the learners the opportunity to meet the L2 words in a new context later in the activities. According to Nation and Webb (2011), generative use can be receptive when encountering the words in new context, or productive, when

using the words in novel contexts. However, I would say that even though the TFA was introduced as a way to compensate for the inadequacies of the ILH and make the framework more quantitative and comprehensive for both researchers and teachers, it is still unclear for some components such as motivation and generation, to be defined properly within the different activities. Hu and Nassaji (2016) reported that TFA had better explanatory power than the ILH in terms of the prediction of vocabulary gains of the activities.

Table 2.5 *Four tasks analysed using TFA and the ILH (Hu & Nassaji, 2016, p. 33)*

	Reading + MC activity	Reading + choosing definitions	Reading+ fill-in-the-blanks	Reading + rewording sentences
Motivation				
Is there a clear vocabulary learning goal?	0	1	1	1
Does the activity motivate learning?	1	1	1	0
Do the learners select the words?	0	0	0	0
Noticing				
Does the activity focus attention on the target words?	1	1	1	1
Does the activity raise awareness of new vocabulary learning?	0	1	1	1
Does the activity involve negotiation?	0	0	0	0
Retrieval				
Does the activity involve retrieval of the word?	1	1	0	0
Is it productive retrieval?	0	0	0	0
Is it recall?	1	0	0	0
Are there multiple retrievals of each word?	0	0	0	0
Is there spacing between retrievals?	0	0	0	0
Generation				
Does the activity involve generative use?	1	0	1	1
Is it productive?	0	0	0	1
Is there a marked change that involves the use of other words?	0	0	0	0
Retention				
Does the activity ensure successful linking of form and meaning?	0	0	1	1
Does the activity involve instantiation?	0	0	0	0
Does the activity involve imaging?	0	0	0	0
Does the activity avoid interference?	1	1	1	0
Total score	6	6	7	6
IL index (need, search, evaluation)	1+1+1=3	1+1+1=3	1+0+1=2	1+0+2=3

The two frameworks, ILH and TFA, can be beneficial when designing any vocabulary pedagogical activities in general, and pushed-output activities in particular. Both frameworks help to identify the level of noticing in designed activities, which is considered one of the functions that can be promoted by pushed-output activities, as discussed in Section 2.3.2.1 (Swain, 1995). In terms of the ILH framework, evaluation can be promoted by the fact that pushed-output activities are better performed in collaborative work, as suggested by Swain (2005). Learners will communicate the input they received with their peers in order to complete the activities; hence, more comparisons would occur between the learners' output in each pair or any given input. With regard to the TFA framework, generation and retrieval are particularly important when designing pushed-output activities, since production itself incorporates some degree of productive retrieval, whether form or meaning retrieval, as well as some generation, as the target items will be used in a new context. Overall, instructors could benefit from the use of either the ILH or TFA. However, it seems that TFA presents more sensitive and quantifiable factors that can serve as a checklist when designing vocabulary activities. The two frameworks, ILH and TFA, will be used for the design of the spoken pushed-output activities in Study 2 in the present thesis.

2.5. Summary and conclusions

Overall, there seems to be a consensus about the need to improve not only the size of vocabulary knowledge but also other components that reflect depth of vocabulary knowledge, such as polysemy. Research evidence has shown that ascertaining the depth of learners' vocabulary knowledge is relevant to pedagogic practices (Read, 2000). This applies to both SWs and FSs. That is because there seems to be an agreement about the importance of FSs in order to become a fluent user of an L2 (Crossley, Salsbury &

Mcnamara, 2014; Durrant & Schmitt, 2009; Kremmel, Brunfaut & Alderson, 2015; Stengers, Boers, Housen & Eyckmans, 2011). Previous research has shown that FSs can be learned either through incidental approaches (e.g., Laufer & Girsai, 2008; Pellicer-Sánchez, 2015; Sonbul & Schmitt, 2013; Szudarski, 2012; Szudarski & Carter, 2016; Webb, Newton & Chang, 2013) or deliberate approaches (e.g., Boers, Dang & Strong, 2017; Boers, Demecheleer, Coxhead & Webb, 2014; Chan & Liou, 2005; Jones & Haywood, 2004; Le-Thi, Rodgers & Pellicer-Sánchez, 2018; Nassaji & Tian, 2010; Peters & Pauwels, 2015; Sun & Wang, 2003). However, pushed-output instruction has received less attention than other deliberate learning approaches in the literature on the pedagogy of FSs. Nassaji and Tian's (2010) study is the only one to have examined the acquisition of one type of FSs, phrasal verbs, through written pushed-output instruction. As for polysemy, unfortunately, there has been a limited amount of pedagogical advice on the teaching of polysemous words and more research is needed to examine the effectiveness of different, incidental or deliberate, approaches for teaching polysemous words. No previous study has investigated the acquisition of polysemous words through spoken pushed-output instruction.

Research evidence has shown that vocabulary can be learned through written pushed-output activities (e.g., Folse, 2006b; Holster & DeLint, 2012; Hulstijn & Laufer, 2001; Keating, 2008; Nassaji & Tian, 2010; Rassaei, 2017; Tahmasbi & Farvardin, 2017; Webb, 2005; Zou, 2017). However, to the best of my knowledge, with the exception of these two studies, Ellis and He (1999) and De La Fuente (2002), there is little empirical evidence to demonstrate that spoken output instruction promotes vocabulary learning. These few studies have mainly focused on the acquisition of nouns. Little is known about the acquisition of other parts of speech, FSs, and other components of lexical mastery, such as polysemy, through spoken pushed-output activities. The current study seeks to

investigate the role of spoken pushed output in the acquisition of polysemous SWVs and PVs.

Chapter 3

Study 1.1: Pushed-output instruction in comparison to traditional instruction: differences in learning gains

As discussed in the previous chapter, the overreaching goal of the Output Hypothesis is to emphasise the role of production in EFL classrooms; that is, in order to learn a language, you need to practise producing that language in the classroom. To date, few published studies have examined pushed-output instruction in vocabulary teaching, and the findings from those few studies have confirmed that vocabulary can be learned through pushed output; however, there has been a distinct lack of research on teaching polysemous words as well as teaching FSs through pushed-output instruction. The present chapter reports the results of a classroom intervention study exploring the effects of spoken pushed-output activities on vocabulary knowledge of polysemous single-word units as well as multiword units. A between-subjects design was used including three conditions: no treatment, traditional[1] treatment, and spoken pushed-output treatment. Both receptive and productive knowledge were examined. The data obtained were analysed using two approaches: first, examining receptive and productive vocabulary gains after instruction, which is the focus of the present chapter (Study 1.1); second, looking beyond vocabulary gains and examining the lexical profile of spoken production after instruction, which will be presented in Chapter 4 (Study 1.2).

[1] The label 'traditional' is used to refer to vocabulary-focused activities typically used in the classroom, such as matching and cloze exercises with keywords.

3.1. Background of the study

As reviewed in Chapter 2 (Section 2.3.3), studies on vocabulary learning through pushed output have investigated output-based approaches from different perspectives. De La Fuente (2002) and Ellis and He (1999) compared spoken output-based approaches to input-based approaches; Tahmasbi and Farvardin (2017) conceptualised written output activities within the Involvement Load Hypothesis (ILH); and Ellis, Tanaka, and Yamazaki (1994), Gass (1988), Gass and Varonis (1994), and Loschky (1994) related the effect of written output activities to the level of negotiation involved in the activities. Other studies have examined the effectiveness of several types of written output activities (e.g., Barcroft, 2006; Holster & DeLint, 2012; VanPatten, 2003). Nassaji and Tian (2010), for example, examined differences in written output activities between individual and collaborative work. Rassaei (2017) and Sun (2017) examined written output activities at the text level, that is, post-reading vocabulary-based output activities. The studies comparing the effect of pushed-output instruction with other types of instruction seem to point in the same direction: learning vocabulary is more successful when using means of activities that involve spoken or written vocabulary production (De la Fuente, 2006; De La Fuente, 2002; R. Ellis & He, 1999; R. Ellis et al., 1994; Nassaji & Tian, 2010; Rassaei, 2017; Sun, 2017; Tahmasbi & Farvardin, 2017). However, the focus has been on examining vocabulary knowledge as a unidimensional construct, rather than exploring the different aspects of knowing a word (Nation, 2001), and highlighting which aspects are indeed emphasised.

As discussed in Section 2.3.3.2, few studies have examined the effect of spoken pushed output for vocabulary learning. For instance, Ellis and He (1999) compared a spoken output condition with an input-only condition and a modified-input condition, and the findings revealed significantly larger learning gains for English words, receptively

and productively, in the spoken output condition compared to the other treatments. De La Fuente (2002) examined differences between a non-negotiated input condition, negotiation without output, and negotiation with spoken pushed-output treatments and suggested that when negotiation incorporated spoken pushed output, it promoted both receptive and productive acquisition of target Spanish nouns. Apparently, the variations in treatments in these two studies merely related to either the variability of input (i.e., input-only or modified input) or the amount of negotiation of that input.

One important aspect of any output activity is that output induces a certain level of engagement with the vocabulary taught. However, an important question needs to be asked – which engagement level is considered more beneficial for learning vocabulary? That is, should the engagement level be kept to a minimum so that the activity will not seem too demanding, especially for low proficiency learners? Or should it be elaborated so that more engagement is induced, and more learning occurs? Induced engagement is important, especially if we consider that teachers are concerned with improving the size of vocabulary knowledge of students. Some studies have examined the level of negotiation involved in output activities. Ellis et al. (1994), Gass (1988), Gass and Varonis (1994), and Loschky (1994) confirmed the greater impact of interactional modifications on lexical development (particularly nouns) in comparison to grammatical development. Nassaji and Tian (2010) and Storch (2008) examined the acquisition of vocabulary through completing activities both collaboratively and individually. These studies reported that a collaborative effort was more effective in the acquisition of PVs. Hence, all in all, collaborative activities were found to be effective in the acquisition of vocabulary.

Another question that needs to be asked, however, is whether FSs can be successfully learned through spoken pushed-output instruction. Overall, previous studies of lexical development through pushed-output were mostly noun-centred treatments. To

the best of my knowledge, only Nassaji and Tian's study (2010) has examined the effect of written output activities on learning FSs, represented in PVs. Nassaji and Tian's study (2010) examined two written output activities: a re-construction cloze activity and a re-construction editing activity, in two different settings: collaborative and individual. They reported a non-significant difference between collaborative output activities and individual output activities in promoting vocabulary knowledge. With regard to activities, they reported that editing activities led to a significantly higher gain in knowledge of PVs than did cloze activities in both settings, collaboratively and individually. However, Nassaji and Tian's main goal was to ascertain whether collaboration would be more beneficial to learning PVs than individual work.

Further, there is, unfortunately, a notable lack of research of pedagogical approaches addressing differences between FSs and SWs in the classroom. Laufer and Girsai (2008) examined three instructions: meaning-focused, contrastive form-focused, and non-contrastive form-focused instructions for learning SWs and collocations. They concluded that a contrastive form-focused group outperformed the other two forms of instruction. However, the study failed to address the question of whether FSs were acquired in a similar pattern of learning to SWs. Alali and Schmitt (2012) examined the differences in acquisition between SWs and idioms in an EFL learning context. They reported that reviewing was effective in teaching both words and idioms at receptive and productive levels; however, there were no differences in the learning gains between SWs and idioms. Peters (2014) examined differences in the written form-recall of SWs and collocations and reported that collocations were more problematic than SWs. Similarly, Kasahara (2010, 2011) reported the same difficulty of learning collocations over SWs; the difference was that Kasahara's studies examined the written recall of meaning. The findings of Peters (2014) and Kasahara (2010, 2011) suggest that learners often have problems with learning collocations, in either written form-recall or written meaning

recall, compared to SWs. To conclude, there is a clear need to examine different instructional methods that are effective for SWs for other types of FSs such as PVs.

Further, as mentioned in Section 2.1.3.2, what we know about the acquisition of polysemy is largely based on cognitive research. A few lexical studies have examined the teaching of polysemous lexical items (e.g., Bensoussan & Laufer, 1984; Khodadady & Khaghaninizhad, 2012; Macis & Schmitt, 2017; Morimoto & Loewen, 2007; Schmitt, 1998). Bensoussan and Laufer (1984) reported that learning polysemous words was difficult only in comparison to monosemous words. Schmitt (1998) concluded that knowledge of multiple meaning senses for a word is rarely possible, and the process of learning these words is patchy and slow, even for advanced learners. Morimoto and Loewen (2007), as well as Khodadady and Khaghaninizhad (2012), concluded that image-schema instruction might be a better pedagogical tool for teaching polysemous words than a translation-based one. Macis and Schmitt (2017) reported that acquiring the figurative meaning senses of idioms is a problematic feature, even for advanced learners. Some studies have suggested that polysemous words should be taught in a piecemeal fashion, one meaning sense per exposure (Shortall, 2002), whereas others have propounded all-at-once instruction in which all the literal and peripheral senses are presented together (e.g., Brodzinski, 2009; Csábi, 2004; Verspoor & Lowie, 2003).

The literature reviewed in the previous chapter and summarised above has shown that empirical evidence for the benefits of spoken pushed-output activities in an instructional context is scant. In addition, most previous studies have focused on the learning of SWs (mainly nouns), and thus our knowledge of the effectiveness of spoken pushed-output activities for other parts of speech and lexical items beyond SWs is rather limited. Importantly, we do not know whether the effectiveness of spoken pushed-output activities will hold true for polysemous lexical items. There is no doubt that knowledge of the diverse meaning senses of a lexical item indicates a broader knowledge of that item,

and perhaps this knowledge is the main difference between a NS and an EFL understanding of English (Schmitt, 2010). The current study is similar to Nassaji and Tian's study in that it examines pushed output in EFL vocabulary instruction for teaching multiword verbs. It also aims to investigate specific aspects of vocabulary knowledge: namely, how the form-meaning link incorporates meaning facets of a word (i.e., polysemy). In addition, it follows Alali and Schmitt (2012) in investigating the effectiveness of a teaching methodology (spoken pushed-output), that has been used for SWs, for the learning of FSs (particularly PVs).

3.2. Research questions

Study 1.1 was designed to answer the following questions:

1. Is there a difference between spoken pushed-output instruction and traditional instruction in vocabulary learning gains, either receptively or productively?
2. Are PVs learned at a similar rate to SWVs, using either spoken pushed-output instruction or traditional instruction, either receptively or productively?
3. Is there a difference in the acquisition of the three meaning senses of the target items within the two treatment conditions, either receptively or productively?

Based on previous research findings, it was expected that spoken pushed-output instruction would result in greater learning gains (e.g., De la Fuente, 2002, 2006; Ellis & He, 1999). Further, Pellicer-Sánchez 2020) suggested that, if the amount and type of exposure are the same for both SWs and FSs, the learnability of both might be similar. Given the fact that the amount and type of exposure of both SWVs and PVs in the present study would be similar, it was expected that the learning gains of the two would also be similar. With regard to differences in meaning senses, Garnier and Schmitt (2016), in

their examination of knowledge of polysemous PVs taken from the PHaVE list, reported that the first, most frequent meaning sense has an advantage over the other two. Thus, it was hypothesised that there would be differences between meaning senses, with the first, most frequent meaning sense being better learnt than the other two.

3.3. Methodology

3.3.1. Participants

Initially, there were 180 female participants who had graduated from secondary school and joined the Preparatory Year programme at King Abdul-Aziz University (Rabigh Branch) in Saudi Arabia (refer to Section 1.2 for more details about the programme). All participants had studied English for a minimum of nine years. They ranged in age between 18 and 21 years old.[2] The English proficiency level of the participants can be considered low to intermediate level within the context of the university, as indicated by their Vocabulary Level Test (VLT) scores (are presented in Section 3.4.1). One-hundred and six participants were excluded from the study for one or more of the following reasons: (1) they were absent in Week 1 (Ethics Approval, VLT, and the pre-test sessions); (2) they were absent from one or more of the eight treatment sessions, as this study examined the effectiveness of a teaching approach through continuous practice and missing one session would have a negative effect on learning gains; (3) they were absent from one or more of the post-test sessions each week; and (4) they had not reached the minimum score for mastering the 2K band in the VLT. After these exclusions, data from 74 participants were included in the analysis.

[2] Students should be between 18 and 21 years of age on admission to the Preparatory Year programme.

3.3.2. Treatment groups

The study design included three treatment groups: a control group (n=60), a traditional group (n=60), and a pushed-output group (n=60). The study was conducted with six classes, with each class being assigned to one of the treatment conditions (two classes per condition). After the appropriate deletions, the number of students in each condition was as follows: the control group (n=21), the traditional group (n=20), and the pushed-output group (n=33). The control group only completed the pre- and post-tests without receiving any experimental treatment. They received their usual language instruction, which did not include any explicit vocabulary activities on the target items or any pushed-output activities. The other two groups (traditional and pushed output) received the same amount of vocabulary instruction, with the only difference being the type of activities they engaged in (traditional vs. pushed-output). That is, both groups received a 10- to 15-minute presentation per session with a focus on target vocabulary. Afterwards, both had to practise the newly taught vocabulary for 50-60 minutes per session. The traditional group completed two activities: a matching activity and a fill-in-the-blanks with keys activity. The pushed-output group completed three activities: matching, sentence reconstruction, and use-in-sentence activities. These three activities were inserted into a spoken information-gap format. Both teaching materials and activities will be explained in Section 3.3.4.2.

3.3.3. Overall schedule

The study took place during normal class time (90 minutes) in the second semester of the Saudi university year, which ran from January to May 2016. The overall study involved six weeks. There were three teaching sessions each week. The timeline of the study is shown in Figure 3.1.

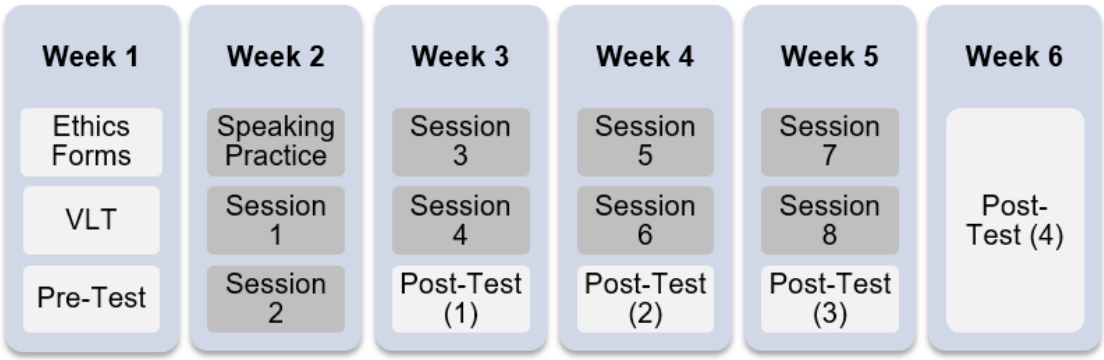


Figure 3.1 Timeline of the study

3.3.4. Materials

3.3.4.1. Target items

The following criteria were considered in the selection procedure: the part of speech, the frequency, and the polysemic nature of the items. The target items consisted of 24 polysemous PVs and 24 polysemous SWVs.

Selection procedure (PVs). The 24 PVs were chosen from the PHaVE list developed by Garnier and Schmitt (2015). The PHaVE list is a pedagogically-based list that presents 150 PVs along with their first, second, third, and fourth most frequent meaning senses (n=288). The PVs included in the list were identified by a previous study (Liu, 2011) as the most frequent in either the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA), at least 10 tokens per million words. The meaning senses of the PVs are presented with percentages in the PHaVE list, which indicates their minimum occurrences in COCA, accounting for at least 75% of all occurrences in that corpus. Certain criteria were considered in the selection of PVs for this study, such as enumerating PVs with only three/four meaning senses in the list, not enumerating PVs with meaning senses that overlap, and preferably not containing the same verb if possible. The last criterion was not feasible for certain common verbs, such as *go, get, take,* and *hold*. The target PVs in the present study are: *back up, break down,*

bring out, come around, cut off, fill in, get off, get through, give out, go along, go off, hold back, hold up, lay down, make up, move on, pass on, put up, set off, take in, take off, take up, throw out, and turn up. The final selection included some PVs including the same verbs (e.g., *hold back* and *hold up*); however, all PVs had a low degree of overlap in their meaning senses, particularly those PVs that overlapped in form. Also, the presentation and practice of these PVs occurred in different sessions.

Selection procedure (SWVs). The selection of target SWVs followed the same criteria as those for the compilation of the PHaVE list: items had to be polysemous and high frequency in the selected corpora. Working from this starting point, two lists of high-frequency verbs were obtained: one from the BNC, the other from COCA. The two lists were compared to compile a list of common high-frequency verbs in both lists, with only slight variations in the numbers of occurrences in the two corpora. For example, the verb *play* has 20,702 occurrences in the BNC and 22,803 occurrences in COCA. Function verbs were excluded from this compiled list as those verbs serve a more grammatical function than simply conveying semantic information, for example, *be, have, do,* etc. The Oxford dictionary was then consulted to check the polysemous status of the verbs selected. Verbs that were not polysemous in the dictionary were excluded, such as *think, say, want*, etc. The final step was examining the BNC's first 100 randomly generated concordance lines to examine the polysemic nature of the selected verbs and decide which were the most frequent senses, and to ensure that minimum coverage of all the senses of these verbs would be at least 75%. Some verbs that did not enumerate multiple meaning senses in the BNC were excluded, such as *mean, ask, keep, call,* etc. The final list consisted of 131 polysemous, high-frequency verbs selected only from the spoken section of the corpus. Twenty-four SWVs were selected from this list, avoiding any overlaps in the meaning senses between the SWVs and PVs. These SWVs are: *address, blow, charge, clear, collect, commit, count, cover, divide, enter, fall, hit, introduce, lead, lift, order,*

play, relate, run, save, serve, settle, train, and treat.

Distribution of the target items. With reference to Figure 3.1, the overall study design involved eight teaching sessions. In each session, three target SWVs and three PVs were taught (three meaning senses for each = 18 per session). There were some criteria to consider when distributing the target items over the weeks. As mentioned before, the 48 target items were distributed taking account of their meaning senses; that is, each week, all SWVs and PVs had a low degree of overlap in their meaning senses. Furthermore, as mentioned before, since having PVs starting with the same verb could not be avoided, these were not included in the same week.

Definitions of target items. The English definitions for the PVs were taken from the PHaVE list as it is the most recent pedagogically-based list that presents 150 PVs along with their first, second, third, and fourth most frequent meaning senses. The meaning senses in the PHaVE list were checked against multiple dictionaries, among them was Oxford dictionaries both British and World English. The English definitions for SWVs were taken from the online Oxford Learners' Dictionary. The lexical profiles of these definitions were checked using *VocabProfile* on the *Lextutor* website (Cobb, 2015) to ensure that the defining words were within the first 1,000- or 2,000-word families. Words in definitions that were less frequent were replaced by more frequent synonyms. These definitions were then simplified after consulting the Cambridge Learners' Dictionary online and the *WordReference* online dictionary.

Arabic one-word equivalents for both SWVs and PVs and their Arabic definitions were obtained online from *WordReference*, *Senseagent*, *Arabdict*, and *Lexicol*. These Arabic definitions were kept as short as possible. However, in some cases, these one-word equivalents were replaced by longer definitions consisting of an equivalent translation and some other significant elements for more comprehensive knowledge of that meaning sense (e.g., connotations, semantic preferences, collocations). These

definitions (n=72, per target type) included 41 one-word equivalents of SWVs, 31 long definitions of SWVs, 44 one-word equivalents of PVs, and 28 long definitions of PVs. This list of Arabic equivalents and definitions was needed for the instructional materials (i.e., designing the measurement instruments). The full list of target items with their English definitions, Arabic definitions, frequency values, meaning senses coverage, and their distribution over the weeks can be found in Appendix 1.

Sample sentences of target items. It was thought that compiling a list of sample sentences for each meaning sense of both SWVs and PVs was needed for designing the instructional materials (i.e., designing the presentation and the activities). Thus, a pool of 576 sample sentences (432 English, 144 Arabic) was obtained from the aforementioned online dictionaries: Oxford Learners' Dictionary, Cambridge Learners' Dictionary, *WordReference*, *Senseagent*, *Arabdict*, and *Lexicol*. Each target meaning sense had one Arabic sample sentence and three English sample sentences. The English sample sentences were also checked for lexical difficulty using *VocabProfile* on the *Lextutor* website (Cobb, 2015), ensuring that all words were within the first 2K.

3.3.4.2. Teaching materials

Target items presentation. The target items for each session were provided to the students by means of a PowerPoint presentation presenting six target items with three meaning senses for each. That is, it included three SWVs (nine meaning senses) and three PVs (nine meaning senses), each with an English definition, a visual stimulus to better illustrate the meaning, and an Arabic sample sentence (see Figure 3.2). The visual stimulus was either a picture or a video segment, and this was used to help consolidate the verb's meaning, as some actions were hard to grasp through a picture only, e.g., *commit a crime*.

Figure 3.2 Sample of target items' presentation

Activities (Traditional group). The 'traditional' treatment involves completing written vocabulary-focused activities that are typically used in the classroom in EFL context, and more specifically in the Saudi context. These activities are mostly form-focused activities which encourage participants to engage in written production by highly controlled activities, such as matching and cloze exercises with keywords. The activities for the traditional group were specifically developed for this study to resemble the activities commonly used in vocabulary textbooks and exams in most EFL contexts, and more specifically in the Saudi context. Two activities were designed: a matching activity and a fill-in-the-blanks activity. In the matching activity, participants were asked to pair the SWVs or PVs with a complement. These complements were selected from the sample sentences mentioned before. Before moving on to the second activity, students were provided with one sample sentence per phrase for the phrases they were exposed to in the matching activity. The second activity was a fill-in-the-blanks activity with the target items provided with their three meaning senses. The sentences in this activity were also taken from the database of sample sentences mentioned before. Figures 3.3 and 3.4, below, are examples of the two activities.

_____	1.	to lay down	A.	to a new place
_____	1.	to serve	A.	in the garden
_____	1.	to move on	A.	him with theft
_____	1.	to play	A.	your pen
_____	1.	to charge	A.	to France for a holdiay
_____	1.	to go off	A.	lunch

Figure 3.3 Sample of the matching activity for the traditional group

Fill in blanks with a word from the box below. Think carefully and remember that each word in the box will be used three times (×3):

lay down (×3)	serve (×3)	move on (×3)
play (×3)	go off (×3)	charge (×3)

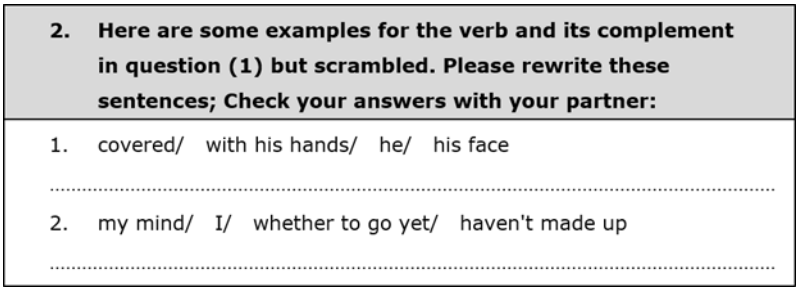
1. You need to forget about your past; just …….. and don't look back anymore.
2. The police may ……..the man with a crime.

Figure 3.4 Sample of the fill-in-the-blanks activity for the traditional group

Activities (Pushed-output group). The aim of setting spoken pushed-output activities in EFL vocabulary instruction is to increase learners' use of words as promoted by the design of these activities. That is, a spoken pushed-output activity does not mean an activity where learners are forced to produce output but rather a more encouraging

activity in which learners feel that they need to produce and practise the language (Swain, 1995). As defined in Section 2.3.3.2, spoken pushed-output vocabulary instruction in the present thesis is defined as spoken activities that are interactive (pair work), communicative (exchange information and give feedback) and encourage spoken production of the target items under time pressure. Further, as explained in Section 2.3.1, pushed output can be incorporated through different elements of the design, such as talking about unfamiliar topics, talking without planning, or some performance conditions, for example, time pressure, amount and nature of feedback, number of collaborations with others, etc. (Nation & Newton, 2008). The elements embedded in the activities for the current study were: nature of feedback (pair-work) and time pressure. This group completed three activities: a matching activity, a sentence-reordering activity, and a use-in-sentences activity. These activities were inserted into an information-gap format which allowed one learner's output to be another learner's input. The three activities were interactive since students worked in pairs (A and B) and they had to complete the activities together, sharing their answers with their partner. Both learners received a worksheet that included: the three activities as well as the answers for the activities of their partners, with a checklist for the third activity. The order of both activities and answers were altered in order to allow for reversing the roles when completing the activities. Hence, each student would practise the newly taught items and share information needed for their pair to complete the activity. The matching activity was exactly the same as the one used in the traditional treatment in which participants were asked to pair the SWVs or PVs with a complement, but only completed in pairs. This activity might not represent the features I chose for an activity to be a spoken pushed-output activity (+interactive, + communicative) but it was deemed appropriate to start be doing an activity that is very common in the Saudi EFL context, exactly the same as the on the traditional treatment, but in a different format (pair-work) just to help in setting the

scene for learners (i.e., working in pairs sharing information orally in order to complete the activities) . The sentence reordering activity required learners to put the words in order to make a sentence (see Figure 3.5). The sentences provided in this activity were the same as those included in the fill-in-the-blanks activity for the traditional group. In the third activity, they had to use each target item with its three meaning senses in sentences and share it with their partner. The vocabulary in all explanations and examples was checked to make sure that it was within the most frequent 2K. The activities were also piloted with two NSs to ensure that the sentences used in the activities were more native-like and did not stand out as odd, especially because some participants were going to practise them in spoken form. The sentences used in the activities were modified after piloting by either replacing them with new ones (n=40) or adding some words essential to grasp the meanings (n=36). All activities in both groups consisted of 18 items, each corresponding to 18 meaning senses of the target SWVs or PVs in each session. All the practice worksheets for each treatment group can be found in Appendix 2.



2. Here are some examples for the verb and its complement in question (1) but scrambled. Please rewrite these sentences; Check your answers with your partner:

1. covered/ with his hands/ he/ his face

.....

2. my mind/ I/ whether to go yet/ haven't made up

.....

Figure 3.5 Sample of the sentence reordering activity for the pushed-output group

To sum up, the two treatment groups differed in multiple aspects; precisely, the traditional instruction can be seen as form-oriented, highly controlled, written production activities with teacher feedback, whereas the pushed-output treatment involves more

meaning-focused, interactive (with peers) spoken activities. Further, the traditional treatment included two activities whereas the pushed-output treatment included three activities, which means that the overall time on one activity was different among the two groups. However, both groups spent the same amount of time learning the items, considering that the traditional group had the opportunity to engage with the words in the feedback for the matching activity on the PowerPoint slides. Learners were provided with one sample sentence per phrase for the phrases they were exposed to in the matching activity. Another fundamental difference is that learners completed the activities in the traditional group individually, whereas learners engaged in pairs while completing the activities in the pushed-output group. So, the two groups differed on multiple aspects: modality (spoken vs. written), feedback (teacher vs. peers), or number of the activities (two vs three times). Although this is considered as a limitation, the aim of the study is to compare the pushed-output instruction, as defined in Section 2.3.1, with the typical instruction that is used in the Saudi context, rather than isolating the effect of a particular variable. This limitation is further discussed in Section 3.6.

3.3.5. Measurement instruments

The instruments used in Study 1 were: the VLT (Schmitt, Schmitt, & Clapham, 2001), a receptive test (form-meaning recognition), and a productive/ oral test (form-meaning recall). The VLT was used to estimate the receptive vocabulary knowledge of the participants. The other two tests were designed for the specific purposes of the present study in an attempt to capture knowledge at both receptive and productive levels that might be differently facilitated by each of the treatment modes. The traditional treatment facilitated the development of receptive knowledge, while the pushed-output treatment emphasised the development of productive knowledge. Moreover, as the study concerns the acquisition of polysemy for target items, testing knowledge of the form-meaning link

was relevant. Forty-eight target items with three meaning senses each gives a total of 144 test items. In the next sections, these instruments are explained in detail. The administration procedures for the tests are explained in Section 3.3.6.

3.3.5.1. Vocabulary level test (VLT)

VLT was developed initially by Paul Nation (1983, 1990) to measure receptive knowledge of vocabulary. This test was then revised by Schmitt, Schmitt, and Clapham (2001), who developed the two latest versions of the test. It is a MC test consisting of four sections, each representing one frequency level of English word families (2K, 3K, 5K, and 10K). There are 30 items in each section presented in 10 clusters of six options in the left-hand column, and three corresponding definitions or synonyms in the right-hand column (see Figure 3.6). Test-takers have to match each definition in the right-hand column with one word in the left-hand column. The results of the responses on each level indicate the receptive vocabulary knowledge they have of that level.

1 birth	
2 dust	_____ game
3 operation	_____ winning
4 row	_____ being born
5 sport	
6 victory	

Figure 3.6 Cluster sample of VLT

VLT was used mainly because it is straightforward to administer multiple levels in one session. Versions 1 and 2 of the VLT (Schmitt, Schmitt, and Clapham, 2001) were used in this study. Only 2K and 3K VLT were used, as the expected level of the participants would be unlikely to allow them to master larger word families (Al-Nujaidi,

2003). The format of this test was an unfamiliar one for learners in the Saudi context; thus, the instructions were translated into Arabic to help learners complete the test. Care was taken to make sure that the participants understood the format of the test and how to complete it correctly, as the same format was used in the receptive test in this study (explained in the next section). Mastery of a level was considered when scores were above 24 out of 30 (80%), as suggested by Schmitt (cited in Xing & Fulcher, 2007, p. 184) and (Xing & Fulcher, 2007), though some studies suggested mastery score of 28/30 (93.3%) words. The two levels of the VLT used are shown in Appendix 3.

3.3.5.2. Receptive pre- and post-tests (form-meaning recognition)

Design. The instrument used for assessing the receptive knowledge of taught vocabulary was a meaning-recognition, MC test similar in format to the VLT. Given the large number of target items in the study, this test format reduced the number of distractors and allowed testing multiple target items within the same test item. Both pre- and post-tests were formatted in the same manner, but their administration was different. The pre-test was administered prior to the treatment to measure previous knowledge of the target items. The pre-test was presented in a 13-page test booklet. The first page of the booklet included a space for participants to write their name, student number, and their university section number. Then, instructions on how to complete the test were presented in both English and the students' L1 (Arabic), along with a test sample. The remaining pages presented the 144 test items. The test items were presented in 48 clusters consisting of 144 keys, 144 definitions, and 144 distractors. Each cluster was designed to assess the knowledge of three target items consisting of six options in the left-hand column: three keys and three distractors, and three L1 definitions in the right-hand column (see Figure 3.7). The same test was divided into four post-tests, each administered a week after the treatment sessions (see Section 3.3.6.3). Each weekly post-test was presented in

a four-page test booklet including the same first page of the pre-test (instructions and sample item), followed by 36 test items in 12 clusters consisting of 36 keys, 36 definitions, and 36 distractors. Learners were instructed to shade in the letter of the most appropriate option from the left-hand column for each definition in the right-hand column. I used a randomiser to create random patterns for both the order in which the six options appeared and the distribution of the keys in each test item (see Figure 3.8).

A. Play **B.** Charge **C.** Cover **D.** move on **E.** lay down **F.** take up	1. يضع شيئاً ما جانباً 2. يُحاسب/ يطلب مبلغ 3. يغُلف/يُغطي	1. (A) (B) (C) (D) (E) (F) 2. (A) (B) (C) (D) (E) (F) 3. (A) (B) (C) (D) (E) (F)

Figure 3.7 Sample of the receptive MC test

Week 1											
1	4	1	4	2	5	2	5	3	6	3	6
2	5	2	5	3	6	3	6	4	1	4	1
3	6	3	6	4	1	4	1	5	2	5	2
A	D	D	A	B	E	E	B	C	F	F	C
B	E	E	B	C	F	F	C	D	A	A	D
C	F	F	C	D	A	A	D	E	B	B	E

* Keys are the shaded boxes in the pattern

Figure 3.8 Sample of random order patterns for the test- items in week 1

One potential limitation of adopting a MC format for a test is that it sometimes tends to allow guessing or overestimates the knowledge of items. Learners could sometimes achieve correct answers by either guessing or eliminating the distractors,

rather than demonstrating their knowledge of an item (Gyllstad, Vilkaitė, & Schmitt, 2015; Stewart, 2014). However, this format is still used for assessing vocabulary knowledge due to its practicality and ease of administration, especially for large samples. To keep guessing to the minimum, three actions were taken. First, three distractors were included in each cluster, as having more options made guessing more difficult (Stewart, 2014). Second, all the test items (keys and distractors) were target items included in the weekly treatments. The reason for the inclusion of only target items is that learners were taught 12 target items per week and were tested on these items a week after the treatment (see Figure 3.1). If the distractors were not target items studied the previous week, they may have been easily picked up by the participants, which might have been a serious detriment to the reliability of the scores. Third, both written and oral instructions encouraged participants not to guess blindly but to choose answers they knew. The three definitions used in each cluster were provided in the participants' L1 (Arabic). Previous literature has shown that L1 definitions are more effective in conveying meaning senses than L2 definitions (Lado, Baldwin, & Lobo, 1967; Laufer & Shmueli, 1997). Moreover, Nation and Webb (2010) listed the main reasons for the effectiveness of using L1 definitions, including the use of well-established knowledge, the use of a shared lexical store for two languages in the early stages of language learning, and the need to avoid overloading participants with L2 listening and reading skills. The full post-test session was 90 minutes, and the time taken to complete the receptive post-test was 30 minutes within each session. The full test can be found in Appendix 4.

Piloting results and modifications. The receptive test was first given to a Saudi PhD student in Psycholinguistics. She was asked to provide feedback on any potentially problematic test items, options, and distractors. She correctly answered 100 items of the test; however, she had four incorrect answers and left 40 items blank. These 44 items for which she could not provide correct responses or did not provide any answer were

discussed in a face-to-face session to ensure that the reason was not any malfunctioning of the test items but rather just a lack of knowledge. She explained that the translations for the four items needed to be modified or expanded to include more details about their meaning. As for the other 40 items, she explained that she did not know the meaning senses for these items as they were less frequent, and she thought there were more frequent alternatives: for instance, *pass away* instead of *pass on*, and *move out* instead of *move on*. In the end, only ten items were modified, by either including more elements to better deliver the meaning (n=7) or modifying L1 definitions to make them more straightforward (n=3).

The second stage of piloting involved running the test with 32 students, all at the same university, enrolled on the Preparatory Year programme as the actual participants, with proficiency levels ranging from beginner to low intermediate. These students were instructed not to guess but to choose answers they knew. By doing this, the potential effect of guessing could be kept to a minimum. Test items were then divided into three categories after piloting based on correct responses: not problematic (over 75%, 32-25 correct responses, n=104), partially problematic (50-75%, 24-16 correct responses, n=23), and definitely problematic (under 50%, 15-0 correct responses, n=17). Forty items were then modified by adding more meaning elements to L1 definitions (n=25), changing L1 definitions either partially or completely (n=8), and altering the order of some items in the test (n=7). It is important to note that some of the participants who took the test in the piloting stage were at a lower level of proficiency than the actual participants, and thus they might have gotten some responses wrong due to their lower level, i.e., not necessarily because of any malfunctioning of the test. Yet, their responses were taken into consideration when modifying the test.

Scoring. This test was scored dichotomously by giving each answer either zero for an incorrect answer or one for a correct answer. An answer was considered correct if a

participant matched the form with the appropriate meaning in each cluster. It was considered incorrect if the participant chose an incorrect form for a meaning sense or if they left the question blank. The maximum score on the test was 144 (one for each meaning sense).

3.3.5.3. Productive post-test (form-meaning recall)

Design. The study used a meaning-recall, oral post-test to measure productive knowledge of the target vocabulary. Previous studies have used various formats to measure productive vocabulary knowledge either in spontaneous-production format (e.g., Peters & Pauwels, 2015) or more controlled-production format such as Vocabulary Knowledge Scale tests (e.g., De la Fuente, 2002, 2006) or translation tests (e.g., Peters & Pauwels, 2015). The purpose of the oral test in the present study is to measure the spoken form-meaning recall of target items and encourage learners to orally use the target items in context. Hence, using tests such as VKS would not reflect the main purpose of the study, beside its common limitations discussed extensively in the literature. Now, considering the learners' proficiency level as well as their lack of familiarity with spontaneous-speaking tests, adopting such tests might result in no learning at all. Hence, it was deemed more appropriate to choose a format that is semi-controlled but, at the same time, encourages some level of freer production. I decided to adopt a Discourse Completion Tasks (DCTs) test, a format which is less common in vocabulary studies but usually used in pragmatics, mostly in studies that examine the elicitation of speech acts, such as apologies, requests, etc. (Bassiouney & Katz, 2012; Blum-Kulka, 1989). DCTs in the present study were developed into culturally-related situations in which participants were asked to respond to these situations, preferably using one of the target items. Such format would help in eliciting the target items and encourage learners to use them orally. Adopting such format could be considered less demanding and more appropriate for the

proficiency level of the participants. At first, the idea was to design a test consisting of DCTs; however, developing a test with only DCTs was not feasible as some verbs held abstract meaning senses such as *divide* (calculate). Thus, for those items, factual questions were developed to elicit answers from the students. The format of the test included 144 test items: 83 DCTs and 61 factual questions. These factual questions and DCTs were presented in both Arabic and English. The Arabic translations were slightly colloquial so as to make them more engaging to the students, rather than presenting the test in standard Arabic.[3] The English ones were examined using *VocabProfile* on the *Lextutor* website (Cobb, 2015) to ensure that none of the vocabulary in the test items was outside the 2K-3K frequency bands. The weekly version of the test included 36 test items. As shown in Figure 3.9, due to the time-consuming nature of this test, it was only used as a post-test. The time taken to complete the productive post-test was 40 minutes within each session. The full test can be found in Appendix 5.

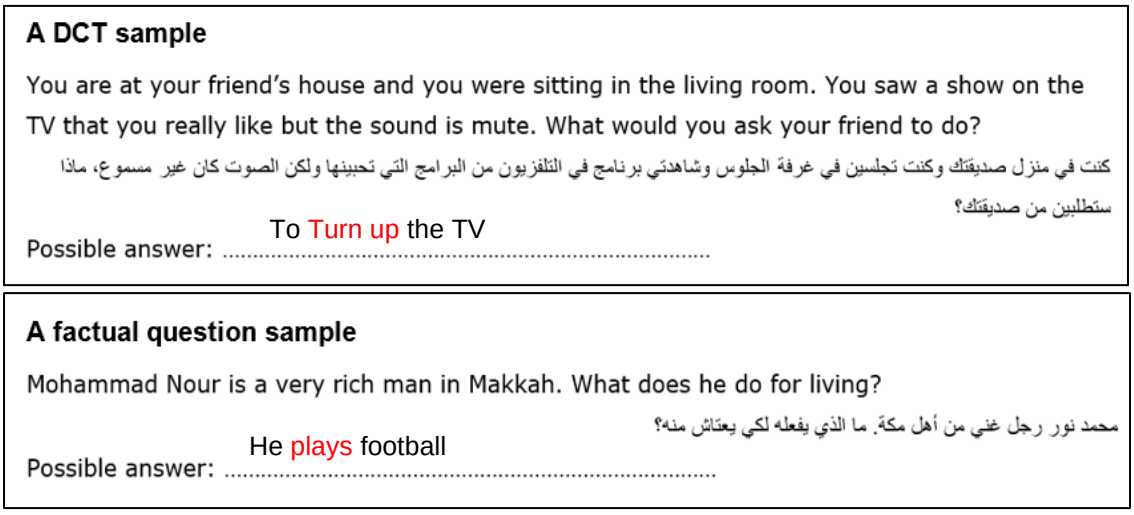


Figure 3.9 Sample of the productive oral post-test

Piloting results and modifications. The productive test was piloted with a

[3] Standard Arabic is the formal version which is used in formal situations and religious rituals.

bilingual English-Arabic speaker. As the design included culturally-related situations that participants could relate to and understand, she was asked to comment on any test items that she felt would not work well for learners. Results from piloting were classified according to the level of changes required:

- test items with no changes (n=86);
- test items which required modifying the context (n=20);
- and test items that elicited a more frequent non-target item for the specified meaning sense rather than target items (n=38).

Consequently, the context was slightly modified in 20 test items to be more informative of the target items. For those test items which could fail to elicit the target item, these could have been modified by either providing the first letter of the target item or asking participants to specifically use items they had learned in class, as suggested by Nation and Newton (2008). However, neither was used, as the idea behind this test is to help learners engage in a free production test, and neither of these solutions seemed to serve this purpose.

Transcribing and Scoring. The total time for recording was 160 minutes per student, which equates to 11,840 minutes of student recordings and 30,762 words. Two approaches were adopted for scoring the productive test: first, accounting for the occurrence of target items in the responses. Each item was scored by giving either one if the target item occurred in the response or zero if it did not. Care was taken when transcribing the data for possible alterations needed to allow for automatic scoring. For instance, instances of PVs which appeared in a non-adjacent form (e.g., *make* it *up*) were changed to the adjacent form (e.g., *make up* it) to allow for automatic scoring and analysis of the data. The total number of altered sentences was 64 (control=1, traditional=1, pushed output=62). Further, all verbs in the transcribed data were altered to the infinitive (e.g., *take* instead of *took* or *taken*). The total number of sentences with altered verbs was

132 (control=10, traditional=10, pushed-output=112). The second approach to analysing the data involved using measures to examine the lexical profile of the utterances produced in the tests. This analysis will be presented in Study 1.2 in Chapter 4. Figures 3.10 and 3.11 show samples of one learner's responses to test items for SWVs and PVs.[4]

1. they will play in the resort
2. I will learn how I play on the guitar
3. the people in Iraq divide into groups
4. he charge Ali to take care of horses
5. they commit him to hospital because he is crazy
6. I will cover him
7. my father will cover my expenses
8. I will divide the salary by months
9. he play football
10. I feel sad because they he did not commit it
11. I will serve food
12. I will cover them in writing
13. we serve in the marking
14. we will divide the price
15. the temperature hit fifty degree
16. he relate what he sees
17. we will hit the street barriers
18. I will order her to leave the room
19. I will order my books by colour
20. I can relate to you
21. please can you lift the bed
22. I will order the pizza
23. train students and teachers
24. don't blow your money
25. the price of the gold is fall
26. I will lead them to her office
27. I will clear the building
28. I will fall on the ground
29. she will address the students and talk to them
30. I will blow the candles
31. she will say that our new collection will be introduce next week
32. to introduce a new colour to the room
33. I will lead the group
34. I will fall asleep
35. I will address the box to her office correctly
36. don't blow your diet
37. I will introduce them to each other
38. I hope they will clear all the charges for my uncle
39. I will say this door will lead you to the cafeteria
40. they should settle this problem
41. I will say to him you should save your money
42. I will say health is count more than money
43. to save him
44. wood would be treat to make it waterproof
45. you need some time then you will settle
46. to collect myself
47. can you collect my things?
48. I will count all the children
49. it save money
50. he treat sick people and save them
51. I will settle in the sofa
52. I will tell him first you should enter your password and your username
53. he likes to collect stamps
54. I will say to her you should treat your grandmother with respect
55. I will advise her to enter the writing group

Figure 3.10 Sample of one learner's responses (SWVs) in the oral productive test

[4] Full transcripts of the productive test responses are available upon request.

1. I will change my clothes and lay down on bed in my bedroom
2. bombs go off in Paris on terrorist attack
3. I will move on to Jeddah
4. to make up for my mother
5. I will lay down the book and relax on my bed
6. to bring out the taste and be delicious
7. I will bring out a dish for her
8. the fire alarm will go off don't do it
9. the study take up all my time
10. you have to make up your mind on this
11. you can move on to another job if you want
12. twelve units make up the book
13. we have to go off to Dubai and relax
14. I will lay down rules to all groups
15. apple every year bring out new iPhones
16. you should move on the past
17. please turn up the sound
18. I will break down the lesson
19. fill in the cup
20. I will take off my jacket
21. your mom pass on to a better life
22. she will break down
23. sorry she take off early
24. she will come around
25. can you fill in
26. the bus break down
27. please pass on the paper
28. I can't get through to her
29. she will cut off her presentation
30. I will go along with my brother
31. we back up her
32. I will throw out my old clothes
33. sorry an accident hold up the road
34. the doctor will cut off his leg
35. sorry the lines was busy I couldn't get through to you
36. he must back up the case
37. before you come, everything were go along just fine
38. I will say back up the car to the exit
39. the line will cut off
40. I will hold up my hand
41. I will go along
42. we wait for them to give out news of Eid Alfitr
43. we must set off early
44. don't be take in by their lies
45. he easily get off the problem
46. you should put up shelves for your books
47. I will give out ads of the event
48. I will hold back my laugh
49. don't hold back anything
50. they set off bombs in mosques
51. we can put up some goals for my brothers
52. I couldn't take in all the information
53. it give out and stopped working

Figure 3.11 Sample of one learner's responses (PVs) in the oral productive test

3.3.6. Treatments and test administration procedures

3.3.6.1. Week 1

Introductory session 1 (Ethical forms). The first introductory session lasted for 90 minutes. All participants received an information sheet that included a detailed explanation of the research in the participants' L1. It started with a brief introduction to the research project, followed by a detailed explanation of the methodology and the rationale for the study. They were told that they would be taking vocabulary lessons on specific target items, probably not from their coursebook, and they would be assessed on what they learnt, with no indication of when they would be assessed. They were also

informed that the study involved some recordings which are not common in the Saudi teaching and examination context. Then, they were given consent forms to sign to give informed consent to participate in the study. They were given the opportunity to ask any questions about the study and all questions were answered. Participants were informed that it was not mandatory to participate, and they could withdraw from the study at any time without any repercussions for their studies or grades. All participants consented to participate in the study[5]. The information sheet and Ethics approval forms are presented in Appendix 6.

Introductory session 2 (VLT). The participants completed the 2K and 3K -word levels of the VLT (see Section 3.3.5.1). They were given 90 minutes to complete the two levels of the test, allowing them to spend the time they needed on each level. At the beginning of the administration procedure, participants were told that they were going to take a vocabulary test which would not affect their grades in the course but rather was a measure of how much vocabulary they knew. The VLT was introduced to the class with a PowerPoint presentation and questions about the test were answered.

Introductory session 3 (Pre-test). The pre-test was administered as a two-part test consisting of 72 test items each. Test booklets for the first part were given to all participants and they were allowed to read the instructions on their own before starting to answer, and any questions about the test were resolved. They were given 35-40 minutes to answer the first 72 items. Then, all test booklets were collected, and participants were given a 5- to 10-minute break before starting the second part of the pre-test, which took 35-40 minutes. The overall pre-test procedure took 90 minutes to complete.

[5] Copies of all signed consent forms are available upon request.

3.3.6.2. Treatment sessions

Speaking practice session. In week 2, before the participants started the treatment sessions, the oral post-test was modelled for the participating classes. The aim was to practise the format of the test. The speaking practice test included 36 test items not included in the post-test. Participants worked collaboratively as a group and were asked to give responses to these test items. Also, they were encouraged to ask questions about anything they did not understand or if they needed guidance on how to answer.

Treatment sessions. Each treatment session lasted for 90 minutes. I delivered the treatment sessions myself. The control group only completed the receptive pre-test and the receptive and productive post-tests. They were held on the same day as the other two groups had their tests. For the other two groups, the session started with a 10-15-minute lesson in which the target items for the session were presented. Then, both groups of learners spent 50-60 minutes completing the activities. In the traditional group, participants first completed the matching activity on their own, checking the answers with me, and reading some sample sentences provided for these verb-complement phrases in the PowerPoint presentation. Afterward, they completed the fill-in-the-blanks activity on their own and then answers were shown in a PowerPoint presentation. All the PowerPoint presentations can be found in Appendix 7. In the pushed-output group, they worked in pairs. They completed the same matching activity, the sentence-reordering activity, and the use-in-sentences activity. Then, they check and share their answers with their partner. The time both groups spent completing the activities was the same to ensure the same length of exposure to the target items in each group (see Table 3.1). One important thing to note is that both groups received feedback on their performance while completing the activities; yet, they differed on the type of feedback received (i.e., teacher feedback vs. peer feedback) as illustrated in Table 3.1.

Table 3.1 *Practice time for traditional and pushed-output groups*

Traditional		Pushed-output	
Matching activity	10 min	Matching activity	15 min
Feedback (teacher)	10 min	Feedback (peer)	
Fill-in-blanks activity	10 min	Sentence Re-ordering activity	15 min
Feedback (teacher)	15 min	Feedback (peer)	
		Use-in-sentences activity	15 min
		Feedback (peer)	
Total	45 min	**Total**	45 min

3.3.6.3. Post-test sessions

It was thought that, instead of having only one final post-test at the end of all treatment sessions, it might be better to simulate an authentic teaching class where learners would be taught some new vocabulary one week, and their learning is assessed the following week. Thus, the post-tests (receptive and productive) for the present study were administered in four parts. Each part was conducted a week after the treatment session and included items learned the previous week; that is, 36 test items per post-test. The productive post-test was administered first to avoid any priming test effect of the receptive post-tests on learners' production, giving participants the opportunity to demonstrate productive knowledge before receptive knowledge.

To begin the weekly testing session, recording devices were first distributed to all students. Instructions on how to use the devices and how to complete the test were presented in the participants' L1 in a PowerPoint presentation. Participants were instructed to state their name and their university section number before they started. They spent 40 minutes completing the productive test. Each situation was presented to the whole group on a screen and they were given one minute to respond. When participants needed more time for one item, they raised their hand so as not to disturb the others. The

test presentation was stopped for the whole group, giving them more time, a maximum of one more minute. After completing all the situations (n=36) in the productive test, the test booklets for the receptive test were distributed. The receptive test was completed individually, and participants took, on average, 30 minutes to complete it. They were given five minutes to review their answers before collecting all the booklets and recording devices. The whole testing session lasted 90 minutes.

3.3.7. Analysis

The following section reports the results of the receptive and productive tests. It first presents the results of the VLT and the pre-test. Afterwards, descriptive statistics for the pre- and post-tests are presented. Then, the results for receptive and productive gains are presented in Sections 3.4.3 and 3.4.4. The data were analysed using SPSS version 22 (IBM Corp., 2013). Absolute vocabulary gains were calculated by subtracting the means of pre-test scores from the means of post-test scores per participants. For receptive vocabulary gains, the means of the receptive pre-test were subtracted from the means of the receptive post-test. For productive gains, in the absence of productive pre-test scores, due to practicality issues, I had to calculate the gains using the receptive pre-test scores. Initially, a productive pre-test was also considered, but the time I was given to collect the data did not allow me to do so. The use of receptive pre-test scores for the calculation of productive gains involves the assumption that those words that were known receptively in the pre-test were also known productively. However, this is obviously not true, and it is well documented in the literature that receptive knowledge of vocabulary tends to be higher than that of productive knowledge (e.g., Chen & Truscott, 2010; Laufer & Goldstein, 2004; Laufer & Paribakht, 1998; Pellicer-Sánchez & Schmitt, 2010; Pigada & Schmitt, 2006; Schmitt & Meara, 1997; Vermeer, 2001; Webb, 2007, 2008). If we consider that receptive knowledge tends to be higher and that it was taken as a pre-test of

productive knowledge, I am actually adopting a more conservative approach in calculating productive vocabulary gains, as it is very likely that many of the words that were known in the pre-test at the receptive level were not known productively. However, in the absence of a productive pre-test, following this more conservative approach was considered more appropriate. It may perhaps involve an underestimation of potential gains, but at least I can be more certain of whatever gains are identified.

Further, relative vocabulary gains were also calculated. Relative vocabulary gains refer to "the difference between post-test and pre-test scores divided by the amount to be learned" (Shefelbine, 1990). Relative vocabulary gains take into consideration the varying opportunities between learners for increases in knowledge, whereas absolute gains do not (Webb & Chang, 2014). For instance, if a participant knew more target items prior to the treatment, then there was be less room for improvement. Relative gains were calculated using the formula: [(post-test score – pre-test score) / (total number of test items – pre-test score) × 100]. Again, in calculating productive relative gains, since there was no productive pre-test, the scores of the receptive pre-test were used. The scores for relative gains were used in the inferential statistics. One important thing to mention is that even though in the inferential statistics, reported in Sections 3.4.3 and 3.4.4, only relative gains were used for the reasons mentioned above, the same analyses were run using absolute gains and the same pattern of findings was found.

Prior to performing any test, the normality of the scores on the VLT, pre-test, and post-tests was checked using Kolmogorov-Smirnov tests, and the data turned out to be non-normally distributed. That means that the data should be analysed using nonparametric tests; however, the most suitable test for the data is the mixed-design ANOVA with repeated measures. There is no nonparametric equivalent to this test. When running multiple Kruskal-Wallis tests, the pattern of the results were similar to the ones shown after running the mixed-design ANOVAs; hence, it deemed more concise to report

the results of the mixed-design ANOVAs. For each type of knowledge (receptive and productive), a mixed-design ANOVA with two repeated measures was conducted. This included the three treatment modes (control, traditional, pushed-output) as the between-subjects variable, while the type of target item (SWVs vs. PVs) and the meaning sense (1, 2, or 3) constituted the within-subjects variables. Table 3.2 shows the different variables used and their levels in the two ANOVAs.

Table 3.2 *Variables used in the two mixed-design ANOVAs for receptive and productive relative gains between control, traditional, and pushed-output groups*

Variable	Type	Levels		
Treatment mode	Independent	Three		
		CG	TG	POG
Target type	Independent, repeated measure	Two		
		SWVs		PVs
Meaning-sense frequency	Independent, repeated measure	Three		
		Sense 1	Sense 2	Sense 3
Receptive relative gains (ANOVA-1)	Dependent			
Productive relative gains (ANOVA-2)	Dependent			

Note: CG = control group, TG = traditional group, POG = pushed-output group.

The assumption of sphericity was checked, and it indicated that none of the effects violated this assumption ($p.>.05$) for either ANOVA. Considering the results of a Levene's test in the two ANOVAs, they indicated that for ANOVA-1 (receptive data) the assumption of homogeneity of variance was violated for three variables ($p.<.05$): SWVs sense-1, SWVs sense-3, and PVs sense-2. For ANOVA-2 (productive data), the assumption of homogeneity of variance was not met for any of the variables ($p.<.05$). The

reason behind this might be, first, the inequality of treatment group sizes, as well as the extreme scores of some participants in the pushed-output group compared to the other two groups. I therefore decided to run Games-Howell post hoc tests, as suggested by Field (2013), for being the most accurate post hoc test when equal variance is not assumed. Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of .025. The effect sizes in this chapter are estimated according to Plonsky and Oswald's (2014) guidelines. Effect size is defined as "an objective and (usually) standardized measure of the magnitude of the observed effect" (Field, 2013, p. 79). Plonsky and Oswald (2014) argued for adopting more conservative field-specific points for effect sizes in the SLA context. They suggested interpreting *r* values as follows: $r > .25$ small effect; $r > .4$ medium effect; $r > .6$ large effect. This interpretation is chosen for the present thesis.

3.4. Results

3.4.1. VLT and previous knowledge of target items

The participants' VLT and pre-test scores were compared to ensure that all groups were homogenous in their vocabulary knowledge level and previous knowledge of the target items prior to performing any further analysis. All participants in the three treatment groups were at a low-intermediate level in the university context, as indicated by their VLT scores. They scored on average 25.09 (max= 27, *SD*= 0.86) on the 2K level, showing that most of them had mastered this level, and 6.66 (max= 17, *SD*= 3.42) on the 3K level, which indicated that they lacked sufficient vocabulary knowledge to master this level (see Table 3.3). A Kruskal-Wallis Test was used to examine potential differences in VLT scores across the three groups. The results showed no significant differences when analysing the scores per level, VLT-2K [X^2 (2, n=74) = .057, p= .972], VLT-3K [X^2 (2,

n=74) = .491, *p*= .782] and when analysing the combined scores for two levels, VLT-2K+3K [X^2 (2, *n*=74) = .669, *p*= .716]. The groups were therefore comparable in terms of their level of vocabulary knowledge. Another Kruskal-Wallis Test was conducted to explore potential differences in learners' previous knowledge of target vocabulary in the study (i.e., combined (SWVs+ PVs) pre-test, SWVs pre-test, and PVs pre-test). Results showed that there were no significant differences across the groups for their combined pre-test scores, [X^2 (2, *n*=74) = 3.469, *p*= .18], SWVs pre-test scores, [X^2 (2, *n*=74) =.854, *p*= .65], or PVs pre-test scores, [X^2 (2, *n*=74) = 4.191, *p*= .12]. Thus, not only were these groups matched in their vocabulary knowledge level but also comparable in terms of their previous knowledge of the target items. Table 3.3 shows a summary of the participants' scores on the pre-tests (combined, SWVs, and PVs) by treatment mode (control, traditional, pushed output).

Table 3.3 *Summary of VLT scores and pre-tests by treatment mode (control, traditional, pushed output) (SD presented in brackets)*

	CG	TG	POG
	M (SD)	M (SD)	M (SD)
Number of students	21	20	33
2K VLT (max=30)	25.05 (0.80)	25.15 (1.04)	25.09 (0.80)
3K VLT (max=30)	6.05 (3.04)	6.40 (2.93)	7.21 (3.90)
Combined VLT (max=60)	31.10 (2.95)	31.55 (3.24)	32.30 (3.70)
Combined pre-test (max=144)	9.00 (4.02)	8.70 (2.58)	9.52 (2.61)
SWVs pre-test (max=72)	6.33 (2.06)	6.20 (2.21)	6.55 (2.62)
PVs pre-test (max =72)	2.67 (2.52)	2.50 (0.89)	2.97 (1.02)

Note: CG = control group, TG = traditional group, POG = pushed-output group.

3.4.2. Descriptive statistics for pre- and post-test scores

Descriptive statistics for the receptive and productive tests (combined, SWVs, and PVs) are presented in Table 3.4. With regard to the differences in scores by treatment mode, the pushed-output scores were the highest, followed by traditional and then control scores. Looking across the target types, the scores on SWVs appeared to be higher than for PVs across the three treatment groups.

Table 3.4 *Descriptive statistics for pre- and post-test scores (combined, SWVs, PVs) for control, traditional, and pushed-output groups (SD presented in brackets)*

	Combined (max= 144)				SWVs (max= 72)				PVs (max= 72)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
Pre-test												
CG	9.00 (4.02)	8	3	17	6.33 (2.06)	6	3	10	2.67 (2.52)	2	0	8
TG	8.70 (2.58)	8	4	16	6.20 (2.21)	6	4	12	2.50 (0.89)	3	0	4
POG	9.52 (2.61)	9	3	17	6.55 (2.62)	7	1	14	2.97 (1.02)	3	1	5
Post-test (R)												
CG	44.29 (12.18)	42	24	77	25.57 (7.88)	24	12	40	18.71 (6.25)	18	10	37
TG	69.25 (20.50)	65	30	111	36.65 (10.87)	36	13	55	32.60 (10.72)	33	15	56
POG	117.91 (9.74)	117	96	135	59.21 (5.35)	60	45	68	58.70 (5.26)	59	45	67
Post-test (P)												
CG	15.52 (9.44)	12	3	38	12.33 (6.84)	10	3	30	3.19 (2.89)	2	0	10
TG	61.90 (21.05)	61	31	115	32.75 (10.22)	32	17	58	29.15 (11.14)	28	13	57
POG	104.61 (24.16)	106	58	140	52.79 (12.46)	53	27	71	51.82 (12.36)	52	29	70

Note: CG = control group, TG = traditional group, POG = pushed-output group. (R) = receptive, (P) = productive.

Tables 3.5 and 3.6 report descriptive statistics for the scores by meaning sense (1, 2, and 3). The descriptive statistics showed slightly higher scores being achieved for

sense-1 compared to the other two meaning senses (2 and 3).

Table 3.5 *Descriptive statistics for pre- and post-tests by meaning sense (SWVs) for control, traditional, and pushed-output groups (SD presented in brackets)*

	Sense-1 (max=24)				Sense-2 (max=24)				Sense-3 (max=24)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Pre-test											
CG	4.86 (2.78)	4	1	10	3.05 (2.27)	2	0	8	2.00 (1.55)	2	0	8
TG	2.00 (1.95)	2	0	7	2.05 (1.47)	2	0	5	2.15 (1.27)	2	0	4
POG	2.64 (1.54)	3	0	6	2.03 (1.42)	2	0	5	1.88 (1.11)	2	0	4
	Post-test (R)											
CG	10.19 (3.37)	10	3	16	8.29 (3.44)	8	3	16	7.10 (2.21)	7	3	13
TG	13.45 (4.24)	14	5	21	11.05 (3.59)	11	4	17	12.15 (4.34)	13	3	21
POG	20.42 (1.84)	21	16	24	19.39 (2.42)	20	13	24	19.39 (2.38)	19	14	23
	Post-test (P)											
CG	6.43 (3.20)	6	2	12	3.67 (2.56)	3	1	10	2.24 (1.84)	2	0	9
TG	10.85 (2.74)	11	6	16	10.30 (4.49)	10	4	21	11.60 (4.74)	11	4	21
POG	18.36 (3.87)	19	10	24	17.03 (4.77)	17	8	24	17.39 (4.49)	19	7	24

Note: CG = control group, TG = traditional group, POG = pushed-output group.
(R) = receptive, (P) = productive.

Table 3.6 *Descriptive statistics for pre- and post-tests by meaning sense (PVs) for control, traditional, and pushed-output groups (SD presented in brackets)*

	Sense-1 (max=24)				Sense-2 (max=24)				Sense-3 (max=24)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Pre-test											
CG	1.10 (1.18)	1	0	3	0.76 (0.77)	1	0	3	0.81 (1.21)	0	0	4
TG	0.85 (0.67)	1	0	2	0.85 (0.75)	1	0	2	0.80 (0.89)	1	0	3
POG	1.45 (0.56)	1	0	2	0.91 (0.77)	1	0	2	0.61 (0.75)	0	0	2
	Post-test (R)											
CG	6.57 (2.84)	7	3	16	6.52 (2.80)	6	2	11	5.62 (2.48)	5	3	12
TG	11.25 (3.73)	12	4	19	11.20 (4.23)	11	5	20	10.15 (3.60)	10	4	17
POG	20.18 (2.59)	20	13	24	19.21 (2.16)	19	15	23	19.30 (2.71)	19	14	24
	Post-test (P)											
CG	1.24 (1.34)	1	0	4	0.81 (0.87)	1	0	3	1.14 (1.49)	1	0	5
TG	9.85 (4.84)	10	2	19	9.75 (3.80)	10	5	19	9.55 (3.49)	9	4	19
POG	17.27 (4.13)	17	8	24	16.97 (4.65)	18	9	23	17.58 (4.38)	17	8	23

Note: CG = control group, TG = traditional group, POG = pushed-output group. (R) = receptive, (P) = productive.

Absolute vocabulary gains were then calculated. Tables 3.7 to 3.9 report the results for the receptive and productive absolute gains for the combined test (SWVs + PVs), by

type of target item (SWVs and PVs) and by meaning sense (1, 2, and 3).

Table 3.7 *Absolute vocabulary gains for combined, SWVs, and PVs for control, traditional, and pushed-output groups (SD presented in brackets)*

	Combined (max= 144)				SWVs (max= 72)				PVs (max= 72)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
CG	35.29 (10.37)	33	18	60	19.24 (7.33)	20	7	31	16.05 (5.05)	15	9	29
TG	60.55 (21.02)	60	14	100	30.45 (11.45)	29	1	46	30.10 (10.93)	31	13	54
POG	108.39 (10.14)	107	87	126	52.67 (5.66)	53	40	64	55.73 (5.48)	56	42	64
	Productive											
CG	6.52 (5.94)	5	-1	21	6.00 (5.53)	5	0	21	0.52 (0.93)	0	-1	2
TG	53.20 (20.75)	53	16	102	26.55 (10.04)	27	7	47	26.65 (11.19)	25	9	55
POG	95.09 (24.22)	97	48	131	46.24 (13.05)	48	19	65	48.85 (12.19)	49	24	67

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Table 3.8 *Absolute vocabulary gains by meaning sense (SWVs) for control, traditional, and pushed-output groups (SD presented in brackets)*

	Sense-1 (max=24)				Sense-2 (max=24)				Sense-3 (max=24)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
CG	5.33 (3.17)	6	0	12	5.24 (3.45)	5	0	13	5.10 (1.76)	5	1	9
TG	11.45 (4.87)	12	0	20	9.00 (4.03)	8	1	16	10.00 (4.48)	10	0	18
POG	17.79 (2.36)	18	12	23	17.36 (2.51)	17	12	23	17.52 (2.64)	18	12	22
	Productive											
CG	1.57 (1.80)	1	0	7	0.62 (0.97)	0	0	3	0.24 (0.54)	0	0	2
TG	8.85 (3.18)	9	4	16	8.25 (4.69)	7	0	19	9.45 (4.72)	9	3	20
POG	15.73 (4.04)	16	8	22	15.00 (4.89)	15	6	23	15.52 (5.01)	17	4	24

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Table 3.9 *Absolute vocabulary gains by meaning sense (PVs) for control, traditional, and pushed-output groups (SD presented in brackets)*

	Sense-1 (max=24)				Sense-2 (max=24)				Sense-3 (max=24)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
CG	5.48 (2.54)	5	2	13	5.76 (2.88)	6	1	10	4.81 (2.06)	4	0	8
TG	10.40 (3.86)	10	2	18	10.35 (4.43)	11	3	20	9.35 (3.54)	9	3	16
POG	18.73 (2.45)	19	12	23	18.30 (2.36)	18	13	22	18.70 (3.06)	19	12	23
	Productive											
CG	0.14 (0.48)	0	0	2	0.05 (0.38)	0	-1	1	0.33 (0.66)	0	0	2
TG	9.00 (5.01)	9	0	18	8.90 (4.02)	9	3	19	8.75 (3.43)	8	4	18
POG	15.82 (4.35)	16	6	24	16.06 (4.51)	16	8	23	16.97 (4.36)	17	7	23

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Relative vocabulary gains were also calculated. Tables 3.10 to 3.12 present the receptive and productive relative gains for all the groups for the combined test, by type of target item (SWVs and PVs) and by meaning sense (1, 2, and 3).

Table 3.10 *Relative vocabulary gains for combined, SWVs, and PVs for control, traditional, and pushed-output groups (%) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
CG	26.23 (8.10)	26	13	47	29.37 (11.46)	29	11	49	23.25 (7.83)	21	13	45
TG	44.69 (15.56)	43	11	75	46.12 (17.45)	43	2	73	43.26 (15.62)	44	19	77
POG	80.59 (7.25)	80	64	93	80.48 (8.05)	82	60	94	80.71 (7.65)	81	61	93
	Productive											
CG	4.94 (4.62)	4	-0.72	17	9.28 (8.72)	7	0	33	0.77 (1.36)	0	-1.41	3
TG	39.35 (15.65)	39	13	78	40.38 (15.75)	39	12	77	38.33 (16.06)	36	13	79
POG	70.71 (17.96)	72	36	97	70.50 (19.20)	72	30	98	70.81 (17.85)	71	36	97

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Table 3.11 *Relative vocabulary gains by meaning-sense (SWVs) for control, traditional, and pushed-output groups (%) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
CG	27.64 (15.61)	30	0	57	24.65 (15.92)	23	0	62	23.24 (7.98)	23	5	39
TG	51.43 (20.90)	51	0	87	40.64 (17.44)	37	5	68	45.63 (20.62)	47	0	86
POG	83.22 (8.88)	86	60	100	79.11 (10.51)	80	52	100	79.14 (10.82)	78	55	95
	Productive											
CG	8.26 (9.49)	5	0	35	3.07 (4.73)	0	0	14	1.17 (2.61)	0	0	9
TG	39.92 (12.88)	39	18	67	37.32 (20.90)	33	0	86	43.24 (21.71)	40	13	87
POG	73.57 (18.03)	75	38	100	68.20 (21.68)	68	30	100	69.70 (21.01)	78	19	100

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Table 3.12 *Relative vocabulary gains by meaning sense (PVs) for control, traditional, and pushed-output groups (%) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
CG	24.00 (11.65)	23	9	62	24.72 (12.31)	25	4	43	20.85 (9.53)	17	0	40
TG	44.83 (16.51)	46	9	78	44.55 (18.67)	46	14	83	40.33 (15.40)	38	13	70
POG	83.15 (11.25)	83	52	100	79.21 (9.52)	78	59	96	79.77 (11.89)	79	55	100
	Productive											
CG	0.64 (2.16)	0	0	9	0.21 (1.67)	0	-4	4	1.47 (2.93)	0	0	10
TG	38.72 (21.51)	37	0	78	38.28 (16.85)	37	14	79	37.73 (14.75)	35	17	78
POG	70.00 (18.61)	70	27	100	69.69 (19.83)	73	36	96	72.58 (18.91)	71	30	96

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Referring to Tables 3.7 to 3.12, the same pattern of variations reported for the pre- and post-tests means occurred for absolute and relative gains; that is, the pushed-output group was the highest, followed by the traditional group, and finally by the control group; relatively similar scores for SWVs and PVs; and some variations amongst the three meaning senses (1,2, and 3). Also, there were negative scores for the productive absolute and relative gains of the control group in PVs sense-2. This might be related to the conservative approach I adopted in the calculation of productive gains using the receptive pre-test scores; thus, in the productive post-test, some learners showed less productive gains than I estimated as their knowledge prior to the treatment. A summary of the relative

vocabulary gains by meaning sense for the three treatment groups (control, traditional, and pushed output) is presented in Table 3.13, which are used in the inferential statistics reported in the following sections.

Table 3.13 *Summary of relative vocabulary gains for control, traditional, and pushed-output groups (SD presented in brackets)*

		CG (n=21)	TG (n=20)	POG (n=33)
		M (SD)	M (SD)	M (SD)
SWVs sense-1	(R)	27.64 (15.61)	51.43 (20.90)	83.22 (8.88)
SWVs sense-2	(R)	24.65 (15.92)	40.64 (17.44)	79.11 (10.51)
SWVs sense-3	(R)	23.24 (7.98)	45.63 (20.62)	79.14 (10.82)
PVs sense-1	(R)	24.00 (11.65)	44.83 (16.51)	83.15 (11.25)
PVs sense-2	(R)	24.72 (12.31)	44.55 (18.67)	79.21 (9.52)
PVs sense-3	(R)	20.85 (9.53)	40.33 (15.40)	79.77 (11.89)
SWVs sense-1	(P)	8.26 (9.49)	39.92 (12.88)	73.57 (18.03)
SWVs sense-2	(P)	3.07 (4.73)	37.32 (20.90)	68.20 (21.68)
SWVs sense-3	(P)	1.17 (2.61)	43.24 (21.71)	69.70 (21.01)
PVs sense-1	(P)	0.64 (2.16)	38.72 (21.51)	70.00 (18.61)
PVs sense-2	(P)	0.21 (1.67)	38.28 (16.85)	69.69 (19.83)
PVs sense-3	(P)	1.47 (2.93)	37.73 (14.75)	72.58 (18.91)

Note: CG = control group, TG = traditional group, POG = pushed-output group.

(R) = receptive, (P) = productive.

3.4.3. Receptive learning gains

The results for the mixed-design ANOVA-1 showed that there was a significant main effect of the treatment mode on receptive relative gain scores, F (2,71) = 211.419, p <.001, η_p^2= 0.856, a large effect size, which indicates that the treatment groups differed significantly in their receptive gains. Games-Howell post hoc tests revealed that scores for the pushed-output group were significantly higher than for the traditional and control groups (both p_s. <.001), as shown in Figure 3.12. There was also a significant difference between the traditional and the control groups (p<.001). This suggests that even though traditional activities were better than the control group, which received no treatment at all, pushed-output activities were superior in improving vocabulary knowledge at the receptive level.

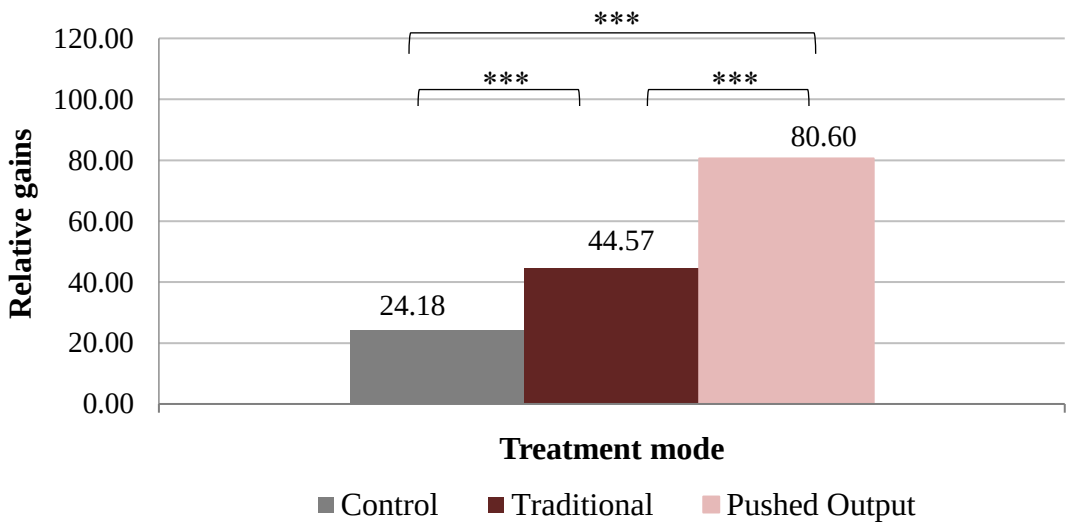


Figure 3.12 Means of receptive overall relative gains by treatment mode (control, traditional, pushed output)

The main effect of target type on receptive gain scores was not significant, indicating that the receptive gains of SWVs and PVs were similar, F (1, 71) = 1.759, p

=.189, η_p^2 = .024, a small effect size. The follow-up contrasts revealed that even though the mean scores of receptive gains for SWVs were slightly higher than for PVs, with only a mean difference of 1.476, 95% CI [-.744, 3.696], *p* =.189, this difference was not statistically significant. Moreover, there was no significant interaction effect between treatment mode and type of target item, *F* (2,71) = .704, *p* = .498, with a small effect size (η_p^2= .019). This suggests that the learning burden for SWVs and PVs is probably the same within the three instructional approaches used in the present study. Furthermore, the results showed that there was a significant main effect of meaning sense on receptive relative gains, indicating that the receptive gains of the three meaning senses were different, *F* (2, 142) = 7.598, *p* = .001, η_p^2= .097, a large effect size. Contrasts revealed that the mean score of sense-1 was statistically higher than that of sense-2, a mean difference of 3.565, 95% CI [.661, 6.469], *p* =.011, and sense-3, a mean difference of 4.217, 95% CI [1.169, 7.264], *p* =.003. However, the difference between meaning sense-2 and sense-3 was not statistically significant, with a mean difference of 0.652, 95% CI [-1.945, 3.248], *p* =1.00 (see Figure 3.13).

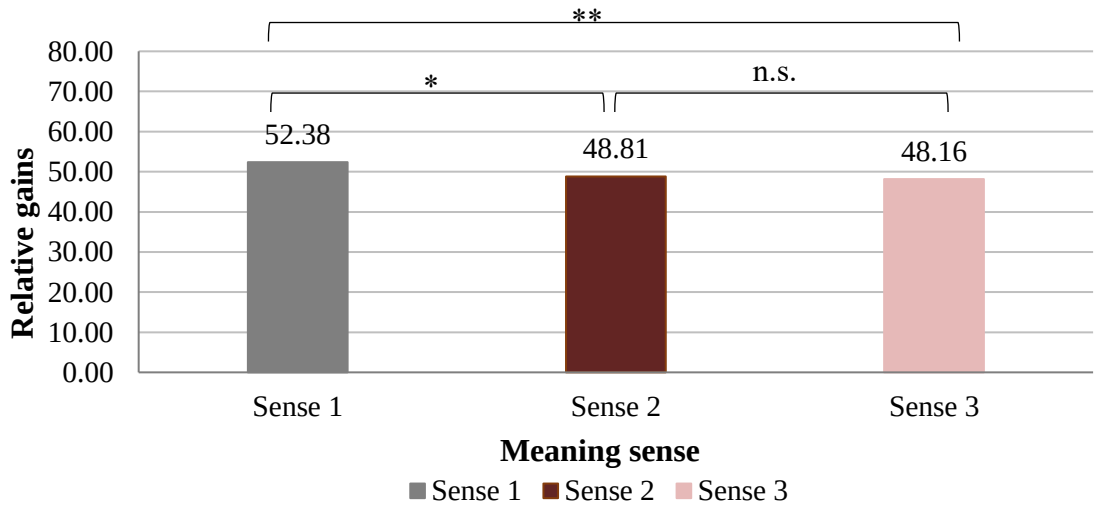


Figure 3.13 Means of receptive relative gains by meaning sense (1, 2, and 3)

There was no significant interaction effect between meaning sense and treatment mode, F (4,71) = .647, p. = .630, η_p^2= .018, a small effect size. The target item's type × treatment mode interaction effect was also not significant, F (2,142) = 2.728, p. = .069, η_p^2= .039, a small effect size. These results indicate that the differences in learning gains of the meaning senses (1, 2 and 3) reported above were not affected by either the treatment they received or the type of target items. Further, the meaning-sense × treatment mode × target item's type interaction effect was not significant, F (4,142) = 1.288, p. = .278, η_p^2= .035, a small effect size. Figure 3.14 presents the mean scores for the three-way interaction effect.

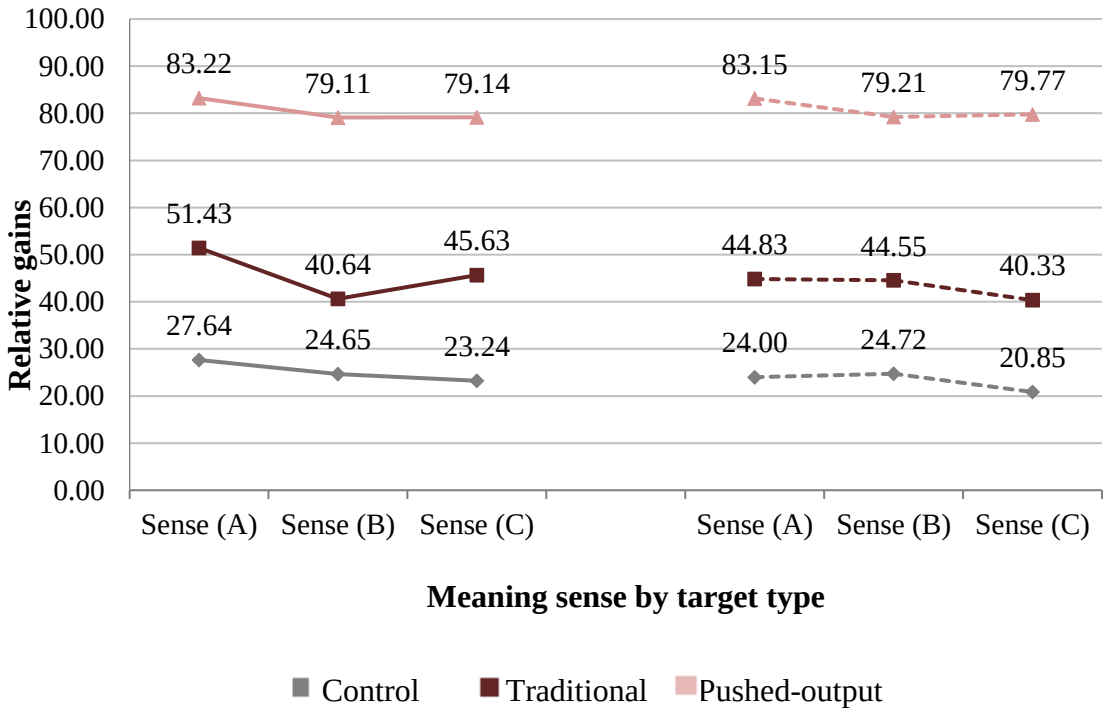


Figure 3.14 Means of treatment mode × target type × meaning sense interaction effect for receptive relative gains

3.4.4. Productive learning gains

The results for the mixed-design ANOVA-2 revealed that the main effect of treatment mode yielded an F ratio of F (2, 71) = 141.252, p < .001, with a large effect size (η_p^2 = 0.799), which indicates that the productive relative gains of the three groups were different. The follow-up Games-Howell post hoc tests showed that the pushed-output group outperformed the other two groups (both p_s. <.001), and the traditional group outperformed the control group (p <.001). Figure 3.15 shows that the mean scores of the pushed-output group were higher than for the other two groups, which proves that the pushed-output treatment had a superior effect over traditional instruction, not only in receptive gains but also in productive gains.

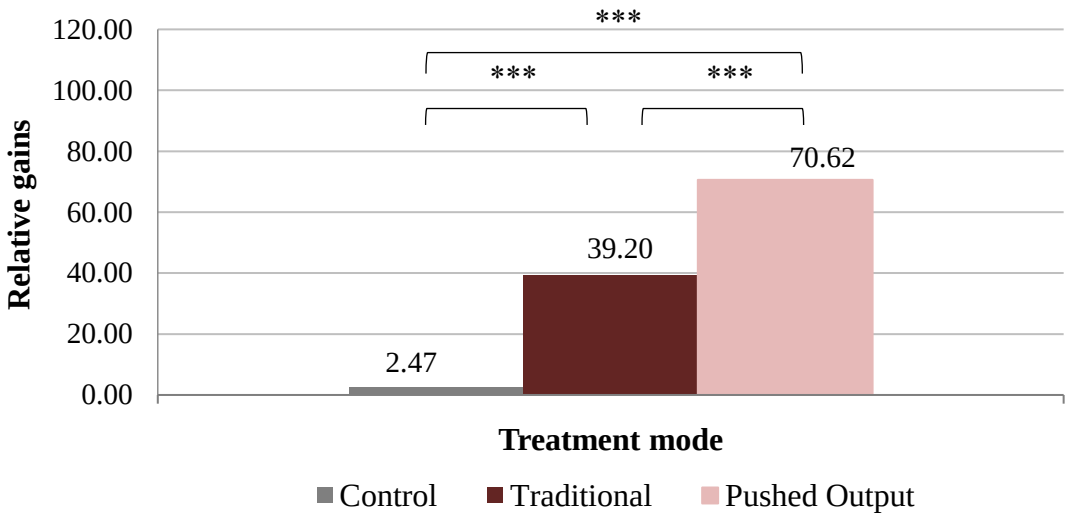


Figure 3.15 Means of productive overall relative gains by treatment mode (control, traditional, pushed output)

Moreover, there was no significant main effect of target type on productive relative gains, F (1, 71) = 3.918, p =.052, η_p^2 = .052, a medium effect size. Further, there was no significant interaction effect between target type and treatment mode, F (2, 71) = 1.772,

p =.177, η_p^2 = .048, a medium effect size. In this context, irrespective of the treatments used in the current study, participants recalled as many SWVs as did PVs. Also, the results showed that there was no significant main effect of meaning sense on productive relative gain scores, F (2, 142) = 2.999, p =.053, with a medium effect size (η_p^2 = .041), indicating that the means of the productive gains for the three meaning senses were very similar. Moreover, neither the meaning sense × treatment mode interaction effect, F (4, 142) = 0.799, p =.528, η_p^2 = .022, a small effect size, nor the meaning sense × target type interaction effect, F (2, 142) = 2.278, p =.106, η_p^2 = .031, a medium effect size, were significant. Further, there was no significant interaction between target type × meaning sense × treatment mode, F (4, 142) = 1.834, p =.126, η_p^2 = .049, a medium effect size (see Figure 3.16). This indicates that productive gains for the three meaning senses were similar for both target types across the three treatment modes.

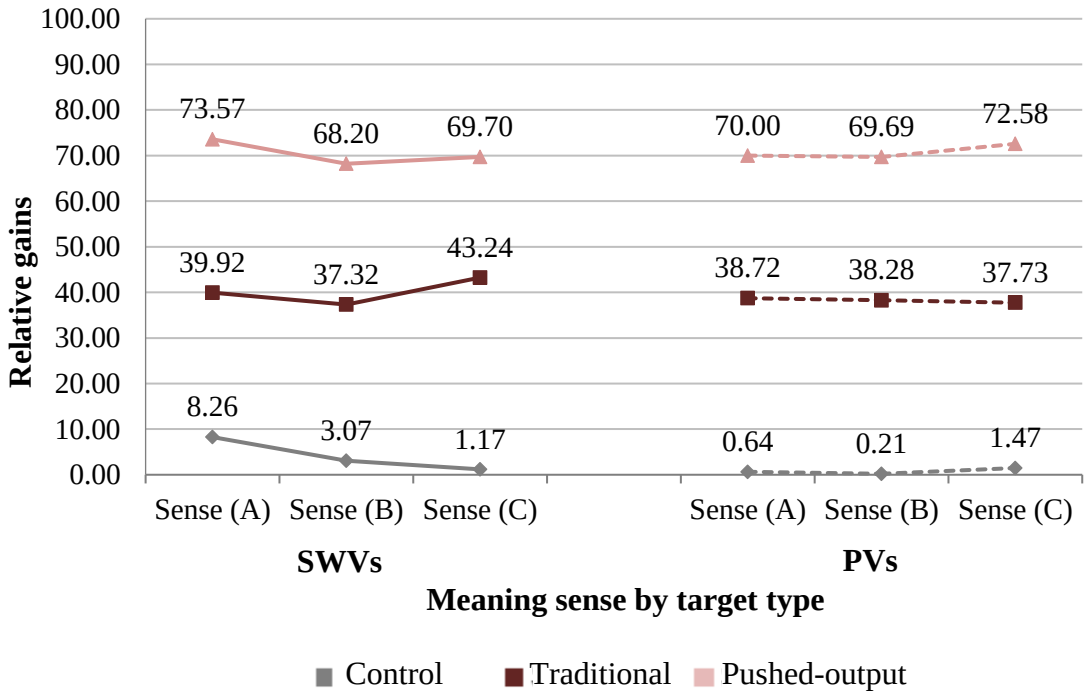


Figure 3.16 Means of treatment mode × target type × meaning sense interaction effect for productive relative gains

3.4.5. Further analysis

Section 3.4.1 showed that there were no significant differences in the scores on the VLT among the three treatment groups. However, it might be the case that small differences in participants' levels of vocabulary knowledge were related to receptive and productive learning gains. The potential relationship between VLT scores, receptive learning gains, and productive learning gains was explored by means of Spearman's rank-order correlations between the combined VLT scores (2K+ 3K) of the participants, the combined SWVs' and PVs' receptive relative gains, and the combined SWVs' and PVs' productive relative gains on the target items. All the correlations were non-significant, probably because there was not enough variation in the VLT scores. This indicates that participants with more vocabulary knowledge did not necessarily have better receptive or productive vocabulary gains compared to those with less vocabulary knowledge in the present study.

3.4.6. Summary of results

The results of the two mixed-design ANOVAs can be summarised as follows:

- Between groups:
    - Significant main effect for the treatment mode in which the pushed-output treatment outperformed the traditional treatment, and the traditional treatment outperformed the control treatment on both receptive and productive gains.
- Within groups:
    - No significant main effect of the target type (SWVs or PVs), which suggests that, under similar instruction conditions, PVs are likely to be learned at a similar rate to SWVs, either receptively or productively.

- Significant main effect of meaning sense on receptive learning gains; however, there was no significant main effect of meaning sense on productive learning gains. In this context, this indicates that the most frequent meaning sense of an item is best learned on a receptive level; however, this difference was much smaller between the three meaning senses with regard to their production.
- No significant interaction between the treatment modes (control, traditional, pushed-output) and the of target items (SWVs, PVs). This suggests that receptive and productive knowledge of SWVs and PVs were similarly affected by the treatment they received.
- No significant interaction between the treatment mode and the meaning senses of the target items (1, 2, and 3). This indicates that receptive and productive knowledge of more or less frequent meaning senses were affected by the treatment mode in a similar manner.
- No significant interaction between types of target items and meaning senses. This indicates that learning the three meaning senses did not differ by target type.
- No significant interaction between treatment mode × target type × meaning sense. This indicates that the three meaning senses of SWVs and PVs were learned in a similar manner both receptively and productively, and this was not affected by which type of treatment they received.

3.5. Discussion

This study has explored the effectiveness of spoken pushed-output instruction for learning polysemous SWVs and PVs, using two different tests: receptive (form-meaning

recognition) and productive (spoken, form-meaning recall) tests, with 74 EFL learners. The results of the study have provided evidence for the positive role of production in vocabulary teaching. In this section, the main results of the study are discussed in light of previous studies. The discussion of the main findings is organised around the three research questions: examining the effect of spoken pushed-output instruction, the differences between SWVs and PVs, and, finally, differences in learning multiple meaning senses of lexical items.

3.5.1. Receptive and productive learning gains in spoken pushed-output instruction vs. traditional instruction

The results from this study suggest that incorporating spoken pushed-output activities does indeed result in significant learning gains of form-meaning(s) link, not only at the level of recognition but also at the level of spoken recall, when compared to traditional teaching approaches. In fact, the recall gains could have been even higher than the ones reported but for the methodological decision discussed in Section 3.3.7. That is, in calculating productive (recall) gains, the scores on the receptive pre-test were used, since no pre-test measuring productive (recall) knowledge was administered. Ideally, it would have been better to have had a productive pre-test to better capture the gains in productive knowledge; however, that was not possible due to time constraints. The conservative approach that was adopted in calculating productive gains is likely to have underestimated actual productive gains and hidden some learning that occurred at the productive level. Thus, even when following such a conservative approach, there were gains in participants' productive knowledge and those gains in the spoken pushed-output treatment were higher than those in the traditional treatment.

Further, it is interesting to note, though, that without the implementation of some procedures that were suggested by Nation and Newton (2008) to encourage the learners'

usage of target items in a production test (such as providing the first letter of the word in the test, as mentioned in Section 3.3.5.3), the learners were able to use them substantially more in their utterances in the pushed-output group and substantially less in the traditional group. Another important thing to note is that the spoken pushed-output group only received peer feedback on their performance while completing activities and still outperformed the traditional group. This further strengthens the effectiveness of the spoken pushed output treatment over the traditional treatment, whose students received direct feedback from the teacher on their performance on the activities.

The results support Swain's Output Hypothesis (1985), which asserted the importance of output in L2 acquisition. The current study provides further empirical evidence for the benefits of pushed-output instruction for learning vocabulary, in line with previous studies. Peters and Pauwels (2015) found that the use of written output instruction had the potential to boost participants' recognition and recall of academic FSs. Nassaji and Tian (2010) concluded that the spoken pushed-output activities improved learners' acquisition of PVs. The present data extend these findings to the learning of polysemous FSs, suggesting that learners' recognition and spoken recall of the multiple meaning senses of FSs (particularly PVs) could benefit from utilising spoken pushed-output activities in EFL vocabulary instruction. With regard to SWVs, previous studies have agreed on the superior effect of spoken pushed-output treatments over other treatments on receptive and productive vocabulary knowledge (e.g., De la Fuente, 2006; De La Fuente, 2002; R. Ellis & He, 1999; Nassaji & Tian, 2010; Rassaei, 2017; Sun, 2017; Tahmasbi & Farvardin, 2017). However, most of these studies were noun-centred or examined vocabulary learning as a unidimensional rather than a multi-dimensional construct consisting of different, yet related, components, such as register, collocation and, in relation to this study, polysemy. Thus, the results of the present study extend those earlier findings to the learning of polysemous verbs.

Further, as discussed in Chapter 2, extensive research has been devoted to examining pushed-output instruction for vocabulary learning in a written mode (e.g., ; Nassaji & Tian, 2010; Peters & Pauwels, 2015; Rassaei, 2017; Sun, 2017; Tahmasbi & Farvardin, 2017). Only a few studies have examined pushed output in a spoken manner (De la Fuente, 2002, 2006; Ellis & He, 1999). Therefore, the current study has confirmed the efficacy of spoken pushed-output instruction for learning SWVs and PVs. In addition, the evidence for the superiority of pushed-output instruction in the oral recall of vocabulary in previous studies (De la Fuente, 2006; De La Fuente, 2002) was only measured through a productive version of the Vocabulary Knowledge Scale (VKS), which can be considered a fairly decontextualized test. The current study provides evidence for the effectiveness of pushed output in the spoken recall of polysemous SWVs and PVs utilising a partially contextualized oral test. If we are to compare the absolute gains in the present study with previous studies, learners' gains in the pushed-output treatment were nearly 75% on the recognition test and 67% on the recall test, which are similar to the percentages reported in Nassaji and Tian (2010) (around 70% for recognition knowledge and 60% for recall knowledge). However, the gains reported here are higher than those experienced by learners in the study by De La Fuente (2002), they only scored 50% on both levels of knowledge: recognition and recall. Perhaps this could be related to differences in the design of the pushed-output activities. In De La Fuente (2002), pushed output was combined with input, and in Nassaji and Tian (2010) a comparison was made between individual vs. collaborative work.

Further, some previous studies examined the effect of pushed-output activities when incorporated with either modified input (Ellis & He, 1999) or negotiated input (De La Fuente, 2002). It was not stipulated whether the superiority of pushed-output instruction was related to pushed output itself or variations in the input in the aforementioned studies. Even more, the possibility to create opportunities for pushed-

output was not limited to the output groups; it could be argued that in modified input/ negotiated input groups, opportunities for pushed output could also be created due to the need to modify or negotiate the input. Thus, the effectiveness of these treatments is not merely related to pushed-output instruction. In the current study, the opportunity for production was more salient in the pushed-output activities, the input for all treatments groups (traditional and pushed output) was controlled to be the same so it would not affect the results.

There are several explanations for the findings of the present study. First, the pushed-output treatment adopted in the current study comprised the three roles advocated by Swain (1995) which were operationalised through pushed output: noticing, hypothesis-testing, and metalinguistic function. From a lexical perspective, it also embraced the three psychological conditions for vocabulary learning advocated by Nation (2013): noticing, retrieval, and creative use. Participants were able to notice the form and meaning(s) of target items in the matching activity; test these vocabulary items within context in the sentence reordering activity; and produce sentences and evaluate the appropriacy of using these vocabulary items in context with their pairs in the use-in-a sentence activity. Further, all these activities were completed in pairs in a spoken manner, which also encouraged the retrieval of these vocabulary items and boosted the engagement with new vocabulary items. However, only noticing was offered in the activities employed within the traditional treatment. Consequently, participants in the pushed-output treatment were not only able to recognise the vocabulary in the receptive post-test but were also able to produce these vocabulary items within contextualised sentences, exhibiting more learning gains in the productive post-test. These findings lend support to both the three functions of output from Swain (1995) and the three psychological conditions from Nation (2013). Pushed output can induce noticing, hypothesis-testing, and metalinguistic function, all of which enable learners to engage in

deeper processing of language and are also beneficial for a higher level of engagement with new words.

Another possible explanation for these results may be related to the ILH. It is widely agreed that learning vocabulary in one's L2 relies on the extent of involvement induced by the activities used, as stated in the ILH (Laufer & Hulstijn, 2001). The higher the involvement induced by an activity, the better it is for the learning and retention of lexical items (Laufer & Hulstijn, 2001). The findings of the present study confirm the predictions of the ILH. The activities used in the pushed-output treatment induced a greater involvement load than that prompted by the activities in the traditional treatment. It could be argued that participants' involvement with the activities in the pushed-output treatment was further enhanced, since completion of the activities required spoken usage of the target items in a two-way information-gap format. This spoken production element in the activities could have induced a greater involvement load than the one in the activities of the traditional treatment. Further, though repeated retrievals occurred in both treatments, but it was a little higher in the spoken pushed-output treatment (three vs. two times). This small difference in number of retrievals could be one factor contributing to the highest ILs in spoken pushed-output treatment. These findings are in line with Tahmasbi and Farvardin's (2017) conclusions about the superior effect of a higher involvement load activity over other activities with lower involvement load indices for both recognition and recall of vocabulary. What Tahmasbi and Farvardin's (2017) study has in common with the current study is the use of relatively similar output activities; however, there are some discrepancies in the involvement load indices induced by the same activity. Particularly, Tahmasbi and Farvardin (2017) examined five output activities: paragraph-writing, sentence-writing, translation, fill-in-the-blanks, sentence re-construction, and form-meaning matching. Tahmasbi and Farvardin reported that both paragraph- and sentence-writing outperformed the other activities, and they related their effectiveness to their

higher involvement load indices. Even though two similar activities were employed in the pushed-output treatment in the current study (matching and use-in-a sentence), it could be argued that participants' involvement with the activities was further boosted since the activities were completed in a spoken manner, which induced more involvement, even for the matching activity.

It seems possible that the superiority of the pushed-output treatment used in the present study is due to the fact that pushed output can be "an internal attention-drawing device" (Izumi, 2002, p. 543), as the learners' attention to vocabulary increases according to their need to produce language. This is also in line with Schmidt's (2001) Noticing Hypothesis, which suggests that noticing and attention are essential for learning to take place. As Schmidt put it, "people learn about the things that they attend to and do not learn much about the things they do not attend to" (2001, p. 30). Pushed-output activities could have boosted the learners' attention to both form and meaning(s). That is, they could have drawn the participants' attention more to both semantic and grammatical features of target words, for example, polysemy, pronunciation, and use in context, which might have led to better recall of the meaning(s) of words.

Furthermore, this increased level of attention, together with the need to use the language to communicate with their partners, in the pushed-output treatment might have also led to more elaborate processing. A pushed-output activity is a two-way activity example in which learners are expected to exchange information for the activity to be completed. This is in contrast to the activities used in the traditional treatment, in which the words were presented first and then learners practised those words in one-way activities with no output involved, which only triggered limited processing of these words. Folse (2006) demonstrated that two-way activities facilitate negotiation of word meaning, which can lead to lexical production and the development of spoken productive knowledge.

Furthermore, another possible explanation for the findings of the current study is related to the TAP framework. The mode in which the target items were learned seems to have some effect on the way these items were orally retrieved. Learners in the spoken pushed-output instruction were offered opportunities to orally use the target items in context in the learning stage whereas the traditional instruction did not. Thus, perhaps learners in the spoken pushed-output instruction performed better in the testing stage which required the spoken retrieval of the target items due to the previous oral practice in the classroom.

Overall, the results are encouraging. Learners were able not only to manifest higher levels of receptive gains in the pushed-output treatment, but also to use these words orally in the production test to a higher degree than learners in the traditional treatment. These findings suggest that teaching vocabulary with the means of spoken production is feasible and might even leads to better gains than traditional vocabulary instruction on both receptive and productive levels of mastery.

3.5.2. Receptive and productive learning gains by target type (SWVs and PVs) within the spoken pushed-output and traditional treatments

The second question in this study is concerned with the differences in learning rates between SWVs and PVs. The results revealed that, regardless of the treatment mode, the learning gains of SWVs and PVs were not significantly different, at neither the receptive level (TG= 33.45 vs. 28.91, POG= 75.59 vs. 71.26) nor the productive level (TG= 26.36 vs. 23.33, POG= 63.42 vs. 64.43). These results suggest that SWs and multiword items are learned at the same level, which contradicts previous findings that FSs seem to be particularly challenging for learners (e.g., Kasahara, 2010, 2011; Peters, 2014). These seemingly contradictory results may be explained by the fact that the difficulty in learning FSs, and PVs in particular, as reported in the literature, might not be due to the intrinsic

characteristics of these verbs but rather to the amount of exposure in general, and more specifically the amount of explicit instruction devoted to them. Previous studies have suggested that a great deal of exposure is needed to acquire FSs (Adolphs & Durow, 2004; Schmitt, 2004). When both SWVs and PVs received equivalent amounts of exposure, equivalent levels of engagement, and equivalent amounts of practice in the current study, PV learning was parallel to SWV learning in terms of recognition and recall.

With regard to teaching PVs, the findings from the current study do not support the assumption that PVs are considered a burden for EFL learners (Boers, 2000; Hulstijn & Marchena, 1989; Kao, 2001; Kurtyka, 2001; Littlemore & Low, 2006; Moon, 1998). Those previous studies based their claims that PVs were more difficult on structural property, syntactic peculiarity, semantical property, and semantic complexity (see Section 2.2.3.4). However, the findings of the present study suggest that polysemous PVs (including those with idiomatic meanings) were not necessarily difficult for learners in the present study and, therefore, the learning of FSs seem to be no different from SWs, provided that the same type and amount of exposure is given.

Further, several previous studies reported a tendency of learners to avoid PVs in a spoken context more than a written one (Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004; Siyanova & Schmitt, 2007). Dagut and Laufer (1985) and Liao and Fukuya (2004) concluded that EFL learners preferred to use SWVs and avoid PVs, especially figurative PVs. They argued that this avoidance could be linked to the structural differences between the students' L1 (e.g., Hebrew, Chinese) and L2 (English), or to the proficiency level of learners (intermediate). Regarding avoidance behaviour, Hulstijn and Marchena (1989) reported that their advanced and intermediate Dutch EFL learners did not show a categorical avoidance behaviour but rather a tendency to adopt "a play-it-safe strategy" (Hulstijn & Marchena, 1989). The findings of the present study show that learners produced PVs with a similar

level of accuracy to that of SWVs. However, learners in the treatments were forced to use the target items; thus, no claims about avoidance behaviour in freer production can be made. It would be interesting though to see if, after having received the same type and amount of exposure, learners would still show a similar usage of PVs in activities that give them an opportunity to avoid using PVs.

The findings of the present study are in accordance with Alali and Schmitt's (2012) conclusion that idioms are learned at a similar rate to that of individual words on the recognition level. Also, these findings support the conclusion offered by Pellicer-Sánchez (2020) that similar learning of SWs and FSs is likely to occur if we account for the number of encounters, the approach of instruction, and the level of attention/ noticing of both being similar. However, they do not support the findings of previous studies indicating the difficulty of FSs over SWs (e.g., Kasahara, 2010, 2011; Peters, 2014). Peters (2014) reported that collocations were more difficult to learn than SWs, as the recall of the written form of SWs was significantly better than that of collocations in immediate, one-week delayed, and two-week delayed post-tests. Kasahara (2010, 2011) found that written recall of the meaning of SWs was significantly better than that of collocations. This discrepancy may be attributable to the type of FSs examined. In both Peters (2014) and Kasahara (2010, 2011), collocations were examined (known-and-unknown collocations in Kasahara's studies and verb+ noun/ adj. + noun in Peters' study), whereas high frequency phrasal verbs were used in the current study. As Wray (2019) pointed out, FSs can be divided into two categories: ones that can be replaced by a single word and those that cannot. For those in the former category, the difference in learning gains between FSs and single-word items may be minimal (Wray, 2019). The PVs used in the present study could be replaced by SWVs in the majority of cases, whereas collocations cannot be replaced by single-word forms. Thus, this could explain the similarity in the learning gains of SWVs and PVs found in the present study, and the higher degree of difficulty in

learning collocations in Peters' (2014) and Kasahara's (2010, 2011) studies.

To conclude, the results of the receptive and productive gains for both SWVs and PVs appears to be similar within the two types of treatments in the present study: traditional and spoken pushed-output. These findings suggests that learning FSs is likely to be similar to that of SWs, given that both received the same amount and type of instruction (Pellicer-Sánchez, 2020).

3.5.3. Receptive and productive learning gains by meaning sense (1, 2, and 3) within pushed-output and traditional treatments

In answer to the third research question, the results displayed differences in the recognition scores for the three meaning senses of the tested items, with the first, most frequent meaning sense being better recognised in SWVs (TG= 51.43 vs. 40.64 vs. 45.63, POG= 83.22 vs. 79.11 vs. 79.14) and PVs (TG= 44.83 vs. 44.55 vs. 40.33, POG= 83.15 vs. 79.21 vs. 79.77). This difference was much smaller for the recall scores of the meaning senses in SWVs (TG= 39.92 vs. 37.32 vs. 43.24, POG= 73.57 vs. 68.20 vs. 69.70) and PVs (TG= 38.72 vs. 38.28 vs. 37.73, POG= 70.00 vs. 69.69 vs. 72.58). Garnier and Schmitt (2016) tested the knowledge of multiple meaning senses of polysemous PVs chosen from the PHaVE list, without providing any instruction. Their results showed that the first, most frequent meaning sense was better known by learners than the other two (or more) meaning senses. They also concluded that learners' knowledge of polysemous PVs was insufficient and consequently needed more attention in the L2 classroom. The present study is one of the first studies to have empirically used the PHaVE list in the classroom, and the findings indicate that even when providing two types of instruction (traditional and pushed-output), the same pattern of knowledge occurred, with the first meaning sense being learned better than other meaning senses of a PV, only on the recognition level.

This pattern of learning the first, most frequent meaning sense can be explained in terms of the "frequency effect", and theoretically related to the common observation of a usage-based approach to language learning; that is, the most frequent items are learned first (Ellis, 2002; Gries, 2008; Waring & Nation, 1997). The frequency of exposure provided for the three meaning senses in the study was the same. However, while it is unlikely that learners had exposure to the least frequent meaning senses in their coursework or textbook, they might have had some initial previous exposure to the most frequent meaning senses in their coursework or textbook (refer to Section 3.3.4.1 for a full description of the selection of target items). This exposure might have been enough to boost the effect of the treatments for the receptive gains of sense-1 more than for the other two meaning senses; yet, it was not enough to show up in the pre-test scores (as shown in Tables 3.5 and 3.6) nor to show up in the productive gains. This difference between receptive and productive learning gains was to be expected, as productive knowledge tend to be more complex than other components of word knowledge (Cheng & Matthews, 2018).

Further, we could speculate that this slight advantage in the acquisition of sense-1 could be related to the fact that the first meaning sense is perhaps the literal meaning sense of an item. For instance, for *hit* and *hold up*, the first meaning senses are the most literal, as shown in Table 3.14. Among these three meaning senses, for both *hit* and *hold up*, sense-1 scored higher than the other two meaning senses.

Table 3.14 *Percentage of learners who responded correctly to the receptive post-test by meaning sense (1, 2, and 3) for traditional and pushed-output groups*

	Meaning sense	TG	POG
hit	1. to touch something quickly and with force	50%	75%
	2. to affect something/someone badly	25%	48%
	3. to reach a place, position, or state	20%	48%
hold up	1. to raise something to a high position so it can be seen	60%	87%
	2. to remain strong /in a good condition after a bad period	25%	57%
	3. to delay/ prevent the progression of something/someone	15%	57%

Note: TG = traditional group, POG = pushed-output group.

Despite the fact that the first meaning sense was acquired better on the recognition level in both conditions, there was also noticeable learning in the other two meaning senses in both conditions. These findings suggest that learning polysemous items at one exposure is possible to some extent, but gains might be lower for the least frequent meaning sense(s). These findings are consistent with Morimoto and Loewen's (2007) and Khodadady and Khaghaninizhad's (2012) conclusion that learning polysemous items is not difficult, though different experimental instructions were used. Both Morimoto and Loewen (2007) and Khodadady and Khaghaninizhad (2012) examined image-schema-based instruction and translation-based instruction for learning polysemous vocabulary (specifically verbs and prepositions), while the current study examined two different approaches, i.e. pushed-output instruction and traditional instruction. This proves that the instruction approaches used in this study are also effective for the teaching and learning of polysemous items at one exposure.

The pedagogy of polysemous vocabulary, SWs or multiword units, has been neglected in previous literature. Most previous studies have examined polysemy from a

more cognitive perspective. For example, Csábi (2004) concluded that explicitly raising learners' awareness of cognitive conceptual mechanisms, such as conceptual metaphors and conceptual metonymies, led to building a better semantic network of targeted polysemous lexicons. The current study provides empirical evidence that explicit instruction, a type of awareness-raising strategy, of the multiple meaning senses of an item is not detrimental but rather beneficial for learning the different meanings correspondingly. Csábi (2004) favoured the all-at-once pedagogical approach and concluded that associating figurative (secondary) meanings with their literal (core) meanings was advantageous for learning them, rather than just presenting one meaning sense per exposure. Boers (2000) demonstrated that learners would benefit from having all the meaning senses together, as learning the core literal meaning sense could encourage them to infer and interpret the other figurative senses on its basis. Verspoor and Lowie (2003) also supported integrating the core sense for an item with its figurative senses at the same time for long-term retention of all the meaning senses. In contrast, Shortall (2002) suggested including only one sense per exposure for explicit instruction. This suggestion may have overlooked the fact that all meaning senses could have been interrelated on semantic grounds and therefore perhaps separating the senses could result in more misunderstandings of these polysemous lexical items. The present data support the claims for an all-at-once approach. Though the design of the present study does not allow me to claim that the learning multiple meaning senses together is better learning them separately, but what I can claim with certainty is that presenting, and teaching literal and peripheral meaning senses together led to considerable receptive and productive gains in all the meaning senses. Future research should compare the learning of the three meaning senses of SWVs and PVs together and separately.

In conclusion, the findings suggests that learning multiple meaning senses together is not detrimental and could be fruitful for better recognition and recall of the multiple

meaning senses of the target items, SWs or FSs.

3.6. Limitations and conclusions

One possible limitation of the present study was the absence of an immediate post-test. It is possible that part of the learning that accrued from the treatments might have been lost by the time of the delayed post-test. However, scores in the delayed post-test provide a more reliable measure of real gains. Having both an immediate and a delayed post-test was not deemed appropriate due to the potential test effect. In addition, having only one post-test session was considered to offer a better reflection of authentic pedagogical practices, which was an important aspect of the design of this study. Replication of this study utilising an immediate post-test could provide a different picture of the learning process.

A second limitation is the selection of only PVs to represent FSs, which did not account for other types of FSs, such as collocation, idioms, etc. More research is needed to examine the efficacy of spoken pushed-output instruction in learning other types of FSs.

Another limitation was the absence of an oral pre-test measuring productive knowledge of the target items, as explained in Section 3.3.7. It would have been better to administer a more controlled oral productive pre-test (e.g., translation tests); however, it was very difficult to administer such a test to the initial sample consisting of 120 learners. Consequently, the productive vocabulary gains were calculated from the receptive pre-test scores, which might be considered an over-conservative approach. The amount of vocabulary gains made at the productive level may be slightly underestimated when using receptive pre-test scores in the scoring of productive gains. It is expected that with a pre-test of productive knowledge, pre-test scores would have been lower and therefore the

gain scores would have been higher. The findings from the productive gains should therefore be interpreted with caution. It might be useful for future studies to use an oral productive pre-test of less time-consuming nature to determine the potential extent of productive gains. However, having one pre-test has been also claimed to have the advantage of controlling for a potential test effect. Having only one pre-test has been a characteristic of previous vocabulary studies. For example, Webb and Kagimoto (2009) argued that that not having a productive pre-test minimises test effects. It would be difficult to ascertain whether learners have knowledge of the target items if they have answered incorrectly as productive tests often allow for multiple correct answers which are semantically related to the target items (Webb & Kagimoto, 2009), such as the DCTs.

Turning to the design of the teaching materials, it is important to acknowledge that the treatment materials used both static and dynamic clips as visual stimuli. While this was the same for both treatment conditions and therefore should not affect the pattern of results and main comparisons made, future studies should look at the effect that such presentation of different types of visual stimuli has on learning gains. Furthermore, the pushed-output activities used in the present study provided learners with opportunities to produce the target vocabulary in a controlled manner, and perhaps allowed for some negotiation between the participants; yet, I acknowledge the fact that the negotiation level of the activities used in the present study is less than what Swain (1995) encouraged implementing within the design of pushed-output activities. The idea behind this was to present activities that pushed learners to produce taught vocabulary in a gradual style, that is, first just saying a lexical item, then practising it in a sentence, and finally producing contextualised sentences with the target items. In an EFL context in which production is hardly the aim of the language classroom, caution is needed to avoid implementing too demanding production activities. Since this was a novel approach for the learners in the Saudi context, I decided to start with the least demanding and more controlled type of

output practice. It would be interesting to look at other types of less controlled pushed-output activities which allow for freer production and more elaborate negotiation. This will be the focus of Study 2. Another possible limitation of the study is that the two instructions differed on multiple aspects: modality (spoken vs. written), feedback (teacher vs. peers), or the number of engagement with words (two vs three times). The aim of the study is to compare the pushed-output instruction, as defined in Section 2.3.1, with the typical instruction that is used in the Saudi context, rather than isolating the effect of a particular variable. With this pedagogically valid design I was able to show that with the same amount of classroom time spent on the teaching of the same number of items, the spoken pushed-output instruction was overall better. This type of evidence is the one that might be able to change classroom practices in this context. However, the disadvantage of such ecologically valid design, is that I cannot make any claims about what exactly is causing the advantage of the spoken pushed-output instruction. Future studies isolating the effect of the different variables should be conducted. Another possible limitation could be related to the fact that in two-way activities, in which the output of one learner serves as the input for a pair, the role of incorrect input should be considered. It is likely that learners will provide input with errors, such as wrong pronunciation, inappropriate collocations, etc., which could affect the results. Future studies could account for this incorrect input in the design of activities, for instance by implementing a checklist that encourages peer feedback and assessing learners' production of the different components of word knowledge, such as pronunciation, collocational knowledge, etc.

All in all, Study 1.1 suggests that incorporating spoken pushed-output activities into classroom instruction could promote the learning of both SWs and FSs. However, this study only examined the effectiveness of spoken pushed-output instruction for low-intermediate EFL learners. It would be interesting to investigate how these findings might be extended to learners of other proficiency levels to gain a better picture of the

effectiveness of pushed-output activities across different levels of proficiency. Further, it suggests that there were no differences between SWs and FSs at either the recognition or recall levels when all received the same instructional approach, ensuring similar exposure and similar practice. Further, with the exception of sense-1 in receptive gains which had an advantage over the other meaning senses, no other differences among the three meaning senses emerged. However, vocabulary learning in this study was assessed by establishing whether learners had used the correct target items or not. This dichotomous scoring of the responses in the vocabulary tests did not take into account of how well those items were used and the characteristics of the sentences in which they were used. Looking at the lexical profile of utterances through measures such as mean length of utterances and lexical richness would allow us to examine not only whether the new target items were used but also how well they were used. This will be the focus of the following chapter.

Chapter 4

Study 1.2: Pushed-output instruction in comparison to Traditional instruction: differences in lexical profiles

In Chapter 3, I claimed that spoken pushed-output instruction leads to better learning gains than non-pushed output instruction for both SWVs and PVs in an EFL context. In addition, it was claimed that no differences existed between the learning of SWVs and PVs. These claims are based on analysis accounting only for the receptive and productive gains, which was also the main goal of the few available studies on pushed-output instruction. Much remains to be known, however, about the effect of spoken pushed-output instruction on the overall quality of EFL spoken production. L2 performance and development are complex in nature (Ellis & Barkhuizen, 2005); thus, a more in-depth analysis of learners' production in spoken pushed-output activities is needed. The value of such analysis lies in the contextualisation provided for the target items that goes beyond traditional vocabulary testing methods and thus the provision for adopting a broader view of vocabulary (Kremmel & Pellicer-Sánchez, in press). Such analysis is neglected in the previous studies of pushed-output instruction in lexical development, which also focused on examining whether or not target items were used. A more in-depth analysis of how learners used the target items in the productive tests would allow me to confirm the advantage of the pushed-output group observed in Study 1.1, as well as to further explore potential differences between SWVs and PVs. Thus, the main aim of Study 1.2 is to compare the lexical profiles of learners' productions in the three different treatment groups (control, traditional, and pushed-output) in the oral test. The dataset employed for the analysis in the present chapter was the one obtained for Study

1.1. Results of the oral test were analysed using a variety of measures including overall text length, mean length of utterances, lexical diversity, lexical density, and lexical sophistication.

4.1. Measuring lexical profiles

Lexical profiling is considered one indirect approach to assess vocabulary in which information about a person's lexical knowledge is gathered from their speech or writing (Kremmel & Pellicer-Sánchez, in press). Laufer and Nation (1995) listed several measures for lexical richness such as lexical density (LD), lexical sophistication (LS), lexical diversity (LDV) and proportion of errors. A variety of measures were used: lexical frequency profiles (LFP) (Laufer & Nation, 1995); Type-Token Ratio (Arnaud, 1984; Daller & Phelan, 2007; Daller & Xue, 2007; van Hout & Vermeer, 2007); D or VoC-D (Durán, Malvern, Richards, & Chipere, 2004); Guiraud's Index, Advanced Guiraud (Daller, van Hout, & Treffers-Daller, 2003); Coh-Metrix (S. A. S. A. Crossley, Salsbury, McNamara, & Jarvis, 2011); and P-Lex (Meara & Bell, 2001), just to name a few. Other measures, such as mean length of utterance (MLU), were used to measure language development in general (e.g., Brown, 1973; De Villiers & De Villiers, 1973; Hickey, 1991; Nice, 1925; Parker & Brorson, 2005) and the degree of relationship between MLU and some of the aforementioned measures was explored (e.g., Dethorne, Johnson, & Loeb, 2005; Miller, 1991). Among the different measures used in the literature, the following measures were used in the present study: LDV, LD, LS and MLU. Each of these measures is now reviewed in turn.

4.1.1. Lexical richness

Lexical richness is a multidimensional concept referring to the vocabulary used in a text (Read, 2000). Lexical richness allows capturing the breadth and depth of an individual's vocabulary use. It is considered an essential construct in L2 teaching as it is linked to the learner's ability to communicate effectively in both written and spoken manner. Engber (1995) and Linnarud (1986) have linked lexical richness of written compositions to the overall text quality, Laufer and Nation (1995) related lexical richness to vocabulary size, and Lemmouh (2008) linked it to grade levels of EFL learners' compositions. Lexical richness has been examined through a variety of measures in studies examining its development in spoken or written L2 texts, such as LDV, LD, LS, and proportion of errors (Read, 2000). The following section discusses in detail findings from previous studies examining different measures of lexical richness.

4.1.1.1. Lexical diversity (LDV)

LDV, also known as lexical variation, involves comparison between tokens and types to measure how often different words are used in a text (Laufer & Nation, 1995). Tokens are the total number of words in a text whereas types are considered the total number of different words (Nation, 2001). Accurate and reliable measures of LDV have been considered a challenge for vocabulary researchers due to the fact that text length is the main source for computing text's diversity. In other words, the longer the text, the less likely it is for a user to produce new words and more likely for it to be less diverse (Durán et al., 2004; Jarvis, 2002; P. M. McCarthy & Jarvis, 2007). There have been a number of measures for LDV used in the literature. One straightforward and intuitive measure of LDV is the number of different words (NDW) which was proved to be a good predictor in child language development (Klee, 1992; J. F. Miller, 1991). Another intuitive and basic measure of lexical diversity is TTR (Arnaud, 1984; Jarvis, 2002;

Laufer & Nation, 1995b; McCarthy & Jarvis, 2007; Tweedie & Baayen, 1998). TTR can be measured by dividing the total number of types by the total number of tokens. For instance, the TTR for this sentence *I am so worried I just need three minutes to collect myself* is 0.92. These two measures are considered the simplest ones to measure LDV; however, both are not without limitations. Both NDW and TTR are highly notorious for being sensitive to the length of texts (Laufer & Nation, 1995; Tweedie & Baayen, 1998). Several transformation techniques have been suggested to overcome this problem. The index of Guiraud (1954) is an example of such techniques, which is calculated by dividing the number of types by the square root of the number of tokens. By doing this, the index of Guiraud is considered independent from the text length. However, for long texts, this procedure might result in a higher LDV score than the simple TTR. Daller, Van Hout and Treffers-Daller (2003) stated that neither the simple TTR nor the index of Guiraud are considered valid measures for LDV at advanced stages of L2 acquisition. Biber (1991) suggested two approaches to overcome this limitation of sensitivity towards text length: either sampling only the first paragraphs or set of words in a text or assigning a limited production task where learners are asked to produce specific number of words (P. M. McCarthy & Jarvis, 2007); however, this may affect the authenticity of the learners' performance. Alternatively, reporting only the NDW, i.e., types, can be considered as a correction to TTR (Ratner & Silverman, 2000). The higher the total count of types, the greater its diversity; yet, the length of that text is still having an effect on its diversity as NDW relies only on the number of types (P. M. McCarthy & Jarvis, 2007).

Durán et al. (2004) proposed another alternative measure of LDV, D or VoC-D, addressing the limitation of sensitivity towards text length. This measure is calculated through a software named VoC-D. What this software does is calculating the statistic D through a formula that selects a sample of 35 to 50 tokens from a 50+ long text 100 times

to be fitted within a theoretical curve. Next, the TTR score is calculated for each sample; then, calculating the D scores out of the mean score of these TTR scores. Finally, these D-scores are averaged together. The whole procedure is repeated three times so that a final average, a more reliable D-score would be obtained (Durán et al., 2004). This process allows us to relatively determine LDV for quite short texts (Durán et al., 2004; Malvern & Richards, 2002). If a text has a higher D score, then lexical diversity is greater for that text (Durán et al., 2004). In a study conducted by McCarthy and Jarvis (2007), the capability of D to overcome the text length effect in the TTR measure was examined. They reported that the D measure had two limitations: first, text length still had an impact on the coefficient for texts of different lengths, and second, the procedure of the random sampling resulted in some different coefficients several times. D is considered a valid measurement tool for LDV; however, it is only reliable for a low diversity sample, for instance, the production of children or low proficiency language learners (McCarthy & Jarvis, 2007).

The Measure of Textual and Lexical Diversity (MTLD) was developed by Phil McCarthy and Scott Jarvis (2007) to reduce the effects of text length. It is the most recent approach to quantifying LDV, also known as Lexical Diversity Assessment Tool (LDAT) "the mean length of sequential word strings in a text that maintain a given TTR value" (0.720; McCarthy & Jarvis, 2010, p. 384). Once this TTR value drops below 0.72, another count (factor count) increases by 1, and TTR counts reset. The process continues from its stop point, and the same procedure is repeated for the whole text. Then, the total number of tokens is divided by the total factor count. The same procedure is performed again but in a reversed order. Then, the average of the two MTLD scores for forward and reversed order is the final MTLD score. The only limitation of this measure is text independency; that is, it does not evaluate the text as a whole unit. For example, a text with five paragraphs each having a high TTR score will be shown as having a great diversity even

if the last four paragraphs are exactly the same as the first one. Still, MTLD is considered the least affected by text length and most reliable measure of LDV (McCarthy & Jarvis, 2007, 2010).

LDV is used as a tool to measure the productive competences in written or spoken production, an "end-product" of language (Yu, 2009, p. 239). It has been considered an essential indicator of language proficiency of learners (e.g., Zareva, Schwanenflugel, & Nikolova, 2005) and an illuminative predictor of the quality of their production in writing (Laufer & Nation, 1995) and speaking (e.g., Jarvis, 2002; Malvern & Richards, 2002; O'Loughlin, 1995). This relationship between lexical diversity and language proficiency reflects the fact that producing a greater amount of different words in a written or spoken script reflects a higher level of language proficiency (Daller, Van Hout, & Treffers-Daller, 2003; Durán et al., 2004; Engber, 1995; Jarvis, 2002; Laufer & Nation, 1995; Yu, 2009). LDV can serve as a good indicator of vocabulary size of L2 learners as the ability to demonstrate variations in vocabulary use within context, either spoken or written, indicate underlying vocabulary growth (Durán et al., 2004). Berman and Nir-Sagiv (2007) argued that advanced vocabulary learning within the context of adolescent lexical development involves more coherent word knowledge repertories including synonyms, polysemy, and also having more words at their disposal to convey various meanings.

Measuring LDV also differs depending on the mode of the text, i.e., written or spoken texts. Biber (1988) found that LDV tends to be higher in written texts than in spoken texts, which can be due to the fact that writing affords learners opportunities to plan, revise, and edit their production, unlike the spoken production. Biber and Biber (1999) stated that high LDV scores in written texts are interpreted as reflecting more coherence and are considered a typical feature of writing. In contrary, Yu (2009) argued that LDV is not affected by the type of the task completed (written vs. spoken) or other

characteristics such as pre-planning and time pressure. Yu (2009) examined the relationship between LDV (using D), the overall quality of written or spoken discourses, and language proficiency in archived data of an international language test. Yu (2009) also examined the differences in LDV of spoken and written samples produced by the same participants. The results showed a positive correlation between LDV with the holistic quality of both written and spoken samples as well the participants language proficiency. Learners' LDV scores of written samples was affected by the topic in which the topics which are more personal, familiar to learners had significantly higher scores than impersonal topics. Further, the LDV of written and spoken samples of the same participants were approximately at the same level.

Several studies examined LDV in speaking and writing production (e.g., Arnaud, 1984; Daller, van Hout, et al., 2003; Engber, 1995; Jarvis, 2002; Linnarud, 1986; O'Loughlin, 1995; Vermeer, 2000; Yu, 2009). With regards to written production, Linnarud (1986) reported clear differences in LDV between compositions written by 17-year-old, Swedish learners of English and NSs. Swedish learners scored lower than English NSs in terms of LDV. Javris (2002) examined several transformation formulae of TTR curves of written narrative descriptions of silent films by 140 Finnish learners of English, 70 Swedish learners of English, and 66 NSs. The study explored the relationship between LDV and L1 background, L2 proficiency, writing quality, and vocabulary knowledge. The results indicated a clear association between LDV and L2 proficiency, but a more complicated association between LDV and L1 background, writing quality, and vocabulary knowledge. Learners with more L2 proficiency tended to have higher LDV scores. Further, NSs produced narratives of higher LDV than NNSs.

With regards to examining LDV in spoken discourse, Vermeer (2000) compared several measures such as NDW, TTR, Guiraud, Herdan and Uber indices of face-to-face interviews between himself and children aged 4–7 years learning Dutch as an L1 or L2.

Vermeer reported that the correlations between LDV measures and learners' scores in receptive and productive vocabulary tests did not show any concurrent validity with both receptive and productive tests, mostly at learners' whose vocabulary size greater than 3K. Simply, the relationship between LDV and learners' performance on the vocabulary tests was not clear. This could be related to the measures used in Vermeer's study due to their limitations, as discussed before. Daller et al. (2003) examined TTR and Guiraud, combined with LFP (Laufer & Nation, 1995), of spontaneous oral speech of German-Turkish bilinguals in Germany and Turkey. Daller et al. (2003) reported strong correlations between the LDV based on LFP and learners' language proficiency.

4.1.1.2. Lexical density (LD)

LD refers to the quantity of lexical words, as opposed to grammatical words, to the total number of tokens in a text (Ure, 1971). Lexical words are the primary carriers of the meaning of the sentence, such as verbs, nouns, adverbs, and adjectives, whereas functional words are those that express how words relate to each other, holding a more grammatical than lexical function, e.g., prepositions, auxiliary verbs, pronouns, particles, and articles. Laufer and Nation (1995) stated that, if a text contains more lexical words compared to the total number of words, that text can be considered "dense" (p. 309). LD can serve as an informative means to reflect on the degree of literacy versus orality of a text (Ure, 1971) as well as the information proportion of a text (Johansson, 2009). That is, texts which are more literate contain a higher number of lexical words and texts which are more oral consist of a higher number of grammatical words. Biber and Biber (1999) concluded that LD can also be reflected on the individual ability to use vocabulary resourcefully in text construction either written or spoken. Halliday (1989) and Ure (1971) reported that LD in spoken texts is below 40%, which is lower than that of written

texts (40%+), as the LD of spoken texts might be affected by different factors such as degree of instructiveness and planning (O'Loughlin, 1995; Ure, 1971).

Previous studies investigated the sources of variation in LD in the different written genres, including scientific and technical texts (Vande Kopple, 2003), translated vs. non-translated texts (Laviosa, 1998; Xiao & Yue, 2009), newspaper texts over periods of time (Štajner & Mitkov, 2011), in different registers in the same language (Yates, 1996), across languages (Neumann, 2013), and compositions of NS and NNS (Linnarud, 1986). More relevant for the purposes of the present study are those studies that examined LD in spoken discourse (e.g. Hasan, 1988; O'Loughlin, 1995; Stubbs, 1986; Zora & Johns-Lewis, 1989). Hasan (1988) examined LD of different spoken interactions in the EFL classroom: formal interviews, formal interaction, informal interaction, informal discussion, and informal conversation. He reported that the informal types of interactions were the highest in regard to LD compared to the formal interactions; more precisely, 43.69% for informal discussion, 41.60% for informal conversations and 40.64% for informal interaction compared to 33.67% for formal interviews and 38.96% formal interactions. Further, O'Loughlin (1995) studied LD of two types of output (live vs. tape-mediated) in three tasks (narration, description, and discussion) in two formats of oral proficiency tests: direct interview and semi-direct simulated interview. The results of O'Loughlin (1995) showed that the degree of LD in the tape-mediated version of candidate output is higher than that of the live ones. Also, the output for the narration task was lower than that of both description and discussion tasks, which were relatively similar in their LD scores. O'Loughlin explained the findings in terms of the degree of interaction as "the degree of interactiveness, rather than test format, emerges as perhaps the single most important determinant of candidate output" (p. 236).

4.1.1.3. Lexical sophistication (LS)

The construct of LS involves "the proportion of relatively unusual or advanced words in the learner's text" (Read, 2000, p. 203). It is the aspect of lexical knowledge that discriminates between the use of frequent and low-frequent vocabulary in the productive lexicon. The ability to use low frequency words can be considered as a sign for L2 lexical and speaking proficiency (Bardel & Gudmundson, 2018). The knowledge of sophisticated vocabulary has been linked to vocabulary size (Ian S P Nation & Webb, 2011) as well as to the knowledge of difficult words (Vermeer, 2000). However, there is still no exact definition of what it might be considered as advanced or sophisticated vocabulary (Kyle & Crossley, 2015). Laufer (1990), for instance, defined sophisticated or advanced vocabulary for Israeli learners with reference to the University Word List (UWL; Nation, 1990). Nation and Laufer (1995) suggested that advanced vocabulary depends on the researcher definition with emphasis on the learners' level.

LS has been measured in the literature following various approaches. In Hyltenstam (1988) and Linnarud (1986) studies, for instance, LS was calculated as the ratio of the number of sophisticated words to the total number of lexical words. Linnarud (1986) and Arnaud (1984) defined sophisticated words with reference to guidelines for the grading of vocabulary in the national Swedish schools (these guidelines are based mostly on frequency). Hyltenstam (1988) defined sophisticated lexical items as those beyond the 7K most frequent Swedish words. Linnarud (1986) found significant differences between NSs English and NNSs Swedish learners of English in the LS of written compositions whereas Hyltenstam (1988) found no differences between NSs and NNSs advanced Swedish written compositions. Kyle and Crossley (2015) emphasised the frequency of the lexical items to be considered in the calculation of LS. Harley and King (1989) suggested using a verb sophistication measure accounting for the ration of the number of

sophisticated verbs (i.e., infrequent verb beyond the 200 most frequent verbs in French) to the total count of verbs in a text. Their results showed significant differences between NSs and NNSs written composition in which NSs had higher LS scores.

The most traditional approach to account for LS is to use corpus-derived frequency counts (e.g., Laufer & Nation, 1995; Crossley et al., 2010; Crossley and McNamara, 2013). It can also be measured through other ways such as range (i.e., the number of the texts that a word occurs in), frequency of multiword items (n-gram), and the frequency of multiword items based on the academic language, and also some psycholinguistic measures such as imageability, familiarity, and concreteness (Kyle & Crossley, 2015). The most attested method so far is to examine the frequency counts of the words. There are two main approaches for analysing LS through frequency. The first approach involves categorising the words in a specific text based on the frequency bands lists, which is established on the basis of a specific corpus. The choice of the appropriate corpus depends on the mode of the text being analysed, written or spoken. McCarthy (1998) stated that frequency lists based on written sources differ from those based on spoken sources; something that would have consequences at the lexical level of the text. Thus, there is a clear risk in the validity of the analysed profiles of spoken texts if a written corpus was used to analyse. Several online tools can be used to provide such counts on the basis of written corpora such as LFP (Laufer & Nation, 1995), *VocabProfile* (Cobb, 2013), and *Text inspector* (2015). For analysing spoken texts on the basis of spoken corpora, i.e., spoken BNC and spoken COCA, *TAALES (V. 2.2)* (Kyle & Crossley, 2015) is a more suitable option. The second approach for measuring LS through frequency is to decide on a representative reference corpus (BNC, COCA, etc.), determine the frequency of each word in a text based on that corpus, and then divide the sum of all frequency scores by the number of tokens in that text to obtain an average frequency score for that text (Kyle & Crossley, 2015). Some tools can be used to calculate such counts as *TAALES (V. 2.2)*

(Kyle & Crossley, 2015).

Morris and Cobb (2004) used LFP to investigate its relationship with the academic success of some L2 teachers in a training course and examined written essays by 112 L2 teachers. They found strong significant correlations between LFP and two training scores in grammar courses. These results are in line with Bardakçi (2016) findings that examined 84 essays written by L2 learners. Further, Bardakçi (2016) reported strong negative correlations between the ratio of words in the General Service List (GSL) and two vocabulary tests: VLT (Nation, 1983) and Depth of Vocabulary Knowledge Test (Qian & Schedl, 2004). More recent studies examined the use LFP to discover the vocabulary needs for students by course designers and teachers (e.g., Aluthman, 2017; Catalán & Llach, 2017; Lin & Morrison, 2010; Mokhtar, 2017). However, LFP is not without limitations as argued by Meara and Bell (2001) such as the processing of errors and proper nouns as well as the accommodation of multiword units. Further, LFP does not allow for discriminating between students with similar sized vocabulary (Meara, 2005).

4.1.1.4. Proportion of errors

Proportion of errors refers to the number of errors in the usage of the words (Read, 2000). Any written or spoken production of EFL learners would probably contain lexical errors of various kinds. For instance, learners might choose the right word for an intended meaning but in the wrong form, or learners might choose an inappropriate word to deliver an intended meaning (Read, 2000). In this sense, proportion of errors is the converse of lexical richness and might help in measuring the quality of vocabulary in a written/spoken production. Previous studies listed some typical errors; for instance, Arnaud (1984) included minor/ major spelling mistakes, derivation mistakes, *faux-amis* (deceptive cognates), interference from another language, and confusion between two lexemes.

Engber (1995) classified lexical errors in several categories, shown in Figure 4.1.

I. Lexical choice

A. Individual lexical items

1. Incorrect – semantically unrelated
   a. It has some *meanings* to study in the US.
   b. We can help ourselves with *doing* as international students.
2. Incorrect – semantically close
   a. They have to *come back* to Indonesia. (*go back*)
   b. We can study some *strange* subjects. (*unusual*)

B. Combinations

1. Two lexical items
   a. Young people can *say* their *ideas*.
   b. I will *bring to go* the development of my country.
2. Phrases
   a. We have *a lot common*. (*a lot in common)*
   b. They can discuss their ideas *from the bottom of mind*.
3. Multiple errors involving core lexical items
   a. Who would want to have more exactly their major.
   b. It is being popular year and year.
   c. I will get English comprehension perfectly.

II. Lexical form

1. Derivational errors
   a. There are a lot of *confliction* between these two countries.
   b. It keeps the class more *activity*. (*active*)
2. Verb forms
   a. It isn't a good way for *looking* a job. (*looking for*)
   b. I want to *make business* with foreign companies. (*do business*)
3. Phonetically similar – semantically unrelated
   a. It is difficult to think in the *wild* horizon. (*wide*)
   b. I *thing* that English will help my country. (*think*)
4. Word distorted – major spelling error
   a. We have learned how to produce a *munifactual* machine. (*manufactured*)
   b. We can see the road shows *stimulously*. (*simultaneously*)

Figure 4.1 Classification of lexical errors (Engber, 1995, p. 146)

4.1.2. Mean Length of Utterances (MLU)

MLU was developed to measure the average length of children utterances at a specific point of time to capture the language development (Nice, 1925). Different approaches were proposed to calculate MLU. Nice's (1925) approach, for instance, involved the measurement of MLU in words. However, Brown (1973) and De Villiers and De Villiers (1973) pointed out that calculating MLU in morphemes was a stronger indicator of the grammatical development of the language children used than their chronological age. Yet, decisions on what constitutes a morpheme are problematic (Neuman, 2003; Thordardottir & Namazi, 2007). To eliminate this difficulty, Hickey (1991) and Parker and Brorson (2005) recommended going back to the initially proposed approach by Nice (1925) of counting MLU in words instead of morphemes. They argued that MLU in words is a faster, easier, more neutral, and more reliable procedure. The only potential limitation of MLU in words is that it underestimates the ability to measure the grammatical complexity of utterances. Further, Dethorne, Johnson and Loeb (2005) stated that counting MLU in words might be strongly influenced by the semantic content due to the lexical calculation base. It is worth noting that this lexical calculation base for MLU provide useful measures of the overall verbal productivity of learners.

MLU has been used extensively as an appropriate measure to capture the progress of grammatical complexity in early childhood language with or without language impairment (e.g., Dethorne, Johnson, & Loeb, 2005; Folse, 2006a; Hoff, 2006; Kaderavek & Sulzby, 2000; Klee et al., 2004; McGregor, Oleson, Bahnsen, & Duff, 2013; Owen & Leonard, 2002; Petersen, Gillam, Spencer, & Gillam, 2010; Piantadosi, Tily, & Gibson, 2011; Prevoo et al., 2014; Rice et al., 2010). The common conclusion of these studies was that high scores of MLU are good indicators of fluency while low MLU scores indicate a sort of syntactic breakdown. Besides, Bates and Goodman (1997) suggested

that there seems to be a strong association between MLU and lexical development in typically-developing children. However, as DeThorne et al. (2005) concluded, it remains unclear what does MLU reflect regarding grammar complexity or lexical development, and perhaps MLU should be considered as a global measure of expressive language. This view is supported by Eisenberg, Fersko, and Lundgren (2001) and Leonard and Finneran (2003).

In the EFL context, many studies examining the lexical richness of learners' written or spoken productions have included MLU as one of their measures to explore the relationship between MLU and measures of lexical richness (e.g., Dethorne et al., 2005; Miller, 1991; & Ukrainetz & Blomquist, 2002). Dethorne et al. (2005) explored the usefulness of MLU as a measure of expressive vocabulary and morphosyntax in a sample of 44 typically developing children (age: 28-37 months). They employed regression techniques to predict MLU from two measures: NDW and a tense accuracy composite (TAC), a measure of morphosyntax. Dethorne et al. (2005) reported a stronger association between MLU and NDW than the one found for TAC. Similarly, Miller (1991) and Ukrainetz and Blomquist (2002) reported the same strong correlation between MLU and NDW.

Very few studies have used MLU to evaluate learners' development of written or spoken production after an instructional intervention. García-Ponce, Mora-Pablo, Lengeling, and Crawford (2018), for instance, looked at the changes in MLU when learners were involved in different types of interaction tasks. Their results suggested that MLU increases within the interaction of the narrative and negotiation tasks more than that of the personal information tasks. Intuitively, one can assume that a greater knowledge of vocabulary after instruction might translate into longer sentences; however, there not enough research examining the effect of specific instructional approaches on improving vocabulary and hence increasing MLU scores.

As reviewed in this section, multiple measures have been suggested for measuring MLU and lexical richness components. All of these measures help in depicting a more detailed, rich picture of the vocabulary that learners use in their production. However, as it has been claimed in the Introduction to this chapter, these measures have been mainly used in the examination of overall quality of written or spoken production. There is a clear need to examine differences in these measures in relation to specific instructional approaches of vocabulary. This is precisely the aim of the present chapter. In order to address this aim, the recordings of the oral post-test of participants in Study 1.1 were analysed according to a range of the measures introduced above. This analysis is discussed in relation to the results of Study 1.1. By doing so, this allows me to have a clearer, more comprehensive picture of the effect of the different instructional approaches on the learning of target items as well as differences between the two target item's types and the different meaning senses.

4.2. Research questions

Study 1.2 aimed to answer the following research questions with regard to the different measures:

1. Is there a difference between spoken pushed-output instruction and traditional instruction in terms of length measures (overall text length and MLU) and lexical richness (i.e., LDV, LD, and LS)?
2. Is there a difference between SWVs and PVs in terms of length measures and lexical density within each treatment condition?
3. Is there a difference between the three meaning senses of the target items in the overall text length and MLU scores within each treatment condition?

No previous studies have actually looked at the effect of spoken pushed-output activities on the lexical profiles of learners (i.e., length measures and lexical richness) or the differences between SWVs and PVs in terms of length measures and lexical richness. However, based on the advantage that has been systematically shown for pushed-output groups in terms of quantity of vocabulary learned in previous studies and results presented in Chapter 3, it makes sense to hypothesise that this advantage will also be present in the examination of length measures and lexical richness. Further, based on the findings of previous studies about the minimal differences in the learning gains of vocabulary between SWVs and PVs if both received the same amount and type of instruction (e.g., Alali & Schmitt, 2012; Pellicer-Sánchez, 2020), along with the results of Study 1.1, it was expected that the same pattern will occur for some of the aforementioned measures. Moreover, with regards to the differences in overall text length and MLU scores by meaning sense, Garnier and Schmitt (2016) in their examination of the knowledge of polysemous items indicated an advantage of the first, most frequent meaning sense over the others, along with the findings of Study 1.1. However, since no previous studies have examined differences in overall text length and MLU between the multiple meaning senses of an item, it was hypothesised that the same advantage for the first, most frequent meaning sense would hold true for overall text length and MLU too.

4.3. Methodology

4.3.1. Data preparation

The data for the current analysis was taken from Study 1.1, which is the transcribed recordings of the productive tests. As explained in Section 3.3.5.3, participants were asked to respond to 144 test items: 83 DCTs and 61 factual questions. They spent 40

minutes completing the productive test. This test was recorded, and recordings were later transcribed in preparation for analysis. The total time for the recording was 160 minutes per student which equals 11.840 minutes of all students' recordings and 30,762 total number of words. The main aim of the study was to explore the features of the utterances in which the target items were correctly used, in order to obtain a better picture of how the newly learned target items were used in production. Thus, prior to any analysis, all the incorrect responses and responses that did not include the target items were deleted. For more details about the design of the study, please refer to Section 3.3.

It must be acknowledged that the data included in the present study is short, fragmented responses to an oral test with different prompts, which might be considered problematic for some of the measures, namely lexical diversity and lexical sophistication. It has been shown in the literature that both lexical diversity and lexical sophistication are highly susceptible to text length and an important recommendation in research has been that the text should be of at least 100 words for a more robust and reliable result (Jarvis, 2013). Hence, a 100-word threshold was applied for the inclusion of the data in the present study. When checking the spoken data, this threshold was only met for the combined test (SWVs and PVs together) in the traditional and spoken pushed-output treatments. Thus, only differences between treatments are examined with no distinction neither between the target types nor between the meaning senses. Only participants who produced +100-words utterances for the combined test (SWVs + PVs) are included. This part of the analysis only compares treatment conditions and not target type or meaning sense. By doing so, we can ensure that texts being considered for analyses are all above the 100-word level. This approach ensures that, while the data is still coming from fragmented speech, all texts included in the analyses are sufficiently long for the analyses reported. This means that, as stated in the research questions, when comparing the treatments, I

looked at overall text length, MLU and lexical richness (i.e., LDV, LD, and LS); when looking at differences between SWVs and PVs, I examined length measures (overall text length, MLU and lexical density; and when looking at the differences between the meaning senses, I examined overall text length and MLU.

4.3.2. Analysis tools

4.3.2.1. Kutools (2015)

Kutools is an add-in tool in Excel, provided by *ExtendOffice*, which combines advanced functions such as calculating the total number of morphemes, words, and utterances. *Kutools* was used in the present study to calculate the overall text length and MLU scores. MLU was calculated by counting the total number of words and the total number of utterances.

4.3.2.2. Text Inspector (2015)

Text Inspector (2015) is an online analysis tool which measures a range of linguistic features of L2 learners' texts. It can help to analyse the text in different ways, for instance, calculating the total number of tokens, types, verbal elements, noun elements, etc. in a text. It also helps in calculating LDV measures (TTR, VoC-D, and MTLD). *Text inspector* also has the Scorecard feature that gives the text a Lexical Profile Score as well as a guide to the text's Common European Framework of Reference (CEFR) level, A1- C2. The metrics in this Scorecard are different depending on whether it is writing, reading, or listening texts. Figure 4.2 provides an example of the Scorecard for a reading text.

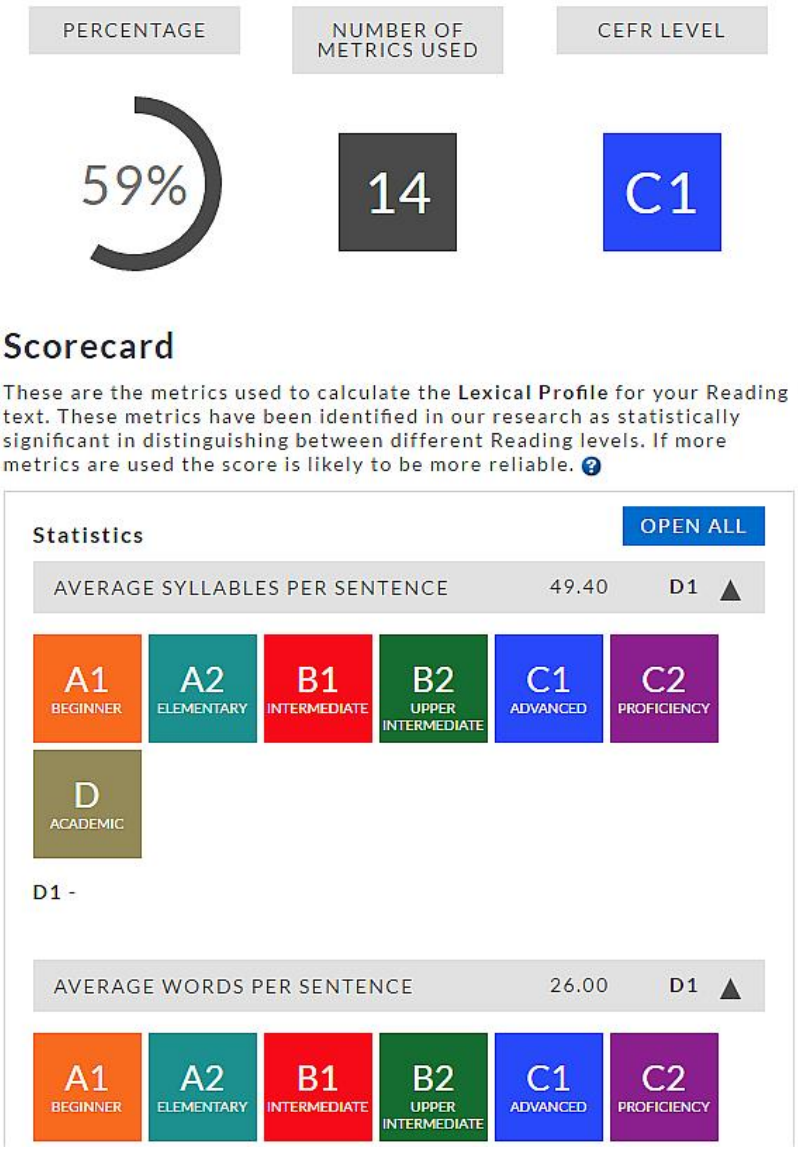


Figure 4.2 A sample of the Scorecard feature provided in *Text Inspector*

Text Inspector can also identify the metadiscoursal markers such *as firstly* and *in conclusion* in a text (Hyland 2004), as shown in Figure 4.3. Those markers either help in the organisation of a text or shows the writers' stance toward the content. *Text Inspector* has been used to analyse 900 EFL essays of test takers in a study conducted by Bax et al. (2013). Bax et al. (2013) also conducted a manual analysis of 200 essays to refine the

examples of the markers used in *Text Inspector*. The finding helped in modifying Hyland's (2004) list by adding more markers to make a 13 metadiscoursal markers list.

Category	Function	Examples
	Textual metadiscourse	
Logical connectives	Express semantic relation between main clauses	in addition / but / thus / and
Frame markers	Explicitly refer to discourse acts or text stages	finally / to repeat / here we try to
Endophoric markers	Refer to information in other parts of the text	noted above / see Fig / in section 2
Evidentials	Refer to source of information from other texts	according to X / (1990) / Z states
Code glosses	Help readers grasp meanings of ideational material	namely / e.g. / such as / i.e.
	Interpersonal metadiscourse	
Hedges	Withhold writer's full commitment to statements	might / perhaps / possible /about
Boosters	Emphasise force or writer's certainty in message	in fact / definitely / It is clear
Attitude markers	Express writer's attitude to propositional content	Unfortunately / I agree / X claims
Relational markers	Explicitly refer to or build relationship with reader	frankly / note that / you can see
Person markers	Explicit reference to author(s)	I / we / my / mine / our

Hyland, K. (2004) *Disciplinary Discourses: p111*

Figure 4.3 List of metadiscoursal markers (Hyland, 2004, p. 109)

The measures used in the present study obtained from *Text Inspector* (2015) scoring are TTR, VoC-D, MTLD, and the total number of nouns, verbs, adjectives, and adverbs.

4.3.2.3. TAALES (V. 2.2) software (Kyle & Crossley, 2015)

TAALES (V. 2.2) includes a wide range of LS indices (114 indices): lexical frequency, lexical range, and some psycholinguistic, word information such as concreteness, imageability, familiarity, and so on. These frequency and range indices draw on different corpora such as BNC (2007), Thorndike–Lorge Corpus (1944), Brown corpus (1967), Brown verbal frequencies (Brown, 1984), SUBTLEXus corpus of subtitles

(Brysbaert & New, 2009) and COCA. The frequency indices are ascertained by determining the reference corpus frequency of each word occurred in the target text and creating an average frequency score for a text by dividing the sum of all frequency scores for the tokens in the text by the number of tokens in that text (Kyle & Crossley, 2015). Additionally, the indices expand to the analysis to all words (AWs), content words (CWs), and function words (FWs). *TAALES* (V. 2.2) was used for calculating LS scores.

4.3.3. Scoring and analysis

4.3.3.1. Length measures

Overall text length. The overall length of the texts was scored using Kutools by counting the total number of words in all responses together by treatment mode, by target type, and by meaning sense. Afterwards, a mixed-design ANOVA with repeated measures was carried out with the treatment mode (control, traditional, and pushed-output) as the between-subjects variable, the target type (SWVs and PVs) and meaning sense (1, 2, and 3) as the within-subjects repeated measures variables (mixed-design ANOVA-1). The assumption of sphericity was checked, and it indicated that none of the effects violated this assumption. The results of the Levene's test indicated that the assumption of homogeneity of variance was significantly violated for all the meaning senses, ($p < .05$). Thus, the Games-Howell post hoc tests were used, as recommended by Field (2013) for being the most accurate post hoc test when equal variance is not assumed. Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of 0.025. Further, the effect sizes are reported according to Plonsky and Oswald (2014) guidelines: $r > .25$ small effect; $r > .4$ medium effect; $r > .6$ large effect.

MLU. As discussed in Section 4.1.2., although MLU is usually a more appropriate

measure for grammatical development, it has also been reported in previous studies to be related to some measures of lexical richness as LDV (e.g., Dethorne et al., 2005, Miller, 1991). Thus, it is used in the analysis of the current study to allow for better comparisons to previous studies. Further, despite different methods being suggested for measuring MLU, Hickey (1991) and Parker and Brorson (2005) claimed that using individual words that was initially proposed by Nice (1925) is faster, easier, more neutral, and more reliable procedure way of measuring MLU; hence, this is the unit used in the present analysis. In this study, a word is taken to be an orthographic word except for PVs which were counted as one word just to avoid bias in results by target type. MLU was calculated by dividing the total number of words by the total number of utterances.

Afterwards, a mixed-design ANOVA with repeated measures was carried out with the treatment mode (control, traditional, and pushed-output) as the between-subjects variable, the target type (SWVs and PVs) and meaning sense (1, 2, and 3) as the within-subjects repeated measures variable (mixed-design ANOVA-2). The assumption of sphericity was checked and it indicated that all effects violated this assumption; thus, the results of the Greenhouse-Geisser are used in reporting these effects, as recommended by Field (2013). The results of the Levene's test did not indicate that the assumption of homogeneity of variance was significantly violated for any meaning sense, ($p > .05$). Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of 0.025. Further, the effect sizes are reported according to Plonsky and Oswald (2014) guidelines: $r > .25$ small effect; $r > .4$ medium effect; $r > .6$ large effect.

4.3.3.2. Lexical richness

As discussed in Section 4.1.1, lexical richness can be accounted for through different measures such as LD, LDV, LS, and proportion of errors (Laufer & Nation,

1995). Lexical richness in the present study is conceptualised only through three measures: LD, LDV, and LS. Although I acknowledge the importance of accuracy (or measuring lexical errors) in task performance (Engber, 1995), lexical errors were not included in the analysis. I believe that identifying the lexical errors from non-lexical ones in the present data can be challenging. This is because identifying errors requires some subjective judgments and it is highly recommended to have several judges and check the inter-rater reliability (i.e., high level of agreement) (Read, 2000). Thus, the inclusion of proportion of errors was not feasible in the analysis of the current study due to practicality issues concerning allocating judges for such large amount of data.

LD. LD is the ratio of number of lexical words (tokens) to the total number of tokens. For the current study, lexical words are nouns, adjectives, adverbs, and verbs (excluding auxiliary verbs "*be*", "*do*", and "*have*", and modal verbs). The total number of these different parts of speech was calculated using *Text Inspector* (2015); then, Excel was used for scoring LD. A mixed-design ANOVA with repeated measures was conducted with the treatment mode (traditional vs. pushed-output) as the between-subjects variable, and the target type (SWVs vs. PVs) as the within-subjects repeated measures variable (mixed-design ANOVA-3). Both the assumption of sphericity and the Levene's test results did not indicate any significant violation, ($p > .05$). Statistical significance was accepted for the simple main effects at a Bonferroni-adjusted alpha level of .025. Further, the effect sizes are reported according to Plonsky and Oswald (2014) guidelines: $r > .25$ small effect; $r > .4$ medium effect; $r > .6$ large effect.

LDV. LDV can be calculated through different measures such as NDW, TTR, VoC-D, and MTLD. As reviewed in Section 4.1.1.1, NDW and TTR were considered problematic due to its sensitivity towards the length of a text whereas VoC-D and MTLD were the ones that are more frequently used. Thus, VoC-D and MTLD are the ones that

are selected for the study. As discussed in Section 4.3.1, only differences by treatment mode (traditional vs. pushed-output) were accounted for in LDV scoring, with no distinction between neither target types nor meaning senses. Prior to performing any test, the normality of the VoC-D and MLTD scores was checked using Kolmogorov-Smirnov tests, and the data turned out to be not normally distributed. Hence, two Mann-Whitney U tests were used to investigate differences in each measure (i.e., VoC-D, and MTLD) between treatment modes (traditional vs. pushed-output). The pattern of the results was the same across the two measures; thus, only the results of the MTLD scores are reported, since MTLD has been shown to be a more reliable measure to account for differences in text length (Fergadiotis, Wright, & Green, 2015; Treffers-Daller, 2013).

LS. As discussed in Section 4.1.1.3, LS can be measured through different approaches: corpus-derived frequency counts, range, n-gram, the frequency of multiword items based on the academic language, imageability, familiarity, and concreteness (Read, 2000; Kyle & Crossley, 2015). The original approach for calculating LS is through frequency and range indices. Thus, the analysis of LS in the present study was investigated using these two scores indices: frequency and range of for AWs, CWs, and FWs of the utterances. The reference corpus chosen for the present study is the BNC spoken corpus since all the meaning senses for the target items (specifically SWVs) were selected relying on the analysis of their frequency in the BNC spoken corpus (refer to Section 3.3.4.1 for more details of the selection of target items).

As discussed in Section 4.3.1, only differences by treatment mode (traditional vs. pushed-output) were accounted for in LS analysis, with no distinction between neither target types nor meaning senses. The dataset was prepared to be inputted in *TAALES* to examine LS by creating text files (.txt) for all the responses per participant. Initially, each of these indices (i.e., AWs, CWs, and FWs) were examined separately in the inferential statistics using multiple Mann-Whitney U tests. The pattern of the results was the same

for all these indices; thus, it seems more concise to report only the analysis of the frequency and range scores of AWs. The descriptive statistics for all the indices are included; i.e. frequency and range scores of AWs, CWs, and FWs.

4.4. Results

The presentation of the results will be divided according to the two measures: length measures (overall text length and MLU) and lexical richness (LD, LDV, and LS).

4.4.1. Length measures

4.4.1.1. Overall Text length

Learners' responses in the oral test varied in overall text length between the three treatment groups as well as between SWVs and PVs. Tables 4.1 and 4.2 present the descriptive statistics of the overall text length scores.

Table 4.1 *Descriptive statistics for the overall text length scores by treatment mode (control, traditional, and pushed-output) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
CG	36.66 (34.50)	26	6	161	27.28 (25.79)	19	6	124	9.38 (9.97)	7	0	37
TG	146.30 (65.48)	111.50	103	307	77.05 (34.98)	60.00	50	167	69.25 (31.07)	51.00	50	140
POG	469.90 (162.09)	448	202	733	240.51 (83.57)	243	92	375	229.39 (81.58)	200	99	358

Note: CG= control group, TG = traditional group, POG = pushed-output group.

Table 4.2 *Descriptive statistics for the overall text length scores of the three meaning senses (1, 2, and 3) by treatment mode (control, traditional, and pushed-output) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	SWVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
CG	13.90 (12.93)	9	2	59	8.14 (6.80)	6	1	29	5.24 (7.40)	4	0	36
TG	26.65 (15.16)	20	7	55	22.10 (16.51)	16	7	60	22.70 (13.10)	19	8	65
POG	86.58 (28.78)	82	25	142	77.21 (29.21)	79	27	125	76.73 (30.10)	80	26	142
	PVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
CG	3.62 (4.34)	2	0	13	2.00 (1.97)	2	0	6	3.76 (5.72)	0	0	20
TG	20.40 (12.62)	18	4	46	20.05 (13.10)	16	5	47	19.55 (13.85)	14	5	52
POG	79.21 (28.92)	76	34	129	75.48 (29.01)	70	22	130	74.70 (27.30)	74	23	115

Note: CG = control group, TG = traditional group, POG = pushed-output group.

The results of the mixed-design ANOVA-1 showed that the main effect of treatment mode was significant, F (2, 71) = 102.833, $p < .001$, $\eta_p^2 = 0.74$, a large effect size. Games-Howell post hoc tests showed that the mean difference in the overall text length scores

from control to pushed-output (72.21, 95% CI [60.31, 84.09], p = .000) was statistically significant as well as the difference from traditional to pushed-output, (56.41, 95% CI [43.09, 69.72], p = .000). The mean difference between control and traditional groups was also statistically significant, (15.79, 95% CI [8.02, 23.58], p = .000). This suggests that learners in the pushed-output group were able to produce overall longer texts than learners in the other two groups. Further, the main effect of target type was significant, F (1, 71) = 22.080, p < .001, η_p^2 = 0.23, a small effect size. The follow up contrasts showed that the utterances with SWVs were longer than these with PVs within each group, as shown in Figure 4.4. The interaction effect between treatment mode and target type was not statistically significant, F (2, 71) = 55.698, p = .565, η_p^2 = 0.016, a small effect size.

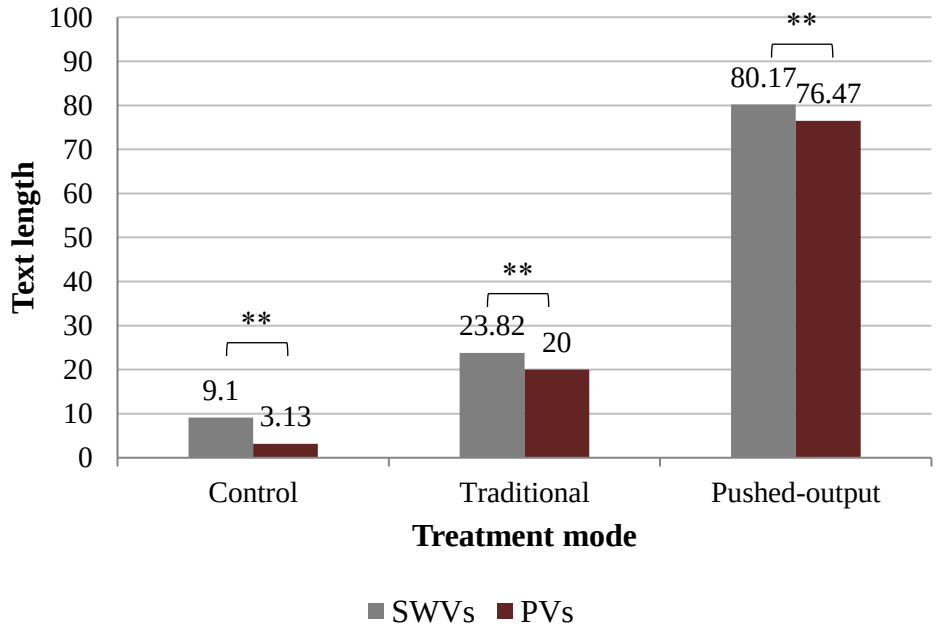


Figure 4.4 Means of overall text length scores for SWVs and PVs by treatment mode (control vs. traditional vs. pushed-output)

Further, the main effect of meaning sense on the overall text length scores was statistically significant, F (2, 142) = 12.183, p < .01, η_p^2 = 0.14, a small effect size. Contrasts revealed that scores for items with sense-1 were significantly higher than items with sense-3, F (1, 71) = 21.112, p < .01, η_p^2 = 0.23, a small size effect, and scores for items with sense-2 were no different from items with sense-3, F (1, 71) = .162, p = .69, η_p^2 = 0.002, a very small effect size. This suggests that when producing texts for the most frequent meaning sense, learners were able to produce longer texts than the ones produced for the other two meaning senses. There were no statistically significant interactions neither between treatment mode and meaning sense, F (4, 142) = 1.199, p = .310, η_p^2 = 0.03, a small effect size, nor between target type and meaning sense, F (2, 142) = 0.634, p = .098, η_p^2 = 0.01, a small effect size. The three-way interaction effect between treatment mode, target type, and meaning sense was not statistically significant, F (4, 142) = .526, p = .69, η_p^2 = 0.01, a small effect size.

4.4.1.2. MLU

Learners' responses in the oral test varied in length between the three treatment groups as well as between SWVs and PVs. Table 4.3 shows samples of these responses by treatment mode (control, traditional, and pushed-output), by target type (SWVs and PVs), and by meaning sense (1, 2, and 3).

Table 4.3 *A sample of learners' responses in the oral test by treatment mode (control, traditional, and pushed-output) for the three meaning senses of SWVs and PVs*

	Run (SWV)		
	CG	TG	POG
Sense-1: Move quickly	run quickly	run to the class	I will run to the room because of the exam and I was so late
Sense-2: Manage		run the company	my brother run the company he is the manager of the company
Sense-3: (liquids)		water will run	the water run in the bathroom
	Break down (PV)		
	CG	TG	POG
Sense-1: Stop working	break down	break down	sorry the bus broke down on way
Sense-2: Separate something		break down to small parts	break down the lesson to small parts
Sense-3: unable to control your feelings		she breaks down	she will break down and start crying

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Tables 4.4 and 4.5 present the descriptive statistics of MLU scores. The scores were higher in the pushed-output group but relatively similar in the other two groups. Also, it shows that MLU scores by target type were relatively similar, except for the control group in which scores for SWVs were higher than for PVs. The means by meaning sense were higher for sense-1 than the other two meaning senses.

Table 4.4 *Descriptive statistics for MLU scores by treatment mode (control, traditional, and pushed-output) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
CG	2.20 (1.30)	1.81	1.20	7.00	2.21 (1.32)	1.75	1.20	7.00	1.44 (1.16)	1.20	0.00	3.63
TG	2.12 (0.95)	1.81	1.16	4.00	2.18 (0.99)	1.93	1.13	4.13	2.05 (0.94)	1.72	1.00	3.85
POG	4.46 (1.10)	4.31	2.35	6.71	4.54 (1.19)	4.36	2.14	7.02	4.38 (1.19)	4.41	2.43	6.38

Note: CG = control group, TG = traditional group, POG = pushed-output group.

Table 4.5 *Descriptive statistics for MLU scores of the three meaning senses (1, 2, and 3) by treatment mode (control, traditional, and pushed-output) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	SWVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
CG	2.07 (1.10)	1.70	1.00	4.92	2.55 (2.53)	2.00	1.00	13.00	1.95 (0.91)	2.00	0.00	4.00
TG	2.44 (1.28)	1.95	1.00	5.00	2.08 (1.03)	1.86	1.00	4.78	1.99 (0.77)	1.96	1.00	3.75
POG	4.66 (1.24)	4.48	2.27	7.89	4.53 (1.23)	4.30	2.25	6.69	4.39 (1.23)	4.24	1.94	7.10
	PVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
CG	1.14 (1.34)	1.00	0.00	4.50	0.95 (1.07)	1.00	0.00	3.00	0.85 (1.39)	0.00	0.00	4.00
TG	2.27 (1.23)	2.00	1.00	5.00	2.03 (1.00)	1.85	1.00	4.14	1.98 (1.00)	1.50	1.00	3.75
POG	4.56 (1.10)	4.53	2.67	6.64	4.41 (1.09)	4.61	2.00	6.10	4.16 (1.06)	4.13	1.93	6.31

Note: CG = control group, TG = traditional group, POG = pushed-output group.

The results of the mixed-design ANOVA-2 showed that the main effect of treatment mode was significant, F (2, 71) = 63.305, $p < .001$, $\eta_p^2 = 0.65$, a large effect size. Games-Howell post hoc tests showed that the mean difference in MLU scores from control to pushed-output (2.87, 95% CI [2.23, 3.50], $p = .000$) was statistically significant as well

as the difference from traditional to pushed-output, (2.32, 95% CI [1.62, 3.02], p = .000). However, the mean difference between control and traditional groups was not statistically significant, (0.54, 95% CI [-.14, 1.23], p = .141). This suggests that learners in the pushed-output group were able to produce longer utterances than learners in the other two groups. Further, the main effect of target type was significant, F (2, 71) = 25.057, p < .001, η_p^2 = 0.26, a small effect size. The follow up contrasts showed that the utterances with SWVs were longer than these with PVs, a mean difference of 0.479 s, 95% CI [0.288,0.669), p <.000. The interaction effect between treatment mode and target type was also significant, F (2, 71) = 13.856, p < .001, η_p^2 = 0.28, a small effect size. In order to further explore this interaction, multiple follow-up Wilcoxon signed-rank tests by group were carried out. The results showed that the difference was statistically significant within the traditional group, z = -2.115, p < .05; however, in both control and pushed-output groups the difference was not statistically significant, as shown in Figure 4.5. This indicates that MLU of SWVs and PVs is affected by the type of instruction they received. This difference is more pronounced when teaching SWVs and PVs through traditional activities; however, the difference is minimal when using pushed-output activities, which suggests that learners in this group were able to produce utterances with more or less the same length for both SWVs and PVs.

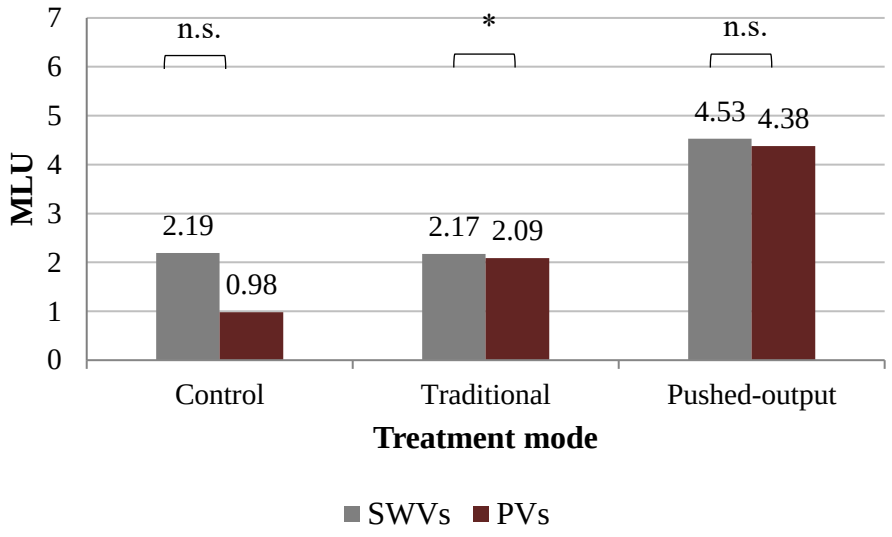


Figure 4.5 Means of MLU scores for SWVs and PVs by treatment mode (control vs. traditional vs. pushed-output)

Further, the main effect of polysemy on MLU scores was statistically significant, F (2, 142) = 5.425, $p < .01$, $\eta_p^2 = 0.07$, a very small effect size. Contrasts revealed that scores for items with sense-1 were significantly higher than items with sense-3, F (1, 71) = 17.087, $p < .01$, $\eta_p^2 = 0.19$, a small size effect, and scores for items with sense-2 were no different from items with sense-3, F (1, 71) = 3.553, $p = .06$, $\eta_p^2 = 0.05$, a very small effect size. This suggests that when producing utterances for the most frequent meaning sense, learners were able to produce longer utterances than the ones produced for the other two meaning senses. There were no statistically significant interactions neither between treatment mode and meaning sense, F (4, 142) = 0.868, $p = .47$, $\eta_p^2 = 0.02$, a small effect size, nor between target type and meaning sense, F (2, 142) = 0.634, $p = .05$, $\eta_p^2 = 0.01$, a small effect size. The three-way interaction effect between treatment mode, target type, and meaning sense was not statistically significant, F (4, 142) = 1.041, $p = .38$, $\eta_p^2 = 0.03$, a small effect size.

4.4.2. Lexical richness

4.4.2.1. Lexical density

The descriptive statistics for LD are presented in Table 4.6. The scores of the pushed-output group were higher than the traditional group. Also, responses which included SWVs were lexically denser than responses with PVs.

Table 4.6 *Descriptive statistics for LD (%) of the traditional and pushed-output treatments by target type (SWVs vs. PVs) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
TG	36.88 (4.25)	36.41	29.67	46.34	44.38 (6.45)	43.99	32.88	61.39	34.61 (2.60)	34.39	29.25	39.92
POG	55.43 (9.10)	55.17	40.44	75.00	66.55 (10.02)	66.67	49.38	94.44	42.96 (6.73)	42.86	30.34	57.14

Note: TG = traditional group, POG = pushed-output group.

The results of the mixed-design ANOVA-2 showed that there was a significant main effect of treatment mode, F (1, 51) = 85.964, $p < .01$, with a large effect size (η_p^2 = 0.63), in which the pushed-output group produced lexically denser utterances than the traditional group. The main effect of target type was also significant, F (1, 51) = 165.315, $p < .01$, η_p^2 = 0.76, a large effect size, showing that the utterances of SWVs were lexically denser than the utterances of PVs. The interaction effect of treatment mode and target type was statistically significant, F (2, 51) = 28.348, $p < .01$, η_p^2 = 0.36, a medium effect size. The follow-up Wilcoxon signed-rank tests by group revealed that the difference was statistically significant in both groups: the traditional group, z = -3.920, $p < .001$, and the

pushed-output group, $z = -5.012$, $p < .001$. Figure 4.6 shows that the difference is more pronounced in the pushed-output group with a difference of 20% when compared to the traditional treatment (10%). This suggests that in both groups' utterances with SWVs were lexically denser than those of PVs, especially within the pushed-output group.

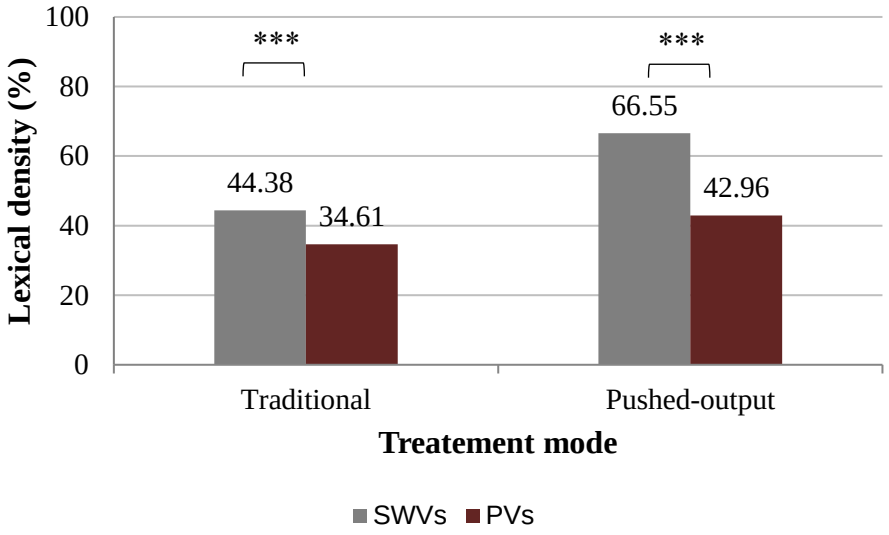


Figure 4.6 Means of LD scores for SWVs and PVs by treatment mode (traditional vs. pushed-output)

4.4.2.2. Lexical diversity

As explained in Section 4.3.1, given the fact that measures of LDV are particularly affected by text length, the analysis of this section only included texts that have a minimum of 100 words long. The application of this threshold did not involve the loss of any data as all participants in the traditional and spoken pushed-output treatments had a combined length above 100 words. The descriptive statistics for the three measures of LDV (TTR, VoC-D, and MTLD) are presented in Table 4.7. The means of TTR in utterances produced by participants in the traditional group were higher than those of

pushed-output group whereas, for the other two measures VoC-D and MTLD, the pushed-output group had higher scores than the traditional group.

Table 4.7 *Descriptive statistics for LDV measures (TTR, VoC-D, and MTLD) of the traditional and pushed-output treatments (SD presented in brackets)*

	Combined											
	TTR				VoC-D				MTLD			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
TG (N=20)	0.55 (0.08)	0.57	0.41	0.71	63.30 (24.52)	64.75	33.84	102.38	49.49 (23.54)	47.41	17.26	86.46
POG (N=33)	0.44 (0.05)	0.42	0.34	0.54	79.99 (15.88)	81.84	51.11	105.43	69.27 (14.17)	64.88	42.92	92.90

Note: TG = traditional group, POG = pushed-output group.

The results of the Mann-Whitney U test showed that the difference in MTLD scores between the traditional and pushed-output groups was statistically significant (U = 166.000, p = .003). Figure 4.7 shows that the pushed-output group produced more lexically-diverse utterances than the traditional group. This suggests that the pushed-output treatment was good enough to exhibit not only longer utterances but also more lexically-varied utterances than the traditional treatment.

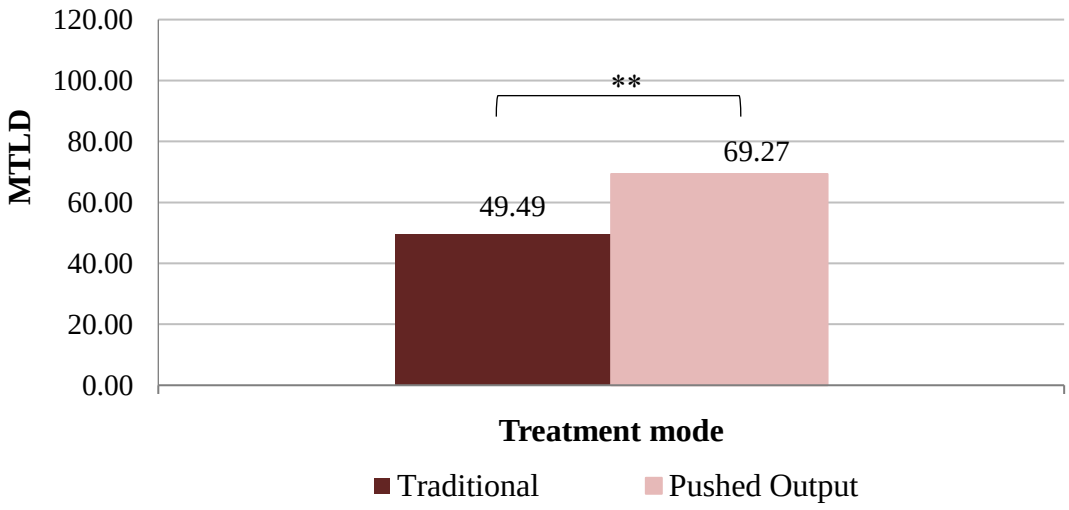


Figure 4.7 Means of MTLD scores by treatment mode (traditional vs. pushed-output)

4.4.2.3. Lexical sophistication

LS-Frequency. The descriptive statistics for LS-Frequency indices are presented in Table 4.8. The mean scores of the pushed-output group were higher than the traditional group.

Table 4.8 *The LS-Frequency scores of the utterances based on BNC spoken corpus for both traditional and pushed-output groups (AWs, CWs, and FWs) (SD presented in brackets)*

	AWs				CWs				FWs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
TG	4.27 (1.77)	3.94	1.76	7.98	0.70 (0.13)	0.66	0.59	1.12	13.79 (3.38)	14.40	6.43	19.45
POG	7.41 (1.33)	7.39	3.76	10.02	0.85 (0.15)	0.80	0.61	1.24	17.13 (2.26)	16.65	12.46	22.13

Note: TG = traditional group, POG = pushed-output group.

The results of the Mann-Whitney U test showed that the difference in LS-Frequency scores between the traditional and pushed-output groups was statistically significant, (U = 49.000, p = .000). Figure 4.8 shows that scores of LS-Frequency for the pushed-output group were statistically higher than the traditional group.

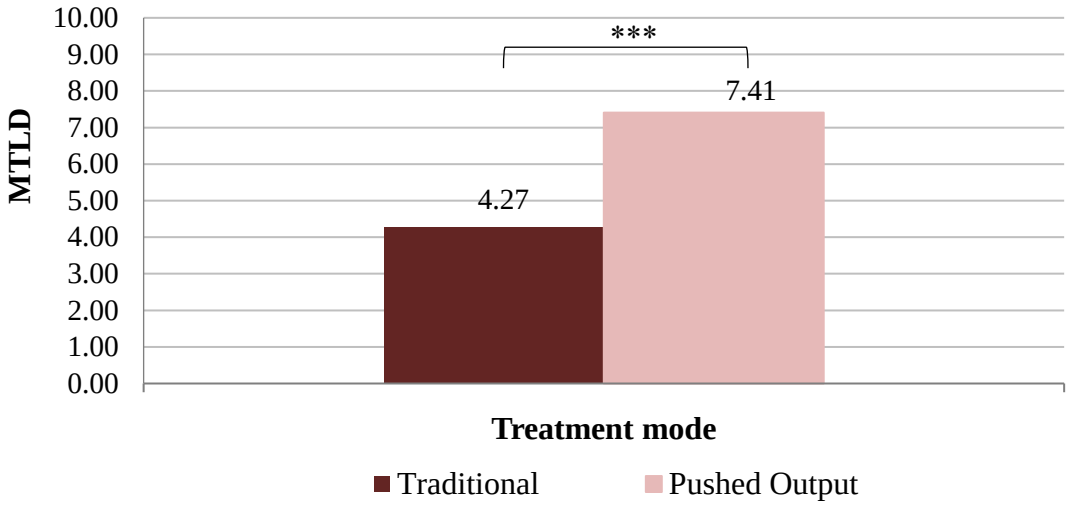


Figure 4.8 Means of MTLD scores by treatment mode (traditional vs. pushed-output)

LS-Range. The descriptive statistics for LS-Range indices are presented in Table 4.9. The range scores of the pushed-output group were higher than the traditional group.

Table 4.9 *The range scores of the utterances based on BNC spoken corpus for both traditional and pushed-output groups (AWs, CWs, and FWs) (SD presented in brackets)*

	AWs				CWs				FWs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
TG	63.94 (3.10)	62.57	59.57	69.31	53.31 (2.33)	52.80	49.28	58.08	89.93 (5.90)	92.03	76.34	98.52
POG	71.06 (3.19)	71.58	63.26	76.17	55.06 (2.57)	55.09	48.34	61.56	95.19 (1.18)	95.36	91.23	97.18

Note: TG = traditional group, POG = pushed-output group.

The results of the Mann-Whitney U test showed that the difference in LS-Range scores between the traditional and pushed output groups, (U = 39.000, p = .000), as shown in Figure 4.9.

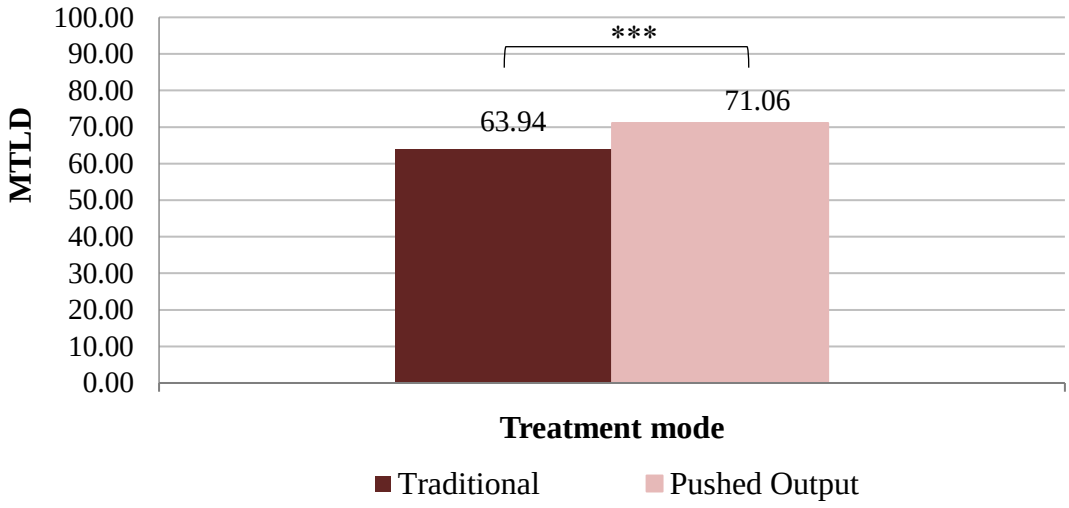


Figure 4.9 Means of MTLD scores by treatment mode (traditional vs. pushed-output)

4.4.3. Further analysis

As we have seen in Section 4.1.2, several studies found a strong correlation between MLU and TTR, as well as between lexical richness and vocabulary size of learners. In order to compare between these previous findings and the findings of the present study, several correlational analyses by target type were computed to examine the relationship between the different measures of lexical richness, MLU, and VLT. Because the data were non-normally distributed, Spearman's correlations were carried out. Table 4.10 shows that, with regards to measures of lexical richness, in LDV measures, VoC-D correlate highly with only MTLD; TTR does not correlate with either of the two.

Regarding the relationship between measures of LDV and other measures such LD, and LS, only TTR correlates negatively with these measures. LD correlates strongly with both measures of LS (Freq. and range). Further, MLU correlates strongly with LD, LS (both Freq. and range), and correlates negatively with TTR. Finally, VLT of the two levels, 2K and 3K, did not correlate with any of the measures of lexical richness. This lack of correlations between vocabulary knowledge and some measures of lexical richness has also been reported in Uchihara and Clenton's (2018) study. They did not find any correlation between receptive vocabulary size and lexical sophistications measures in spoken production of EFL learners. They used the whole VST with all levels so the fact that no correlations were found between the higher levels of VST and the learners' use of vocabulary in the activity is expected. They concluded that the reason for this lack of correlation might be related to the fact that in spoken discourse learners would not probably be using low frequency words beyond the most frequent 2,000 words (Milton, 2009), as the goal of conversational communication is to be comprehensible to listeners. Following their interpretation, it would make sense to expect a correlation between learner knowledge of high frequency vocabulary and the lexical richness measures. However, no correlations were found in the present study between the 2k and 3k scores of VLT and lexical richness measures. The small range in 2k and 3k scores and the lack of enough variation in these scores could explain the lack of correlation in the present study.

Table 4.10 *Correlations between lexical richness measures (LDV, LD, and LS), MLU, and VLT (2K and 3K)*

		LDV			LD	LS		VLT	
		TTR	VoC-D	MTLD		Freq.	Range	2K	3K
LDV	TTR								
	VoC-D	0.205							
	MTLD	-0.084	**.803****						
LD		**-.554****	-0.062	-0.057					
LS	Freq.	**-.590****	-0.275	0.028	**.611****				
	Range	**-.737****	-0.129	0.231	**.583****	**.870****			
VLT	2K	0.103	-0.014	-0.153	0.140	0.107	-0.045		
	3K	-0.139	-0.056	0.057	0.001	0.109	0.197	-0.151	
MLU		**-.646****	-0.128	0.268	**.576****	**.883****	**.964****	0.067	0.117

4.4.4. Summary of the results

Overall, the effect of treatment mode in favour of the pushed-output group compared to the traditional group was consistent within all measures. This is in line with the results of the receptive and productive gains reported in Chapter 3. The effect of target type was only investigated in overall text length, MLU, and lexical density and it was significant in all these measures showing that scores for SWVs were higher than PVs. The meaning sense effect was only investigated for the overall text length and MLU scores, and it showed that utterances with SWVs or PVs holding the most frequent meaning sense were longer than the other two less frequent meaning senses. This effect was not affected by either the target type or the treatment mode. Results of the ANOVAs

showed that none of the other interactions were significant. Table 4.11 presents a summary of the results for the different measures used in the study.

Table 4.11 *Summary of the results for LDV, LD, LS-Frequency, LS-Range, text length and MLU*

	Text length	**MLU**	**LD**	**LDV**	**LS**	
					Freq.	**Range**
Treatment mode	***	***	***	**	***	***
	p <.001	p <.001	p <.001	p <.01	p <.001	p <.001
Target type	***	***	***			
	p <.001	p <.001	p <.001			
Treatment mode × Target type	n.s.	***	***			
	p =.565	p <.001	p <.001			
Meaning sense	***	**				
	p <.001	p <.01				
Treatment mode × Meaning sense	n.s.	n.s.				
	p =.310	p =.470				
Target type × Meaning sense	n.s.	n.s.				
	p =.098	p =.498				
Treatment mode × Target type × Meaning sense	n.s.	n.s.				
	p =.690	p =.380				

4.5. Discussion

This study explored the effect of spoken pushed-output instruction on the lexical profiling of utterances including the target polysemous SWVs and PVs. The data confirmed the pattern of results found in the literature and in Study 1.1 about the superior

effect of pushed-output instruction in terms of vocabulary learning. In this section, the main results of the study are presented in response to each of the three main research questions and interpreted in light of previous research.

4.5.1. Lexical profiles of the pushed-output production vs. the traditional production

The findings from the present study suggest that learners in the spoken pushed-output instruction produced longer, more lexically-rich utterances than learners in the traditional treatment as indicated by the measures examined (i.e., text length, MLU and lexical richness: LDV, LD, and LS). Learners in the pushed-output instruction not only improved the vocabulary gains on receptive and productive levels, as reported in Study 1.1, but also used those learned words in longer and more complex utterances. Previous studies of spoken pushed-output instruction only reported the effectiveness of spoken pushed-output instruction in terms of the recognition and the usage of the target items in production, without examining the features of the utterances in which those target items were used (e.g., De la Fuente, 2002, 2006; Ellis & He, 1999). The present study shows that the effects of spoken pushed-output instruction go beyond learning the receptive or productive form-meaning links to improving the quality of usage of target items in context. Such approach of analysis could be employed in future vocabulary studies.

One possible explanation for the findings of the current study is that the mode in which the target items were learned seems to have some effect on the way these items were orally retrieved. That is, the activities in the spoken pushed-output instruction offered learners opportunities to orally practice the usage of the target items in context whereas activities in the traditional instruction did not. Learners in the traditional treatment only practiced the target items in reading manner to help them complete the

assigned matching and fill in blanks activities. Perhaps learners in the spoken pushed-output instruction became more efficient with oral practice and thus were more successful in the oral test due to the previous oral practice in the classroom. This could be explained in relation to the Transfer-Appropriate Processing framework (TAP), which was developed as an expression of such general relationship, and was specifically applied to memory (e.g. Bransford, Franks, Morris, & Stein, 1979; Morris, Bransford, & Franks, 1977). Advocates of the TAP framework affirm that information will be more easily retrieved if the way of retrieval of the information is similar to the way it was processed and stored (e.g., Blaxton, 1989; Morris et al., 1977). Thus, learners in the spoken pushed-output instruction had an advantage over the traditional instruction when they were tested orally in the same way they processed the information.

So far, the current findings support the TAP hypothesis by looking at the sum score of the four post-test sessions. It is also interesting to look at the scores per testing session. Table 4.12 presents the scores per session and it shows a steady increase in the pushed-output group in all the measures throughout the four post-tests sessions, whereas for the traditional group the increase seems to fluctuate throughout the four post-tests sessions. That proves that when learners were offered opportunities to orally practice the target items, their performance was gradually improving week by week when compared to the traditional group in which such gradual development is absent. That again supports the TAP hypothesis.

Table 4.12 *Progress development of the scores for lexical profiling measures throughout the four post-tests for the traditional and pushed-output groups (SD presented in brackets)*

	TG				POG			
	Post-test (1)	Post-test (2)	Post-test (3)	Post-test (4)	Post-test (1)	Post-test (2)	Post-test (3)	Post-test (4)
SWVs								
MLU	1.94 (1.11)	2.29 (1.50)	1.71 (1.11)	2.25 (1.14)	3.99 (1.39)	4.38 (1.62)	4.45 (1.44)	5.00 (1.77)
LDV	4.94 (1.12)	7.64 (2.00)	4.48 (0.24)	5.74 (1.02)	7.50 (1.27)	8.40 (2.00)	9.46 (2.05)	10.12 (3.14)
LD	43.90 (6.24)	44.07 (5.23)	42.57 (4.25)	47.06 (7.68)	52.42 (12.02)	61.52 (10.23)	66.67 (9.25)	68.97 (8.25)
LS (Freq.)	10.03 (5.02)	11.39 (6.23)	11.17 (1.02)	8.49 (4.32)	11.27 (2.54)	11.65 (1.65)	11.87 (0.23)	11.96 (1.45)
LS (Range)	74.68 (3.21)	73.54 (8.35)	73.28 (10.21)	72.50 (10.27)	72.67 (9.56)	73.11 (7.25)	74.36 (6.35)	80.03 (9.33)
PVs								
MLU	2.04 (1.14)	2.04 (1.55)	1.30 (1.17)	2.89 (1.61)	3.01 (1.61)	4.00 (1.41)	4.11 (1.11)	4.49 (1.43)
LDV	5.84 (2.00)	7.55 (1.05)	5.05 (2.45)	6.20 (2.75)	6.50 (1.12)	7.80 (2.19)	9.97 (1.25)	10.75 (1.09)
LD	33.65 (7.25)	34.77 (10.85)	35.86 (9.23)	38.37 (6.23)	46.67 (8.12)	48.00 (7.25)	50.00 (6.42)	53.35 (7.63)
LS (Freq.)	7.73 (4.33)	8.87 (0.25)	8.64 (2.02)	7.33 (3.45)	7.27 (2.02)	8.16 (1.23)	8.59 (1.75)	8.84 (0.32)
LS (Range)	85.15 (7.65)	81.73 (5.32)	85.97 (4.21)	85.71 (1.02)	80.7 (12.01)	84.46 (14.23)	87.39 (11.02)	88.36 (13.02)

Note: TG = traditional group, POG = pushed-output group.

Further, Laufer and Nation (1995) also pointed out that there are other factors that

affect the lexical richness of learners' performance in speaking and writing besides their vocabulary size, and among these is the skill of writing or speaking; that is, how well are the written or spoken production by the same learners at different times. As it can be seen in Table 4.11, learners in the pushed-output group were performing better than the other two groups whereas the traditional group were relatively similar (sometimes slightly higher) at different point of times (and different target items). The other factors they mentioned are familiarity with the topic and communicative purpose and these two are controlled to be the same since both groups completed the same oral test.

The findings of the current study can also be explained in terms of the Level of Processing framework proposed by Craik and Lockhart (1972) which posit that the nature and duration of memory trace are determined by the level of depth at which input was processed. Input that is subject to superficial analyses is assumed to be more poorly retained than input that receives a deeper level of analysis. In relation to the present study, the differences in the level of processing between the two instructional approaches might explain the results. It could be argued the spoken pushed-output instruction offered learners opportunities for a deeper level of processing of the target items through the oral practice. In contrast, the level of processing that the activities in the traditional group involved was more superficial.

Further, it seems logical to expect more lexically-rich utterances of the spoken pushed-output learners, when compared to learners in the traditional treatment, as they produced longer utterances. Previous studies reported that MLU correlated strongly with LDV measured in NDW (e.g., DeThorne, 2002, Dethorne et al., 2005, & Miller, 1991) and VoC-D (Durán et al., 2004), indicating that longer utterances entail more resourceful usage of vocabulary. Although it was not an aim of the study to explore the relationship between the different measures, the analysis also looked at the relationship between the different measures of lexical richness (i.e., LDV, LD, and LS) and MLU. The results of

the correlation analyses are in contrast with this finding as negative correlation was only found between TTR and MLU, but no correlations between VoC-D and MTLD with MLU. One explanation for these findings might be related to the type of data analysed in Durán et al. (2004) and the present study: the data in the present study come from transcriptions of an oral test taken by EFL learners, whereas in Durán et al. (2004) the data come from transcriptions of GCSE tests taken by British secondary school students. On the other hand, there were strong positive correlations between LD and LS with MLU, suggesting that with longer utterances, more lexically dense and sophisticated utterances were produced.

Moreover, Laufer and Nation (1995) suggested that vocabulary use is linked to the vocabulary size of the learners and that was extensively examined in the literature (e.g., Bardel & Gudmundson, 2018; Daller et al., 2003; Durán et al., 2004; Engber, 1995; Jarvis, 2002; Laufer & Nation, 1995; Zareva et al., 2005; Yu, 2009). Even though, again, exploring the relationship between VLT and the measures used in the present study (i.e., MLU, LDV, LD, and LS) was not the aim of the study, several correlation analyses revealed that no significant correlations were found between VLT and any of the measures. These findings could be related to the fact that learners in both groups, traditional group and pushed-output group, were matched in terms of their VLT scores (see Section 3.4.1). The lack of enough variation in learners' vocabulary level could have led to this lack of relationship between VLT and the rest of measures.

Overall, the results are in favour of the spoken pushed-output instruction for improving the lexical profiles of utterances produced by learners for either SWVs or PVs.

4.5.2. Lexical profiles within the pushed-output and traditional productions by target type (SWVs and PVs)

The results of the present study suggest that there were differences between SWVs and PVs in terms of text length, MLU and lexical density. Utterances with SWVs scored higher in all these measures. In Chapter 3, I reported that both SWVs and PVs were similarly improved in terms of the recognition and spoken recall of the form-meaning link. No significant differences were found between SWVs and PVs in any of the analyses reported in Chapter 3 (see Sections 3.4.3 and 3.4.4). However, the results of the measures used in the current analysis suggest that there are indeed some differences in how learners used SWVs and PVs and this potentially points to the additional difficulty of PVs over SWVs. This possible higher level of difficulty in learning PVs, when compared to SWVs, was not reflected in the gain scores presented in Chapter 3 but in the more in-depth analysis of the utterances in which the target items were used, as reported in this chapter. This also further justifies the need to conduct this type of analysis in vocabulary learning studies involving production.

As discussed in Chapter 3 (Section 3.5.2), Alali and Schmitt (2012) reported that similar learning gains at the recognition level occurred for FSs (idioms) and SWVs when they were taught in a similar manner, controlling for the amount and type of exposure; yet, for the recall level of mastery, these differences existed in terms of the learning gains. Peters (2014) and Kasahara (2010, 2011) reported that FSs imposed a further learning burden than SWs in terms of the written recall of form and meaning. The analysis conducted in the present study supports Peters (2014) and Kasahara (2010, 2011) conclusion that FSs might be considered challenging for L2 not in terms of the actual recall (production of the words itself) but in terms of the quality of these utterances produced for both SWVs and PVs. The usage of PVs seems to lag behind SWVs in terms of measures of lexical profiling in the present study.

The current findings could be related to the frequency of exposure of PVs compared to SWVs. Pellicer-Sánchez (2020) suggests that intrinsic properties of FSs that make learning them more challenging for L2 learners can be overridden if the quantity and quality of exposures are controlled to be similar. Precisely, if we account for the number of encounters, the approach of instruction, and the level of attention/noticing, then the learnability of both FSs and SWs is likely to be similar. While the findings of the analyses in Chapter 3 seem to be in line with this conclusion; yet, the few available studies that were reviewed in Pellicer-Sánchez (2020) paper were only concerned with the recognition and recall of target items. Further, though the exposure to both PVs and SWVs in the current study was controlled to be equal, it might be the case that learners had some external exposure to SWVs outside the treatments but not to PVs. This external exposure did not significantly affect the quantity of SWVs and PVs learned receptively or productively but it might have had some effect on the measures of lexical profiling of SWVs and PVs.

Another possible explanation for these findings could be that learners might have felt insecure when using PVs in context. They were able to recognise the meaning(s) of PVs, retrieve these PVs, but not to the extent to use them in a more complex manner in context. For instance, for the target items *turn up* and *collect*, the same student constructed these two sentences:

POG3: I am so worried I just need three minutes to ***collect*** myself

POG3: please ***turn up*** the TV

It could be the case that learner POG3 felt more confident to construct a longer sentence for the SWV *collect* adding a reflexive pronoun and adverbs whereas for the PV *turn up* the sentence consists of only an exclamation, a PV, and an object. Another learner in the traditional group had produced these utterances for the same target items:

TG2: five minutes to ***collect*** ideas

TG2: ***turn up*** TV

Though both sentences are considered simple; yet, we can see that learner TG2 for the PV *turn up* only produced the PV along with an object while for the SWV *collect* learner TG2 added words other than the object, i.e., *five minutes*. Several studies reported the tendency of learners to avoid using PVs in context and how they alter to use their counterpart SWVs, especially in the spoken context (e.g., Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004; Siyanova & Schmitt, 2007). Such tendency was not observed in the relative gains in Study 1.1; yet, this tendency could be conceptualised in the way they used the PVs in context in the current study. The difficulty that PVs impose might not only affect learning the form or the multiplicity of meanings but also the structuring of utterances of such items in context, specifically spoken context.

To conclude, the findings suggests that SWs still holds some advantage over FSs in terms of the lexical features of the spoken production and suggest the need to look at these measures in future vocabulary learning studies.

4.5.3. Lexical profiles within the pushed-output and the traditional productions by meaning sense (1, 2, and 3)

The third research question in the current study is concerned with the differences in text length and MLU scores of the three meaning senses of the target items. The results reported in Study 1.1 revealed that the first meaning sense was better recognised than the other two meaning senses among both SWVs or PVs. Results also showed that recall of the three meaning senses was the same among SWVs or PVs. Thus, it was expected that a similar advantage of the first meaning sense may occur for text length and MLU to some extent. The results of the present study showed that sense-1 was used in longer contexts

than the other two meaning senses. However, there were no interaction effects neither between meaning sense and treatment mode nor between meaning sense and target type. Learners in both treatment groups produced longer utterances when the meaning of the target items (SWVs or PVs) was the most frequent meaning sense. For instance, for *break down*, here is what two learners from the two treatment groups produced:

POG2: sorry the bus ***broke down*** on the way (eight words)

POG2: ***broke*** it ***down*** to parts (five words)

POG2: my sister ***broke down*** (four words)

Learner POG2 produced longer sentence for the first meaning sense of the PVs *break down* whereas for the other two meaning senses, utterances were relatively within the same length. Another learner in the traditional group produced these utterances for the same target item:

TG3: the bus ***broke down*** on way (six words)

TG3: ***break down*** to sections (four words)

TG3: she ***break down*** (three words)

Learner TG3 produced a slightly longer sentence for the first meaning sense than the other two meaning senses in which the length were relatively similar. These findings can be explained in terms of the frequency effects that I discussed in Section 3.5.3; that is, the most frequent meaning senses are better learned and used in a longer context. Perhaps the most frequent meaning sense was also the one that learners were exposed to before the present study. This potential, previous exposure was not enough to show declarative knowledge in the pre-test but was perhaps enough to boost the learners' knowledge of the usage of these target items with sense-1 which was shown by the MLU scores. However, this view is rather limited as it is only based on the MLU scores only. Future research should examine differences between the multiple meaning senses using

other measures such as LDV, LD, LS-Frequency, and LS-Range. Overall, it seems that items with the most frequent meaning senses can be used in longer utterances.

4.6. Limitations and conclusions

All the limitations discussed in Chapter 3 (Section 3.6) for Study 1.1 (i.e., the absence of an immediate post-test, the use of only PVs to represent FSs, the absence of an oral pre-test, and the use of visual stimulus) are still applicable to the present study since it involves the same dataset but employing different approach of analysis. Yet, there are inevitably some limitations of these analyses that should be considered, and that could be the foundations for future research. Though the in-depth analysis used in the present chapter, examining lexical richness and MLU, allowed to reveal further differences between treatment modes and target types that cannot be captured by only looking at the receptive and productive learning gains; yet, it is not without limitations.

First of all, as discussed in Section 4.3.3.1, inputting the spoken data as a whole text is considered a major limitation of the present study. The dataset consists of responses to different questions of an oral test, and they were inputted and scored as a whole text per participant. This scoring approach did not account for the textual homogeneity, which is usually found in a spontaneous speech. Thus, it would be fruitful for future studies to investigate more textual spontaneous datasets of L2 oral production when taught using spoken pushed-output activities.

Moreover, the present study is based only on the spoken dataset, and therefore it prevents me from knowing whether the findings could be generalised to the learners' written vocabulary. Although these findings are probably the most comprehensive view to date of the effect of spoken pushed-output instruction on learning polysemous SWVs and PVs in terms of learning gains, MLU, and lexical richness; yet, future research should

investigate the impact of written pushed-output instruction on learning polysemous SWVs and PVs.

Another limitation is the fact that the quantity of words of data did not allow to examine differences between the three meaning senses with all the different measures. The amount of data extracted from participants in the control group was not enough to allow for scoring the specified measures in the present study; thus, this group was eliminated in the analysis. Further, the data for the three meaning senses were also less than the minimum in the traditional group, which did not allow me to score and then run any comparison tests between the utterances of the three meaning senses produced by the traditional and pushed-output groups in any of the measures for lexical richness. Future studies should account for the requirements of scoring the different measures of the spoken data extracted from L2 learners; for instance, ensuring that the text length allows for scoring the different measures.

Further, the data were scored for lexical richness using different tools, i.e., *Kutools* (2015), *Text inspector* (2015), *TAALES (V. 2.2)* (Kyle & Crossley, 2015). Each tool has its own limitation; for instance, *Text inspector*, was built to score data extracted from reading, writing, and listening but not speaking. However, I believe that the scores that were taken from the analysis carried out on *Text inspector*, i.e., TTR, VoC-D, MTLD, and the total number of nouns, verbs, adjectives, and adverbs, would not have been different if they were scored in the mode of speaking. That is because, as stated by the developers of the website, specifying the mode of text in *Text inspector* is specifically useful in the Scorecard for lexical profiles of the texts, which were not included in this study (refer to Section 4.3.2.2 for more details about the Scorecard). Perhaps it would be fruitful if *Text inspector* would allow for scoring data extracted from a spoken mode. *TAALES (V. 2.2)*, on the other hand, was used for calculating LS-Frequency and LS-

Range scores of the utterances and it is not without its own limitations. *TAALES (V. 2.2)* is a fairly redundant tool which uses a limited size of the corpora database (i.e., Brown, 1984, frequency values); thus, perhaps not all the words in a target text were included in calculating LS-Frequency and LS-Range scores, which may have affected the scores.

All in all, the present study demonstrated that the advantage of spoken pushed-output instruction that had previously been found in the learning of the form-meaning link was also reflected in the lexical profiles of the utterances in which those learned items were used. The results also displayed differences between the target types (SWVs and PVs) in terms of the overall text length, MLU and lexical richness, with SWVs having an advantage over PVs in almost all the measures. It also revealed differences in the overall text length and MLU produced for each meaning sense, with the most frequent meaning senses entailing longer utterances. The present study demonstrated another advantage of spoken pushed-output instruction, further justifying its use in the classroom. However, given the many different types of spoken pushed-output activities that could be implemented in the classroom, a logical question would be which the most effective types of spoken pushed-output activities are. Having a clear answer to this question has important pedagogical implications and is needed if we are to recommend and implement the use of spoken pushed-output activities in EFL teaching. This will then be the focus of the following experimental study presented in Chapter 5 and Chapter 6.

Chapter 5

Study 2.1: Exploring differences between three spoken pushed-output activities: differences in learning gains

Study 1.1 explored the effectiveness of two treatment conditions, spoken pushed-output (PO) instruction and traditional vocabulary instruction, on the learning of multiple meaning senses of high-frequency SWVs and PVs. The study explored receptive and productive learning gains, and the results showed that spoken PO activities led to higher learning gains on both receptive and productive levels. If we are to implement spoken PO activities in the classroom context, given the different types of PO activities that could be used, it is important to know which type of PO activity is more effective. Thus, an interesting question is whether different types of PO activities lead to similar learning gains (receptive or productive). This chapter reports the results of Study 2.1, which investigated differences between three spoken PO activities for learning polysemous SWVs and PVs. Over five weeks, EFL learners (N=49) were taught the multiple meaning senses of high frequency SWVs and PVs using three activities that differed in their involvement load, as defined by the Involvement Load Hypothesis (ILH), and Technique Feature Analysis (TFA) aspects: i.e. sentence reconstruction, listen-and-retell meaning and picture-description activities. Knowledge of the multiple meaning senses was measured through receptive MC pre- and post-tests and a productive oral post-test. The data obtained were analysed using two approaches. The first approach involves examining the receptive and productive vocabulary gains after instruction, which is similar to Study 1.1 reported in Chapter 3 and is the focus of the present chapter (Study

2.1). The second approach involves examining the lexical profile of the spoken production after instruction, which is similar to Study 1.2 reported in Chapter 4 and is presented in Chapter 6 (Study 2.2).

5.1 Background of the study

Chapter 3 reported the results of a study that provided empirical evidence for the effectiveness of spoken PO instruction over traditional vocabulary instruction for learning the multiple meaning senses of SWVs and PVs. However, there remains a clear need to examine the effectiveness of the available variety of spoken PO activities to determine which ones are more beneficial for vocabulary learning. This section reviews the literature on PO instruction, with a focus on the type of activities used and on the findings of previous studies. The following section provides a brief overview of the different PO activities, first those in the written mode, then in the spoken mode.

5.1.1. Written PO activities

Much previous research has provided evidence that vocabulary can be learned from a range of written PO activities. One PO activity commonly examined in the literature is **composition writing**, in which learners are required to write a composition that coherently connects a given set of words. **Sentence writing** is another common activity among PO studies, which simply asks learners to write original sentences using the given items. These activities have been found to be the most effective for receptive and productive learning of vocabulary (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017). One common explanation offered by all these studies for this effectiveness is the involvement load (IL) that each activity induced. Composition writing and sentence writing induce identical IL indices (IL=3, +need, –

search, ++evaluation), higher than those of other activities, such as cloze and fill-in-the-blank exercises (IL=2, +need, –search, +evaluation). Thus, they resulted in greater vocabulary gains at both the receptive and the productive levels. Further, Zou (2017), in her investigation of these two activities, argued that sentence- and composition-writing are highly beneficial for vocabulary learning for two reasons. First, the degree of pre-planning involved in the activities encourages learners to create possible scenarios in their virtual mental space before writing down their answers; thus, they have the opportunity to practice using the target items twice (once while thinking and once while writing their answers). Second, writing exercises require the chunking of words; that is, information is processed in meaningfully grouped units rather than in individual units, which makes it easier to memorise. The use of chunking may contribute to the successful recall of the meanings of these words. Further, Zou (2017) highlighted that, among the two, composition writing was more effective than sentence writing, as it requires a higher cognitive load as learners might employ more coherent, hierarchical connections between the various chunks. In sentence-writing, on the other hand, no such load is required, as each sentence is independent from the other. Hence, Zou (2016) suggested the inclusion in writing activities of this additional level of evaluation (+++ very strong) that distinguishes composition-writing from sentence-writing.

Other PO activities that have received much attention in the PO literature include **cloze** and **fill-in-the-blank activities**, which involve filling in blanks in a passage with target words, either at a sentence level in fill-in-the-blank activities or at a reading passage level in cloze activities. The two have been found to be less effective, resulting in lower vocabulary gains at both the receptive and productive levels, than composition and sentence writing activities in earlier studies (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017) due to inducing lower ILs compared to

composition writing and sentence writing (IL=2, +need, –search, +evaluation). Further, as Zou (2017) demonstrated, cloze activities involves neither the pre-planning nor the chunking and hierarchical organisation that contribute to the high effectiveness of sentence- and composition writing.

Another activity commonly used in PO studies is the **text reconstruction** activity. It involves learning ordering words, sentences, or paragraphs to create a coherent text. The text reconstruction activity can be adapted to different formats, such as the cloze text reconstruction activity, in which there are missing words from the text and learners must fill in the missing parts with some given words. Another format is the editing text reconstruction, in which learners are given a text containing some errors and are tasked with identifying the errors and correcting them. Both formats were investigated in Nassaji and Tian's (2010) study on the learning of PVs. Their research is unique in the way it examined two different conditions for PO activities: individual work versus collaborative work. Most of the studies mentioned above examined the effect of written PO activities only in individual work (i.e., learners complete the activities by themselves). Nassaji and Tian's (2010) study is also the only study examining the acquisition of English PVs through PO activities. Their results suggested that both individual and collaborative conditions resulted in similar vocabulary learning gains. In terms of differences between PO activities, the editing activity led to significantly higher gains of PVs than the cloze activity. They explained the results in terms of the degree of negotiation and scaffolding generated by each activity. Their study also involved analysis of transcriptions of interaction among learners during the collaboration condition, which revealed that the editing activity generated more instances of form-focused talk and feedback than the cloze activity.

Another type of text reconstruction activity is **dictogloss**, which has been extensively studied in the PO literature. Dictogloss is a dictation technique in which

learners hear the text read out multiple times, take notes and reconstruct that text from memory. It has been investigated in PO studies both in relation to language learning in general (Swain & Lapkin, 2001) and for grammar acquisition in particular (e.g., Kowal & Swain 1994; Malmqvist, 2005; Nabei, 1996; Swain, 1998). The findings suggested that dictogloss has the potential to promote focus on form (e.g., Kowal & Swain, 1994, 1997; Malmqvist, 2005; Nabei, 1996; Swain, 1998; Swain & Lapkin, 2001) and to expose learners to various amounts of meaningful input, output and feedback during the different stages of the activity (Kowal & Swain, 1994, 1997; Malmqvist, 2005; Nabei, 1996; Swain, 1998; Swain & Lapkin, 2001). Specifically, Nabei (1996) reported that learners completing the dictogloss were engaged in both meaning-based and grammar-based communication and provided with opportunities to hypothesise about how grammar and vocabulary work.

Text reconstruction activities have been shown to lead to more noticing, be it noticing the gap in learners' interlanguage (Swain, 1998) or noticing the discrepancy between correct target language use and their own language use (Doughty & Williams, 1998). Specifically, dictogloss has been shown to trigger metalinguistic talk on the form-meaning link while producing language (Kowal & Swain, 1997). In other words, while completing the activity, learners first notice the gap of another form for the target vocabulary, and that noticing may trigger them to talk about which form best fits the intended meaning. They may also compare their own production to the given sentences. The nature of the activity is the production of a new text, which contributes to development at not only the lexical level but also the discoursal, syntactic and phonological levels (Thornbury, 2002).

Text reconstruction has been investigated not only in studies examining vocabulary but also in studies examining grammar acquisition through PO activities. However, the

results of these studies were inconclusive about its effectiveness. Izumi et al. (1999) and Izumi and Bigelow (2000), for instance, reported insignificant differences between a text reconstruction activity and a text comprehension activity. On the other hand, Izumi (2002) and Nobuyoshi & Ellis (1993) reported superior performance of learners who engaged text reconstruction activities over those in text comprehension activities.

Jigsaw, another type of PO activity, refers to a split information activity which learners rely on one another to help them complete (Nation & Newton, 2008). Jigsaw has been shown not only to improve production accuracy but also increase instances of negotiation between learners (e.g., Pica, Holliday, Lewis & Morgenthaler 1989: Pica, Lincoln-Porter, Paninos & Linnell 1995). Swain and Lapkin (2001) compared dictogloss and jigsaw for language learning by evaluating their effectiveness in contributing to a range of skills: content, organisation, vocabulary, morphology and syntax. The learners in the jigsaw activity were asked to construct a story based on a series of eight pictures in a two-way information gap activity. Each held four pictures and worked through the cards sequentially, alternately telling each other what their pictures contained. In the dictogloss activity, learners listened to a passage read twice at normal speed, took notes on its content and worked in pairs to reconstruct and rewrite the passage based on the two sets of notes. The study primarily focused on syntactic accuracy but also included a brief analysis of lexical learning. The learners were assessed using three measures: a MC test on article use (based on grammatical gender), a grammaticality judgement test on a given set of sentences and a picture-word matching test. They reported that the dictogloss was more effective than the jigsaw in improving accuracy of both grammatical form and target vocabulary.

Sentence-combining is a type of PO activity which requires learners to combine segments into a grammatically correct sentence. **Translation** is also another type of PO activity, which requires learners to translate a given sentence from the L1 to the L2 or

vice versa. These were found to be less effective than sentence or composition writing but more effective than meaning-matching activities for vocabulary learning (Tahmasbi & Farvardin, 2017). The two activities, sentence-combining and translation, both induced a lower IL index (IL=2, +need, –search, +evaluation) than composition- and sentence writing (Tahmasbi & Farvardin, 2017).

In summary, all of the aforementioned studies indicate general advantages of composition writing and sentence writing activities for vocabulary learning at the receptive and productive levels. The studies explained this advantage in relation to the ILH framework: both activities induce higher ILs indices than other PO activities. The higher the IL induced, the higher the vocabulary gains, as contended by Hulstijn and Laufer (2001). However, does this conclusion hold true for spoken PO activities?

5.1.2. Spoken PO activities

Very few studies have examined the use of spoken PO activities for receptive and productive vocabulary learning (e.g., Ellis & He, 1999; De la Fuente, 2002). These two studies had the common goal of investigating the effectiveness of input-based versus output-based instruction. Both studies examined one activity, a **picture description** activity which involves learners making statements about pictures: descriptions, comparisons, predictions, pointing out differences, explanations for what happened and so on. This activity can be carried out in various formats: picture-ordering, same-or-different, listening to pictures, listen-and-do, etc. Pica, Lincoln-Porter, Paninos and Linnell (1996) stated that picture description activities help to encourage the learner's output, as they are characterised by a one-way flow of information in which one learner holds all the information and supplies it to the other. As a result, such activities may place a heavy demand on learners to produce accurate and clear information about the picture,

pushing them to externalise their hypotheses about how the target language works (Pica et al., 1996).

Both Ellis and He (1999) and De la Fuente (2002) examined a picture description activity in a listen-and-do format under three different conditions: an input-only condition (learners listen to directions about where to place the furniture in the apartment), an interactionally-modified input condition (listen and negotiate with the teacher) and an output condition (learners produce the directions to their peers). The findings showed that the output condition outperformed the other input conditions, and the authors related these results to the beneficial role of dialogic interaction (Ellis & He, 1990) and the importance of the negotiation process (De la Fuente, 2002) for receptive and productive vocabulary learning. Both studies, however, were more concerned with revealing whether L2 learning through output-based instruction would be as effective as L2 learning through input-based instruction. In other words, the central focus of both studies was not on examining differences between multiple PO activities. Further investigation is needed to examine the effectiveness of various spoken PO activities for vocabulary learning, both SWs and FSs.

There are many general speaking activities used in language instruction, including retelling, ask and answer, 4/3/2 and pass and talk. These activities can be designed with vocabulary-focused goals and implemented as spoken PO activities, using factors such as topic, time pressure and information distribution. However, there is very little empirical evidence for or against their effectiveness. In order to help better assess the learning conditions of the different activities, there are two frameworks that might be used: ILH (Hulstijn & Laufer, 2001) and TFA (Nation & Webb, 2010)[6]. Their use allows for predictions about the relative effectiveness of the various activities. There have been

[6] Refer to Section 2.4 for a comprehensive review of the ILH and TFA.

indeed studies examining some of the written PO activities in relation to the ILH (e.g., Hulstijn & Laufer, 2001; Zou, 2017; Keating, 2008; Rassie, 2017; Sun, 2017); however, no previous study has examined the effectiveness of spoken PO activities for learning vocabulary while conceptualising the activities within either ILH or TFA frameworks. Thus, it might be beneficial to explore the effectiveness of multiple spoken PO activities, which differ from one another in IL or TFA score, in the acquisition of vocabulary.

Another aspect that has not received much attention in studies investigating PO for vocabulary learning studies is the examination of PO activities for teaching FSs. Only Nassaji and Tian (2010) have examined the use of PO activities for the acquisition of FSs. With the limited empirical evidence available, it is not yet known whether the findings about differences among the activities featured in previous studies (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017) on teaching SWs can also be applied to FSs. Finally, the teaching of polysemy is another aspect that has been neglected in the literature of teaching vocabulary in general, and more specifically within the PO literature.

5.2 Research questions

Having examined the effectiveness of spoken PO instruction in comparison to traditional instruction for receptive and productive learning of polysemous SWVs and PVs, I wanted to compare the efficacy of different spoken PO activities concerning the receptive and productive learning of polysemous SWVs and PVs. Differences among three PO activities were first conceptualised using the ILH and TFA frameworks, and it was found that the three activities differed in their ILs and TFAs indices (will be presented in Section 5.3.5). The effectiveness of these activities was then assessed in a study

involving a five-week treatment in which learners were exposed to 12 target SWVs and 12 target PVs through three spoken PO conditions (activities): sentence reconstruction, listen-and-retell meaning and picture description. Vocabulary learning gains were assessed by means of receptive and productive test(s). The primary objective was to examine whether any of the activities prove to be more effective than the others for the learning of polysemous SWVs and PVs. The data was analysed following the same procedure used in Study 1: examining the learning gains of the target items in the present chapter (see Chapter 3), as well as examining the lexical profiling of participants' responses, which will be presented in Chapter 6 (see Chapter 4). The following research questions were addressed:

1. Is there a difference between the sentence reconstruction, listen-and-retell meaning and picture description activities in terms of learning gains?
2. Is there a difference between SWVs and PVs in terms of receptive and productive learning gains within each treatment condition?
3. Is there a difference between the three meaning-senses of the target items in terms of receptive learning gains and productive learning gains within each treatment condition?

Based on previous research findings, it was expected that the activity that induces the highest IL and TFA score would result in higher learning gains on both the receptive and productive levels (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017). Further, it was expected that the learnability of SWVs and PVs would be relatively similar due to the fact that both would receive a similar type and amount of instruction (Pellicer-Sánchez, 2020). The knowledge of the meaning senses was expected to be different, with the first meaning sense more being better learned than

the other two meaning senses of a target item, due to the "frequency effect" observed in previous studies in the literature of usage-based approaches (Ellis, 2002; Gries, 2008; Waring & Nation, 1997) and also in the receptive learning gains in Study 1.1.

5.3 Methodology

5.3.1. Participants

The participants in this study were 90 EFL learners of English in the Preparatory Year programme at King Abdul-Aziz University (Rabigh Branch) in Saudi Arabia (refer to Section 1.2 for more details about the programme). These participants did not participate in Study 1. The minimum time the participants had studied English in a formal setting was nine years. They ranged in age between 18 and 21 years old, as required for admission to the Preparatory Year programme. Their proficiency level can be considered low-intermediate, as indicated by their vocabulary levels test (VLT) scores (will be discussed in Section 5.4.1). Forty-one participants were excluded from the study due to failure to complete one of the tests, failure to attend one of the treatment sessions or failure to reach the minimum score for mastering the 2K band on the VLT. The final pool of participants included in the analyses consisted of 49 participants (females, mean age = 19).

5.3.2. Treatment groups

The study design included three treatment groups, which corresponded to the three PO activities: sentence reconstruction, listen-and-retell meaning and picture description. The study was conducted with three intact classes, with each class being assigned to one of the treatment conditions. After the exclusion procedure, the number of participants in

each group was as follows: the sentence reconstruction group (n=18), the listen-and-retell meaning group (n=18) and the picture description group (n=13). All the three groups completed the pre- and post-tests. In addition, all three groups received the same amount of vocabulary instruction, the only difference being the activity they practiced. They received a 10- to 15-minute presentation per session, with a focus on the target vocabulary, similar to what was done in Study 1 (refer to Section 3.3.4.2 for more details on the target items presentation). Afterwards, they were made to practise the taught vocabulary for one hour per session. The three PO activities were inserted in spoken information-gap format, which will be explained in detail in Section 5.3.5.

5.3.3. Timeline of data collection

The data collection for Study 2 took place over a five-week period during the second semester of the Saudi university year, which ran from January 2017 through May 2017. In the first week, a pre-test was administered, along with the VLT and ethics forms. In addition, a practice session with non-target items was conducted in order to familiarise the students with the type of activity and the speaking test. To do so, I taught the students a short mini-lesson on non-target items (2 SWVs and 2 PVs) and led each class in either sentence reconstruction, listen-and-retell meaning and picture description activities that foreshadowed the data-gathering sessions. At the end of the class, a speaking practice test was carried out. The instructions on how to complete the test were given, and learners completed it as group work. In the second week, the treatment sessions started. Two teaching sessions were conducted per week in which three SWVs and three PVs (each with three meaning senses) were taught in each treatment session (n=18 meaning sense per session). The treatment sessions were 90 minutes each, the normal class time at the university. In the third week, the first post-test was administered for the items that had been taught in the previous week. The treatment session continued in the fourth week,

and the second post-test was given to the students in the fifth week.

5.3.4. Target items

The target items consisted of 12 polysemous PVs and 12 polysemous SWVs (each with three meaning senses), which were selected from the set of target items used in Study 1. One criterion was considered in the selection of the target items: the included items were the ones for which participants received the lowest scores in the pre-test in Study 1. The cut-off point for inclusion was a maximum of five correct answers (25%) per meaning sense within each group, as the highest possible score of correct answers per group was 20. Items which were scored as correct more than five times within each group were excluded. With this selection process, I aim to ensure that most of the target items would be unfamiliar to the students, since the participant groups in Study 1 and 2 were enrolled in the same programme at the same university. The target SWVs in the present study are *blow, clear, collect, commit, count, fall, hit, introduce, relate, run, settle* and *train*. The target PVs are *back up, come around, fill in, get through, hold back, hold up, make up, pass on, put up, set off, take in* and *take up*. The content of the target items' presentation was the same as in Study 1; only the order was modified so as to follow the new distribution of target items. For more details about the target items and design of the presentation materials, refer to Section 3.3.4.

5.3.5. Spoken PO activities

For the purpose of the present study, the participants were assigned to three spoken PO activities: a sentence reconstruction activity, a listen-and-retell meaning activity and a picture-description activity. These activities were chosen for three reasons: first, the three activities differed slightly in the lexical aspects they tap into. In other words, the

lexical retrieval process involved in each of the activities to complete them is different, i.e. retrieval of form, retrieval of meaning, or retrieval of form-meaning link. Secondly, both sentence reconstruction and picture description were investigated in previous studies examining vocabulary learning, through written output for the sentence-reconstruction activity (e.g., Nassaji & Tian, 2010) and through spoken output for the picture description (e.g., Ellis & He, 1999; De la Fuente, 2002). Thirdly, the three activities were designed to operationalise different levels of IL. The following section begins with a description of these activities, then an overview of their conceptualisation within the ILH and TFA frameworks. Finally, a pilot study was conducted to evaluate the activities, the results of which are presented in the last part of the following section.

5.3.5.1. Description of the activities

The three activities meet the characteristics of the pushed-output instruction examined in the present thesis: interactive (learners had to work in pairs (A and B) using an information-gap format to complete the activities), communicative (learners need to sharing the information they had with their partners and give feedback), and of spoken modality that are implemented under time pressure (the time spent on the activity was controlled to be same for each group, one hour), as discussed in Section 2.3.1. Apart from these elements, the context provided in the activities was taken from the BNC spoken corpus to ensure that the context given in the sentences represent typical sentences with the target items that learners may encounter in everyday life. Further, the context was checked using *VocabProfile* on the *Lextutor* website (Cobb, 2015) to ensure that all words were within the first 2K of the VLT.

The sentence reconstruction activity. Working in pairs (A and B), the learners were asked to reconstruct some sentences read aloud to them, with a focus on replacing the verbs in the given sentences, preferably with the target items. Learner A read the

sentences (ranging from 5 to 7 in number) twice, one by one, while Learner B took notes if desired. Then, Learner B attempted to retell the sentence orally to Learner A. The roles were then reversed, and Learner B would read a different set of sentences aloud, and Learner A listened, took notes and retold the sentences orally, attempting to use the target items. This way, both learners had the opportunity to practice using the target items with their meaning(s) in spoken output. The target items were not provided in the activity. Examples of sentences used in this activity include:

> I can't believe that he spent 10 years in prison for a crime that he didn't *do* (commit).
>
> The news of her friend's death strongly *affected* her, so I tried reminding her that her friend *moved on* to a better place (hit, pass on).
>
> The temperature *reached* 53°C yesterday, so the football players could not *exercise* as much as usual (hit, train).

As explained in Section 5.1.1, the text reconstruction activity has been one of the activities most commonly used in research on vocabulary learning through written PO instruction. Therefore, it was selected as one of the activities in this study, though modified to allow for reconstruction on a sentence level. In this way, the design of the activity would allow the incorporation of the different target items with their multiple meanings. The type of retrieval required to complete the activity is the retrieval of the form. It could be argued that the activity may also be considered a meaning-based activity, as learners expand their understanding of the language available to reconstruct the sentence they have (Wajnryb & Maley, 1990). Yet, the instruction in the present study requested learners specifically to replace the existing verbs in the given text, which further emphasises the activity's focus on lexical forms. The full activity can be found in

Appendix 8.

The picture description activity. The study used the same or different format of the picture description activity with keywords. The learners needed to decide whether they held the same pictures by describing them to their peers. In each pair, Learner A had a sheet which included the six target items as keywords and 18 pictures representing the three meaning senses of each of these target items. Learner B received a similar sheet with the same keys and pictures, but with a different order: some of the pictures were in the same positions, and some were put in different positions as Learner A's sheet. The pictures in the activity served as vessels for the intended meaning(s) for each item. The keywords (target items) were provided to help learners retrieve the form-meaning(s) links of the target items. A star next to a picture would indicate whether it was Learner A or Learner B who would describe the first picture. For instance, if Learner A had a star next to Item 1, she would begin describing the first picture. Learner B listened to the description, asked Learner A any questions if needed and looked at the first picture on their sheet to decide whether or not the picture was the same as Learner A's. If it was the same, Learner B said so, and both wrote 'S' on their sheets next to Item 1. If it was different, both wrote 'D' next to Item 1 and Learner B described the picture she had for learner A. Then, they moved to Item 2, for which Learner B began describing the picture and Learner A listened, asked questions and decided whether to write 'S' or 'D'. Each pair worked through their items in this way. The successful completion of the activity depended in the first place on the learner's ability to provide a clear and accurate description of the picture she saw.

As explained in Section 5.1.2, the picture description activity was investigated in studies examining vocabulary learning through spoken PO activities (e.g., De la Fuente, 2002; Ellis & He, 1999), and the results showed it to be beneficial for vocabulary learning. The format investigated in both Ellis and He (1999) and De la Fuente (2002) was the

listen-and-do, while the present study investigated the same-or-different format. Further, the type of retrieval required to complete the activity is the retrieval of the form-meaning link, as learners were provided with the target items as keys as well as the multiple meaning senses (represented in the pictures) and were encouraged to use the keys to describe the given pictures. The complete materials for this activity can be found in Appendix 9.

The listen-and-retell meaning activity. In this activity, learners received a worksheet that included nine sentences for the meaning senses of the target items. Learners A and B received different sheets with different sentences. Learner A read a sentence which included the target items twice, while Learner B listened. Learner A asked about the meaning and/or translation of the target items, and Learner B had to orally provide them. Then, the roles were reversed, and Learner B read and asked about the meaning, while Learner A listened and orally retrieved the meaning for the target items. The sentences in the activity were the same used in the sentence reconstruction group; the only difference was that the target items were provided within the context of the sentences. Below are some samples for the same target items provided in the previous section.

I can't believe that he spent 10 years in prison for a crime that he didn't *commit*.

The news of her friend's death strongly *hit* her, so I tried reminding her that her friend *passed on* to a better place.

The temperature *hit* 53°C yesterday, so the football players could not *train* as much as usual.

The selection of the listen-and-retell meaning activity was mainly motivated by the specific features of the activity. Learners were provided with the forms in the given sentences and needed to recall their meanings: the type of retrieval required to complete the activity is the retrieval of meaning. It must be noted that the listen-and-retell meaning activity has not been investigated in previous PO studies or may not always be considered an output activity. However, it was deemed that the elements of the design discussed in Chapters 2 and 3 for developing the "push" elements of an activity (such as amount and nature of feedback, information distribution, time pressure) could enhance the design of the activity such that it could be considered a PO activity. The reason behind including this activity relates to the lexical aspect learners tap into to complete the activity, which is different from that of the other two activities, i.e. the retrieval of the meaning.

When recalling the meaning, learners can make use of what was already learned in the presentation session or the cues in the sentence context. Thus, they may employ some guessing from context in order to complete the activity. Guessing from context involves "making informed guesses as to the meaning of a word in the light of all available linguistic cues in combinations with the learner's general knowledge of the world, her awareness of context and her relevant linguistic knowledge" (Haastrup, 1991, p. 40). This has been recognised as an effective strategy for vocabulary learning (e.g., Chern, 1993; Clarke & Nation, 1980; Fraser, 1999; Huckin & Bloch, 1993; Kanatlar, 1995; Mondria & Boer, 1991; Rott, 1999; Soria, 2001). However, some studies have reported that it relies heavily on aspects other than lexical knowledge of the target item, such as extant lexical knowledge, effective recall and memory usage (e.g., Jenkins, Matlock & Slocum, 1989; Liben & Posnansky, 1977; Nassaji, 2006; Nation, 1982). The instruction provided in the activity required learners to only recall the meaning of the target items, relying on the contextual clues on a sentence level. Furthermore, as pointed out by Paribakht (2005), when single sentences are provided as the primary sources for inferring meaning, learners

effectively utilise the clues in these sentences. The complete activity can be found in Appendix 10.

In order to better understand potential differences among the activities selected for the study, I have applied the ILH and TFA frameworks. The next sections present the conceptualisation of the activities within the ILH framework first followed by their conceptualisation within the TFA framework.

5.3.5.2. Conceptualising the three activities within the ILH framework

According to Hulstijn and Laufer (2001), the Need component of the ILH framework is a motivational aspect of the construct related to the reason for knowing a word. This aspect is operationalised at three levels: absent, if the activity does not impose any requirements for learning the target items; moderate, if the need is imposed by an external factor such as the teacher or the requirements of an activity; or strong, if it is imposed by learners themselves. The Search component refers to the attempt to determine the meaning of unknown words. It can be either present, if the activity requires the learners to seek the meaning of unknown words, or absent, if no such effort is required and the meaning can be inferred from the context. The third component, Evaluation, relates to making comparisons and decisions about the word's suitability in a given context. It can be either absent; moderate, if it involves choosing the best meaning to fit a given context; or strong, if the activity requires deciding how to combine a word with additional words in an original sentence, as explained in Chapter 2 (Section 2.4.2) for more details on the ILH framework).

In order to ensure the accuracy of the estimated ILs, three ESL teachers in UK were asked to rate each activity based on the ILH framework. The rating sheets included a sample of each activity and a comprehensive definition of each component, as defined by

Hulstijn and Laufer (2001). The scores for both Need and Evaluation ranged between 0-2 (0 if absent, 1 if moderate and 2 if strong) whereas the scores for the Search component ranged between 0-1 (0 if absent or 1 if present). The rating sheet can be found in Appendix 11.

The three raters agreed on the scores of each component for the three activities, as presented in Table 5.1. In all three activities, Need was estimated to be 'moderate'. This was expected, as the Need is imposed by an external factor, i.e. the requirement of the activity itself (Hulstijn & Laufer, 2001). The Search component was scored as 'present' only in the picture description activity but not in the other two activities. The explanation for this scoring offered by the three raters is that the picture description activity requires learners to search for the correct word to represent the given picture within the given keys. The Evaluation component was scored 'moderate' in the listen-and-retell meaning and picture description activities and 'strong' in the sentence reconstruction activity. The raters explained that learners were not engaged in any use of the target items in the listen-and-retell meaning activity and did not evaluate their use of the target items in other given contexts in the picture description activity. Consequently, Evaluation is moderate in both activities. In contrast, in the sentence reconstruction activity, learners were engaged in comparing their own use of the target items with the ones given in the activity.

Table 5.1 *Raters' scores for the three activities (sentence reconstruction, listen-and-retell the meaning and picture description) within ILH framework*

	SRG			LRMG			PDG		
	R 1	R 2	R 3	R 1	R 2	R 3	R 1	R 2	R 3
Need	1	1	1	1	1	1	1	1	1
Search	0	0	0	0	0	0	1	1	1
Evaluation	2	2	2	1	1	1	1	1	1
Total ILs	3	3	3	2	2	2	3	3	3
Average ILs*		3			2			3	

Note: SRG = sentence reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture-description group.

R= rater

Average ILs = average total ILs scores of the three raters

My hypothesis in Section 5.2 was that an activity that induces higher ILs would lead to higher learning gains on both the receptive and productive levels, as evidenced in the literature (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017). Both sentence reconstruction and picture description are estimated to have similar ILs in total (though they differ in scores per component). It was therefore expected that both would result in similar learning gains, whereas the listen-and-retell meaning activity, with its lower IL, would be the least effective. Overall, the pattern of the effectiveness for the three activities according to the ILH is expected to be as follows:

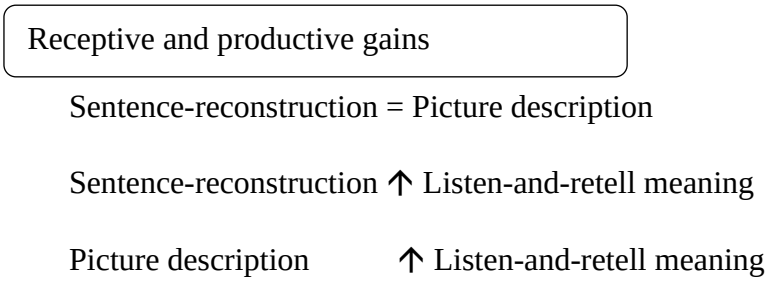


Figure 5.1. Pattern of the effectiveness for the three activities within ILH framework

5.3.5.3. Conceptualising the three activities within the TFA framework

With reference to the TFA framework (Nation & Webb, 2011; Webb & Nation, 2017), five criteria are used to evaluate the conditions of vocabulary learning of an activity: motivation, noticing, retrieval, generations (varied encounters and varied use) and retention. Motivation relates to the purpose of the task; noticing deals with the learner's attention on the target word in the task; retrieval deals with whether the task involves receptive or productive retrieval, recognition or recall retrievals, single or multiple retrievals and spaced or massed retrievals; generation (also termed as varied encounters and varied use) may be receptive when the new words are encountered in a new context or productive when using the words in an original context; and retention deals with ensuring a successful form-meaning link and whether forming this link requires instantiation or imaging with no interference (refer to Chapter 2, Section 2.4.3 for more details on the TFA framework).

In order to ensure the accuracy of the estimated TFAs, the same three raters who rated the activities in terms of the ILH framework were also asked to evaluate the activities according to the TFA framework. The rating sheets included an example of each activity and comprehensive definition of each component as defined by Nation and Webb

(2011). The TFA framework consists of a checklist of features that covers the five components: motivation, noticing, retrieval, generations (varied encounters and varied use) and retention. Several questions are asked for each component, and point values are used in evaluating the features, with each question answered with either 0 or 1. The maximum possible score differs for each component, and the highest possible number of points is 18. Table 2.4 in Chapter 2 presents the TFA checklist (Nation & Webb, 2011, p. 7; Webb & Nation, 2017, p. 236). The rating sheet can be found in Appendix 12.

The three raters provided different scores for some of the questions, so multiple Fleiss' kappa tests per activity were carried out to determine if there was agreement between the three raters' judgements on each of the features. Fleiss' kappa is a statistical measure used for assessing the inter-rater reliability of agreement among more than two raters (Fleiss, 1971). The three raters reached high inter-rater reliability for each activity: the sentence-reconstruction activity (κ = .911 (95% CI, .644 to .1.178), p < .0005), the listen-and-retell meaning activity (κ = .926 (95% CI, .659 to 1.192), p < .0005) and the picture description activity (κ = .932 (95% CI, .657 to 1.190), p < .0005). Table 5.2 presents the rating per activity and shows that the instances of disagreement were three in total, all regarding the 'noticing' feature: the question about negotiation for both the sentence reconstruction activity and the listen-and-retell meaning activity, and the question about raising awareness for new vocabulary learning in the picture description activity. The average of the three TFAs scores per rater was the final TFA score for each activity. It can be seen in Table 5.2 that the sentence reconstruction activity received the highest score, followed by the picture description and the listen-and-retell meaning activities, respectively. This ranking is different from the one created using the ILH framework. Hence, my hypothesis was different here as well: it was expected that the sentence reconstruction activity would result in higher learning gains than the picture

description and the listen-and-retell meaning activities, respectively. This difference between the ILH- and TFA-based rankings was predicted, as ILH is a much simpler framework than TFA and TFA was developed to include several features that other research has shown to be important for designing vocabulary teaching techniques (Nation & Webb, 2011).

Table 5.2 *Raters' scores for the three activities (sentence reconstruction, listen-and-retell meaning and picture description) within TFA framework*

Criteria	SRG			LRMG			PDG		
	R 1	R 2	R 3	R 1	R 2	R 3	R 1	R 2	R 3
Motivation									
Is there a clear vocabulary learning goal?	1	1	1	1	1	1	1	1	1
Does the activity motivate learning?	1	1	1	1	1	1	1	1	1
Do the learners select the words?	0	0	0	0	0	0	1	1	1
Noticing									
Does the activity focus attention on the target words?	1	1	1	1	1	1	1	1	1
Does the activity raise awareness of new vocabulary learning?	1	1	1	1	1	1	**0**	**1**	**1**
Does the activity involve negotiation?	**1**	**0**	**1**	**1**	**0**	**1**	1	1	1
Retrieval									
Does the activity involve retrieval of the word?	1	1	1	1	1	1	0	0	0
Is it productive retrieval?	1	1	1	0	0	0	0	0	0
Is it recall?	1	1	1	1	1	1	0	0	0
Are there multiple retrievals of each word?	0	0	0	0	0	0	0	0	0
Is there spacing between retrievals?	0	0	0	0	0	0	0	0	0

Criteria	SRG			LRMG			PDG		
	R 1	R 2	R 3	R 1	R 2	R 3	R 1	R 2	R 3
Generation (varied encounters and varied use)									
Does the activity involve generative use (varied encounters and use)?	1	1	1	0	0	0	1	1	1
Is it productive?	1	1	1	0	0	0	1	1	1
Is there a marked change that involves the use of other words?	1	1	1	0	0	0	1	1	1
Retention									
Does the activity ensure successful linking of form and meaning?	1	1	1	1	1	1	1	1	1
Does the activity involve instantiation?	1	1	1	1	1	1	0	0	0
Does the activity involve imaging?	0	0	0	0	0	0	1	1	1
Does the activity avoid interference?	0	0	0	1	1	1	0	0	0
Total TFAs	13	12	13	10	9	10	10	11	11
Average TFAs*		13			10			11	

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

R= rater

Average TFAs = average total TFAs scores of the three raters

Therefore, the pattern for the effectiveness of the three activities according to the TFA is expected to be as follows:

Receptive and productive gains

Sentence-reconstruction ↑ Picture description

Sentence-reconstruction ↑ Listen-and-retell meaning

Picture description ↑ Listen-and-retell meaning

Figure 5.2. Ranking of the effectiveness of the three activities within the TFA framework

5.3.5.4. Piloting of the activities

Prior to data collection, the three activities were piloted with 10 postgraduate students: four NSs of English, one bilingual Arabic/English speaker, one Spanish, one Chinese and three Saudis. They were asked to report on the clarity of the instructions, the appropriateness of the pictures, the naturalness of the sentences and the appropriateness of the answers. Some adjustments were made based on this feedback. For instance, a sentence was reworded because it caused confusion, and a picture was replaced when it did not reflect the intended meaning.

5.3.6. Measurement instruments

The instruments used in Study 2 are the same ones used in Study 1: the VLT (Schmitt, Schmitt & Clapham, 2001), the receptive pre- and post-test (form-meaning

recognition) and the productive/oral post-test (form-meaning recall). Both the receptive and productive tests were adapted to the study with a few changes; this section provides a description of these changes. For a full description of the measurement instruments, refer to Section 3.3.5.

5.3.6.1. The receptive pre- and post-test (form-meaning recognition)

Design. The form-meaning recognition test was the same as the one used in Study 1, but the test items were distributed in a new order, since only half of the target items in Study 1 were selected. It presented learners with 72 test items in 24 clusters: 72 keys, 72 definitions and 72 distractors. Each cluster was designed to assess knowledge of three target items consisting of six options in the left column – three keys and three distractors – and three L1 definitions in the right column (see Figure 5.3). The form-meaning recognition test was used in both the pre-test and post-test. Though the format of the pre- and post-tests were similar, the administration was altered to more accurately simulate authentic pedagogical practices, with the pre-test presented in a seven-page booklet containing the full test. The time taken to complete the receptive pre-test was 90 minutes (session time). The same test with 72 items was divided into two test booklets of four pages and 36 items each. There were two testing sessions (refer to Section 5.3.3). The full post-test session was 90 minutes, and the time taken to complete the receptive post-test was 30 minutes within each session. The full test can be found in Appendix 13.

A. commit **B.** run **C.** blow **D.** make up **E.** take up **F.** hold back	1. يركض 2. يستحوذ على شيئاً ما (مساحة/وقت) 3. يبدد المال/يُبذر	1. (A) (B) (C) (D) (E) (F) 2. (A) (B) (C) (D) (E) (F) 3. (A) (B) (C) (D) (E) (F)

Figure 5.3. A sample of the receptive form-meaning recognition test

Piloting results and modifications. Though the test was comprehensively piloted for Study 1 to ensure the clarity and feasibility of all of test items and formats, it was piloted again for Study 2 to ensure the appropriateness of the new distribution of target items without any overlap between the meaning senses. The piloting was done in two stages. First, the test was given to an Arabic/English bilingual PhD student in Applied Linguistics. She did not report any problematic issues with the new distribution and correctly answered all the questions. Secondly, the test was given to 15 students enrolled in the same Preparatory Year programme as the participants. These students were encouraged not to guess blindly but to answer only the items about which they were confident. In general, the results of the piloting indicated that only seven test items were found to be problematic due to overlapping in their meaning senses in the new distribution and needed to be altered.

Scoring. The scoring of the receptive form-meaning recognition test was dichotomous, with responses scored as correct only if the appropriate meaning was selected and incorrect if a wrong answer was chosen or if the question was left blank. The highest possible total score was 72 (one for each of the meaning senses).

5.3.6.2. The productive post-test (form-meaning recall)

Design. The oral form-meaning recall post-test was identical to the one used in Study 1 aside from the deleting of irrelevant questions about the non-selected target items. This test included two types of questions: discourse completion tasks (DCTs) and factual questions. In the DCTs, learners were asked to respond to some culture-related situations. The factual questions were questions for target items for which creating a DCT was not feasible, and learners were asked to respond to these questions. Both the DCTs and factual questions were presented in a written mode in a PowerPoint presentation, and learners responded orally to the questions. The test included 72 test items: 51 DCTs and 21 factual questions. As shown in Figure 5.4, due to the time-consuming nature of this test, it was only used as a post-test. The weekly version of the test included 36 test items. The time taken to complete the productive post-test was 40 minutes within each session, 80 minutes in total over the two post-test sessions. The full test can be found in Appendix 14.

A DCT sample

Imagine that you are a police officer and you found out that there is a bomb in a building that you are in it. What is the first thing you will do?

تخيلي أنك تعملين كضابطة شرطة ووجدت ان هناك قنبلة في المبنى الذي انت فيه حالياً فما هو أول شي ستقومين بفعله؟

Possible answer: Clear the building …….

A factual question sample

Lately, there were some terrorist attacks in some of the famous mosques in Saudi Arabia, what did they do?

في الفترة الأخيرة كانت هناك أعمال إرهابية بالقرب من بعض المساجد في المملكة العربية السعودية فما الذي فعله الإرهابيون ؟

Possible answer: set off bombs in the mosques …….

Figure 5.4. A sample of the productive form-meaning recall post-test

Transcribing and Scoring. The duration of each student's recording was 80 minutes, totalling 3,920 minutes in all and 20,231 words. The test was scored using the two approaches adopted for Study 1.1 and 1.2. The first approach involved accounting for the occurrence of the target items in the responses by giving either a score of 1 if the target items occurred or a 0 if they did not. The second approach of the analysis involved examining the lexical profiles of the utterances in which the target items were used (i.e., MLU and lexical richness). Results for the first approach are presented in this chapter. The results of the second approach will be presented in Chapter 6.

5.3.7. Analysis

The following section reports the results of the receptive and productive gains. It first presents the results of the VLT and the pre-test. Then, the descriptive statistics of the receptive and productive pre- and post-tests are presented. The results of the analysis are presented in two sections: one on the receptive learning gains and one on the productive learning gains. The data were analysed using SPSS (IBM Corp., 2013).

The same procedure for calculating the absolute and relative gains used in Study 1.1 was also followed in the present study. For receptive gains, the means of the receptive pre-test were subtracted from the means of receptive post-test. For the productive gains, the scores of the receptive pre-test were used, due to the absence of a productive pre-test, and were subtracted from the means of the productive post-test. This creates the same limitation as in Study 1.1 of using a more conservative approach in calculating the scores, as discussed in Section 3.3.7. Further, only relative gains were used in the inferential statistics; however, when the same analyses were run using absolute gains, the same pattern was found.

Prior to completing any tests, the normality of the scores on the VLT, pre-test and post-tests was checked using Kolmogorov-Smirnov tests, and the data was found to be non-normally distributed. That means that the data should be analysed using nonparametric tests. However, the most suitable test for the data is the mixed-design ANOVA with repeated measures. There is no nonparametric equivalent to this test. When running multiple Kruskal-Wallis tests, the patterns of the results were similar to the ones shown after running the mixed-design ANOVAs. Therefore, it was judged more efficient to report the results of the mixed-design ANOVAs. Two mixed-design ANOVAs with two repeated measures were carried out for the receptive relative gains (mixed-design ANOVA-1) and the productive relative gains (mixed-design ANOVA-2). The between-subjects variable was the PO activity (sentence reconstruction, listen-and-retell meaning

and picture description) while the target item's type (SWV or PV) and meaning sense (1, 2, or 3) were the within-subjects variables.

The assumption of sphericity was checked for the two ANOVAs and showed that none of the effects violated this assumption (*p.*>.05). The results of the Levene's test in the two ANOVAs indicated no violation in the assumption of homogeneity of variance (*p.*>.05) except for the receptive relative gains of sense-1 in SWVs (*p.*<.05). The Games-Howell post-hoc tests were then used, as they are recommended by Field (2013) for being the most accurate post-hoc tests when equal variance is not assumed. Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of .025. The effect sizes are estimated according to Plonsky and Oswald's (2014) guidelines. They suggested interpreting r values as follows: r > .25 small effect; r > .4 medium effect; r > .6 large effect. This interpretation was chosen for the present thesis.

5.4 Results

5.4.1. The VLT and previous knowledge of target items

Prior to conducting any analysis, potential differences in participants' vocabulary knowledge scores were examined through the comparisons of their VLT scores and their pre-test scores for the target items. Participants scored 25.29 on the 2K level (max= 30, *SD*= 2.02) whereas they scored 13.37 on the 3K level (max= 28, *SD*= 4.26). These results indicated that all participants had mastered the 2K level and not yet mastered the 3K level. The differences in VLT scores across the three treatment groups were examined using multiple Kruskal-Wallis tests. The results showed no significant differences between the three groups, neither when analysing the scores per level, VLT-2K [X^2 (2, *n*=49) = 0.23,

p= 0.89], VLT-3K [X^2 (2, *n*=49) = 0.97, *p*= 0.62], nor when analysing the combined scores for the two levels, VLT-2K+3K [X^2 (2, *n*=49) = 2.26, *p*= 0.32]. To evaluate differences among the three treatment groups on their previous knowledge of the target items, multiple Kruskal-Wallis tests were conducted for the combined pre-test (i.e., SWVs pre-test + PVs pre-test), SWVs pre-test and PVs pre-test scores. The results showed no significant differences between the three groups in the combined pre-test scores [X^2 (2, *n*=49) = 0.02, *p*= 0.99], the SWVs pre-test scores [X^2 (2, *n*=49) =0.13, *p*= 0.94] or the PVs pre-tests scores, [X^2 (2, *n*=49) = 1.68, *p*= 0.43]. All groups were comparable not only in terms of their level of vocabulary knowledge but also their previous knowledge of the target items. Table 5.3 presents a summary of the learners' scores on the VLT (2K, 3K and combined) and the pre-tests (combined, SWVs and PVs) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description).

Table 5.3 *Descriptive statistics for VLT scores and pre-tests by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	SRG	LRMG	PDG
	M (SD)	M (SD)	M (SD)
Number of students	18	18	13
2,000 VLT (max=30)	25.39 (2.00)	25.28 (2.11)	25.15 (2.08)
3,000 VLT (max=30)	12.39 (5.61)	14.00 (2.87)	13.85 (3.74)
Combined VLT (max=60)	37.78 (6.30)	39.28 (3.10)	39.00 (4.51)
Combined pre-test (max=72)	12.67 (6.38)	15.11 (11.31)	13.15 (5.27)
SWVs pre-test (max=36)	8.44 (3.82)	8.50 (5.79)	8.46 (2.90)
PVs pre-test (max =36)	4.22 (3.98)	6.61 (6.40)	4.69 (3.38)

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

5.4.2. Descriptive statistics for the pre- and post-test scores

The descriptive statistics for the receptive and productive tests (combined, SWVs and PVs) are shown in Table 5.4. Looking at the differences by PO activity, the scores of the receptive post-test for learners completing the sentence reconstruction activity were higher than learners completing the other two activities, i.e. the listen-and-retell meaning and picture description activities. The same pattern occurred for the combined productive post-test but not within the SWVs or PVs productive post-test scores. In both SWVs and PVs, the listen-and-retell meaning group scored higher than the other two groups. With regards to differences by target type, the scores for SWVs appeared to be slightly higher than those for PVs across the three treatment groups.

Table 5.4 *Descriptive statistics for pre- and post-test scores (combined, SWVs, PVs) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Combined (max= 72)				SWVs (max= 36)				PVs (max= 36)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
Pre-test												
SRG	12.67 (6.38)	11	5	29	8.44 (3.82)	8	4	19	4.22 (3.98)	3	0	15
LRMG	15.11 (11.31)	15	3	44	8.50 (5.79)	8	2	23	6.61 (6.40)	5	0	26
PDG	13.15 (5.27)	11	7	20	8.46 (2.90)	8	6	15	4.69 (3.38)	4	0	11
Post-test (R)												
SRG	55.72 (10.75)	59	35	69	28.94 (5.27)	30	18	36	26.78 (5.93)	29	15	35
LRMG	36.50 (18.11)	33	9	66	19.06 (8.81)	17	6	33	17.44 (9.82)	15	3	33
PDG	44.85 (13.41)	48	26	69	22.77 (6.95)	24	12	35	22.08 (6.54)	24	14	34
Post-test (P)												
SRG	36.83 (11.92)	39	12	55	18.67 (6.95)	18	6	27	15.83 (6.81)	16	3	26
LRMG	27.94 (13.75)	24	9	53	21.28 (5.11)	22	9	30	19.22 (5.15)	20	5	27
PDG	34.92 (13.27)	31	17	58	19.54 (5.41)	19	12	30	16.46 (7.83)	15	5	32

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

(R) = receptive, (P) = productive.

Tables 5.5 and 5.6 present the descriptive statistics for the scores by the meaning sense (Senses 1, 2 and 3). There are minor differences amongst the three meaning senses, and the scores for items with sense-1 were higher than for the other two meaning senses (2 and 3) in both the receptive and productive post-tests.

Table 5.5 *Descriptive statistics for pre- and post-tests by meaning sense by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SWVs) (SD presented in brackets)*

	Sense-1 (max=12)				Sense-2 (max=12)				Sense-3 (max=12)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Pre-test											
SRG	2.39 (1.75)	2	0	6	2.44 (1.20)	2	1	5	3.61 (2.81)	3	0	9
LRMG	2.56 (2.48)	2	0	9	3.06 (2.48)	3	1	9	2.94 (2.26)	3	0	9
PDG	3.54 (2.03)	3	2	9	2.00 (1.35)	1	1	5	2.92 (1.19)	3	1	5
	Post-test (R)											
SRG	10.72 (1.36)	11	7	12	8.94 (2.62)	10	4	12	9.28 (2.19)	10	5	12
LRMG	7.22 (3.46)	6	2	12	5.78 (3.44)	6	1	11	6.06 (2.58)	6	3	11
PDG	9.46 (1.90)	10	6	12	7.00 (2.89)	7	2	12	6.31 (2.78)	7	2	11
	Post-test (P)											
SRG	6.94 (2.98)	8	0	11	5.61 (2.75)	5	2	10	6.11 (2.40)	6	2	10
LRMG	7.06 (2.31)	7	3	12	7.50 (1.89)	8	3	10	6.72 (2.05)	7	3	9
PDG	6.69 (1.75)	7	4	10	6.23 (2.49)	6	3	11	6.62 (2.75)	5	4	11

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

(R) = receptive, (P) = productive.

Table 5.6 *Descriptive statistics pre- and post-tests by the meaning sense by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (PVs) (SD presented in brackets)*

	Sense-1 (max=12)				Sense-2 (max=12)				Sense-3 (max=12)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Pre-test											
SRG	1.39 (1.33)	1	0	5	1.22 (1.48)	1	0	5	1.61 (1.61)	1	0	6
LRMG	2.33 (2.45)	2	0	9	1.78 (2.32)	1	0	9	2.22 (2.32)	2	0	8
PDG	1.69 (1.38)	1	0	5	1.54 (1.20)	1	0	4	1.46 (1.05)	1	0	3
	Post-test (R)											
SRG	9.39 (2.15)	10	5	12	8.50 (2.43)	9	3	11	8.89 (2.00)	10	6	12
LRMG	6.06 (3.86)	6	1	12	5.94 (3.51)	6	1	12	5.44 (2.94)	5	1	10
PDG	7.15 (2.79)	8	3	12	7.77 (2.39)	9	3	11	7.15 (2.23)	7	4	11
	Post-test (P)											
SRG	5.33 (2.14)	6	1	9	4.78 (2.78)	5	1	9	5.72 (2.99)	6	1	10
LRMG	6.17 (2.09)	6	2	10	5.89 (1.75)	6	1	9	7.17 (2.43)	8	2	11
PDG	5.31 (2.66)	6	1	11	5.08 (3.28)	4	1	11	6.08 (2.75)	6	3	11

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

(R) = receptive, (P) = productive.

Absolute receptive and productive vocabulary gains were then calculated for the combined test, by type of target item (SWVs and PVs) and by the meaning sense (1, 2 and 3) (see Tables 5.7–5.9).

Table 5.7 *Absolute vocabulary gains for combined, SWVs and PVs by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Combined (max= 72)				SWVs (max= 36)				PVs (max= 36)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
SRG	43.06 (12.71)	47	20	60	20.50 (6.64)	23	9	30	22.56 (6.88)	24	10	34
LRMG	21.39 (13.72)	20	4	57	10.56 (6.99)	9	2	28	10.83 (8.38)	9	2	29
PDG	31.69 (16.01)	29	10	62	14.31 (7.39)	11	6	28	17.38 (8.98)	18	4	34
	Productive											
SRG	24.17 (13.36)	29	0	45	10.22 (5.77)	9	2	22	11.61 (6.62)	11	2	23
LRMG	12.83 (11.94)	7	0	39	12.78 (5.64)	13	4	23	12.61 (6.62)	12	1	27
PDG	21.77 (14.87)	20	2	48	11.08 (5.75)	10	3	22	11.77 (9.00)	9	1	29

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Table 5.8 *Absolute vocabulary gains by meaning sense for sentence reconstruction, listen-and-retell meaning and picture description groups (SWVs) (SD presented in brackets)*

	Sense-1 (max=12)				Sense-2 (max=12)				Sense-3 (max=12)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
SRG	8.33 (2.57)	9	1	12	6.50 (3.05)	8	0	10	5.67 (3.40)	6	0	11
LRMG	4.67 (3.38)	4	0	11	2.72 (2.80)	2	0	9	3.11 (2.56)	2	1	10
PDG	5.92 (2.93)	6	1	10	5.00 (3.06)	5	0	11	3.38 (2.47)	3	0	9
	Productive											
SRG	4.56 (2.77)	5	0	10	3.17 (2.62)	2	0	8	2.50 (3.07)	1	0	10
LRMG	4.50 (3.28)	4	0	11	4.44 (2.57)	5	0	9	3.78 (2.37)	4	0	8
PDG	3.15 (2.08)	3	0	7	4.23 (2.62)	4	0	8	3.69 (3.07)	2	0	9

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Table 5.9 *Absolute vocabulary gains by meaning sense for sentence reconstruction, listen-and-retell meaning and picture description groups (PVs) (SD presented in brackets)*

	Sense-1 (max=12)				Sense-2 (max=12)				Sense-3 (max=12)			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
SRG	8.00 (2.40)	9	3	11	7.28 (2.78)	7	3	11	7.28 (2.37)	7	3	12
LRMG	3.72 (3.41)	2	0	12	4.17 (3.19)	4	0	10	3.22 (2.71)	2	0	9
PDG	5.46 (3.45)	6	1	12	6.23 (3.32)	7	0	11	5.69 (2.78)	5	2	11
	Productive											
SRG	3.94 (2.04)	4	0	8	3.56 (2.97)	3	0	9	4.11 (2.89)	4	0	9
LRMG	3.83 (2.71)	3	0	9	4.11 (2.27)	4	0	8	4.94 (3.00)	4	1	11
PDG	3.62 (2.99)	2	0	10	3.54 (3.78)	3	0	10	4.62 (2.99)	4	0	10

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

In addition to absolute gains, and following the procedure employed in Study 1.1, receptive and productive relative vocabulary gains were also calculated for all the groups for the combined test, by type of target item (SWVs and PVs) and by meaning sense (1, 2 and 3). Tables 5.10–5.12 present these results.

Table 5.10 *Relative vocabulary gains for combined, SWVs and PVs by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (%) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
SRG	72.09 (18.61)	78	38	95	73.63 (19.55)	79	42	100	70.59 (18.96)	77	34	97
LRMG	39.99 (26.07)	32	6	86	40.13 (26.39)	33	6	88	38.65 (28.38)	27	6	85
PDG	52.84 (24.05)	55	19	95	51.53 (24.65)	48	20	97	53.69 (24.33)	60	16	94
	Productive											
SRG	39.97 (21.92)	45	0	73	37.81 (20.48)	35	6	71	36.42 (19.33)	34	6	70
LRMG	22.15 (19.68)	15	0	67	45.93 (16.02)	46	16	68	40.86 (17.27)	42	10	75
PDG	36.28 (23.59)	35	4	77	39.85 (19.86)	43	13	79	36.49 (26.21)	30	4	88

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Table 5.11 *Relative vocabulary gains by meaning sense of SWVs for sentence reconstruction, listen-and-retell meaning and picture description groups (%) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
SRG	84.94 (19.72)	90	17	100	66.76 (30.20)	75	0	100	63.03 (30.48)	72	0	100
LRMG	51.37 (33.60)	48	0	100	32.10 (29.82)	22	0	89	33.65 (22.75)	27	10	91
PDG	66.75 (24.71)	75	29	100	49.52 (29.82)	50	0	100	37.99 (27.06)	38	0	90
	Productive											
SRG	48.16 (26.68)	50	0	83	32.95 (27.72)	23	0	80	24.33 (28.18)	11	0	83
LRMG	42.74 (28.73)	41	0	100	46.30 (22.91)	55	0	82	39.58 (23.25)	48	0	73
PDG	35.73 (18.83)	33	0	70	41.69 (26.75)	40	0	88	39.76 (31.94)	25	0	90

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Table 5.12 *Relative vocabulary gains by meaning sense of PVs for sentence reconstruction, listen-and-retell meaning and picture description groups (%) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
	Receptive											
SRG	75.35 (21.22)	82	30	100	66.93 (22.12)	68	25	92	69.51 (18.41)	71	44	100
LRMG	40.99 (34.22)	28	0	100	42.77 (31.53)	40	0	100	33.13 (24.42)	24	0	78
PDG	50.98 (29.80)	56	10	100	57.22 (27.91)	70	0	92	52.94 (22.22)	55	20	92
	Productive											
SRG	37.09 (17.98)	39	0	73	32.18 (25.20)	32	0	75	39.48 (26.40)	35	0	82
LRMG	36.58 (22.58)	30	0	82	37.50 (18.85)	38	0	67	48.18 (24.25)	47	11	92
PDG	34.02 (26.10)	29	0	91	32.00 (33.52)	25	0	91	43.11 (27.08)	33	0	91

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Referring to Tables 5.7–5.12, the same pattern of results reported for the pre- and post-tests means occurred for absolute and relative gains – that is, the sentence-reconstruction group was the highest in receptive gains, whereas the listen-and-retell meaning group was the highest in productive gains. Further, for differences by target type, higher vocabulary gains occurred for SWVs over PVs. For variations among the three meaning senses, no pattern was found in which one meaning was the highest in all groups.

However, in some cases, sense-1 scored higher (e.g., SWVs receptive absolute and relative gains), whereas in other cases, sense-3 scored higher than the other two meaning senses (e.g., PVs productive absolute and relative gains). A summary of the relative vocabulary gains by meaning sense across the three treatment groups (sentence reconstruction, listen-and-retell meaning and picture description) is presented in Table 5.13, as it is used in the inferential statistics reported in Sections 5.4.3 - 5.4.4.

Table 5.13 *Summary of the relative vocabulary gains by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	SRG (n=18)	LRMG(n=18)	PDG (n=13)
	M (SD)	M (SD)	M (SD)
	Receptive		
SWVs sense-1	84.94 (19.72)	51.37 (33.60)	66.75 (24.71)
SWVs sense-2	66.76 (30.20)	32.10 (29.82)	49.52 (29.82)
SWVs sense-3	63.03 (30.48)	33.65 (22.75)	37.99 (27.06)
PVs sense-1	75.35 (21.22)	40.99 (34.22)	50.98 (29.80)
PVs sense-2	66.93 (22.12)	42.77 (31.53)	57.22 (27.91)
PVs sense-3	69.51(18.41)	33.13 (24.42)	52.94 (22.22)
	Productive		
SWVs sense-1	48.16 (26.68)	42.74 (28.73)	35.73 (18.83)
SWVs sense-2	32.95 (27.72)	46.30 (22.91)	41.69 (26.75)
SWVs sense-3	24.33 (28.18)	39.58 (23.25)	39.76 (31.94)
PVs sense-1	37.09 (17.98)	36.58 (22.58)	34.02 (26.10)
PVs sense-2	32.18 (25.20)	37.50 (18.85)	32.00(33.52)
PVs sense-3	39.48 (26.40)	48.18 (24.25)	43.11 (27.08)

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

5.4.3. Receptive learning gains

The results of the mixed-design ANOVA-1 revealed that the PO activity had a significant main effect on receptive relative gains scores [F (2,46) = 9.106, p <.001, η_p^2= 0.284] – a small effect size, according to Plonsky and Oswald (2014). The post hoc tests revealed that the scores of the sentence-reconstruction group were significantly higher than those of the listen-and-retell meaning group (p. <.001). However, no significant differences were found between the sentence-reconstruction and the picture description groups (p. =.088) or between the listen-and-retell meaning and picture description groups (p. =.319), as shown in Figure 5.5. This indicates that the sentence reconstruction group had some advantage over the listen-and-retell meaning group for the receptive learning of both SWVs and PVs, but not over the picture description group.

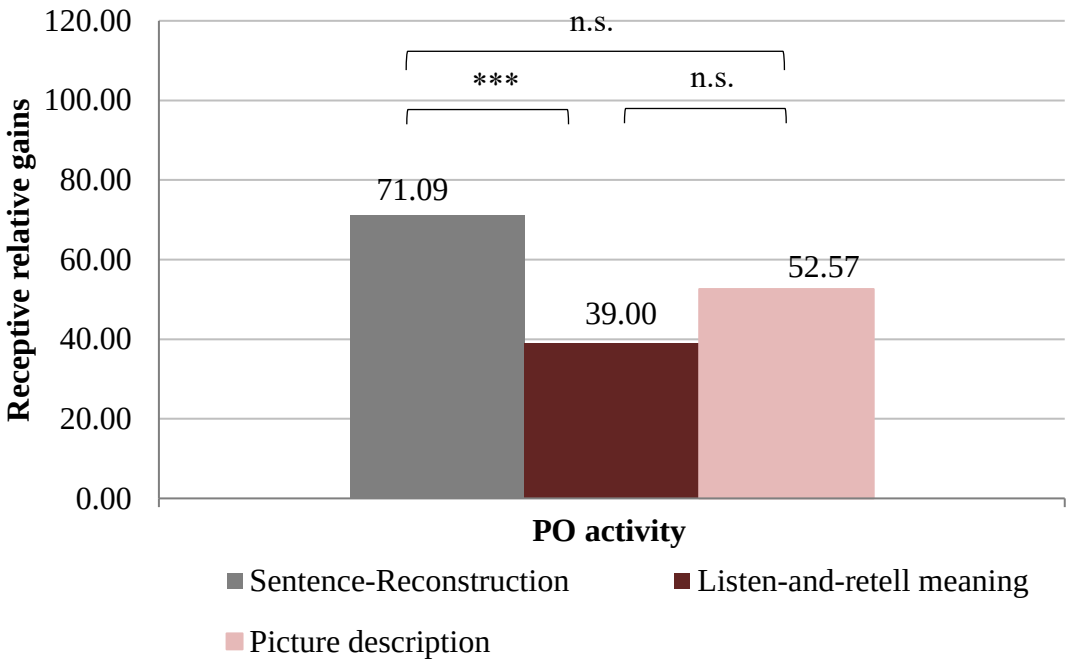


Figure 5.5. Means of overall receptive relative gains by PO activity (sentence reconstruction, listen-and-retell meaning and picture description)

The main effect of target type on receptive relative gains was not statistically significant [F (1, 46) = .040, p =.842, η_p^2 = .001]. Moreover, no significant interaction effect was found between PO activity and type of target item: F (2,46) = .205, p = .816, η_p^2= .009 (a small effect size). These results suggest that receptive gains for PVs were similar to gains for SWVs, regardless of the treatment received, which is the same pattern of findings as those of Study 1.1. Further, the results showed that the main effect of meaning sense on receptive relative gains yields an F ratio of F (2, 92) = 15.680, p < .001, with a small effect size (η_p^2 = 0.254). The follow-up contrasts revealed that Sense 1 was significantly different from the others (i.e., Sense 2 and Sense 3), whereas the difference between Sense 2 and Sense 3 was not statistically significant, as shown in Figure 5.6.

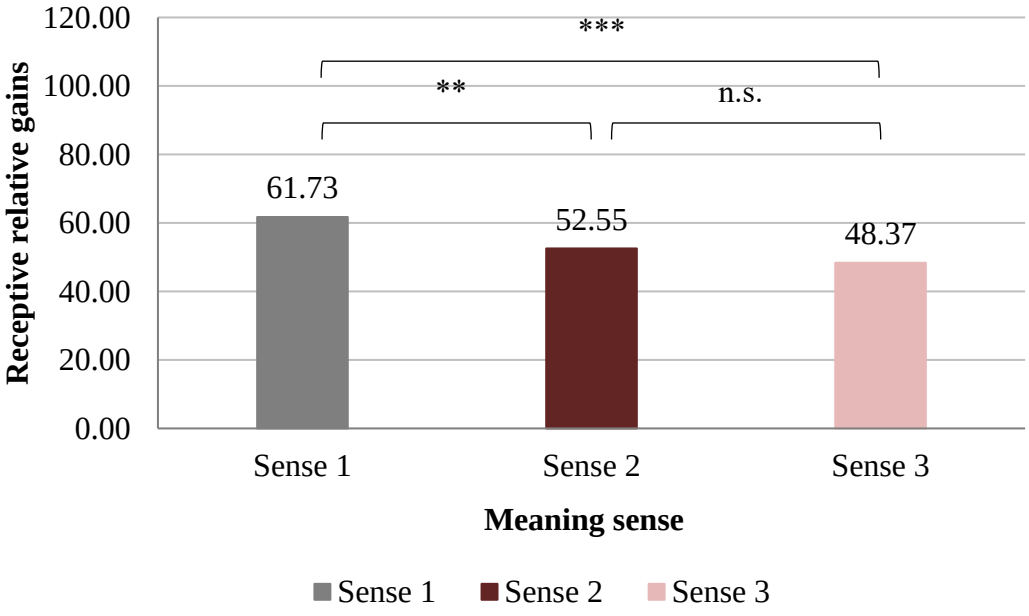


Figure 5.6. Means of receptive relative gains by meaning sense (1, 2 and 3)

There was no significant interaction effect between the meaning sense and the PO activity: F (4,92) = 0.524, p. = .702, η_p^2= .022. The meaning sense × target item's type interaction effect was significant: F (2,92) = 12.055, p <.001, η_p^2= .208, a small effect size. These results suggest that the differences among the three meaning senses differ depending on the type of target item. Figure 5.7 shows that the mean scores for the three meaning senses for PVs were relatively similar, whereas in SWVs, the difference in the mean scores for the three meaning senses is more noticeable. Further, Sense 1 scored higher in SWVs than PVs, whereas for the other two meaning senses, an opposite pattern was found in which PVs scored higher than SWVs. The meaning sense × PO activity × target item's type interaction effect was not statistically significant: F (4,92) = 1.615, p. = .179, η_p^2= .066.

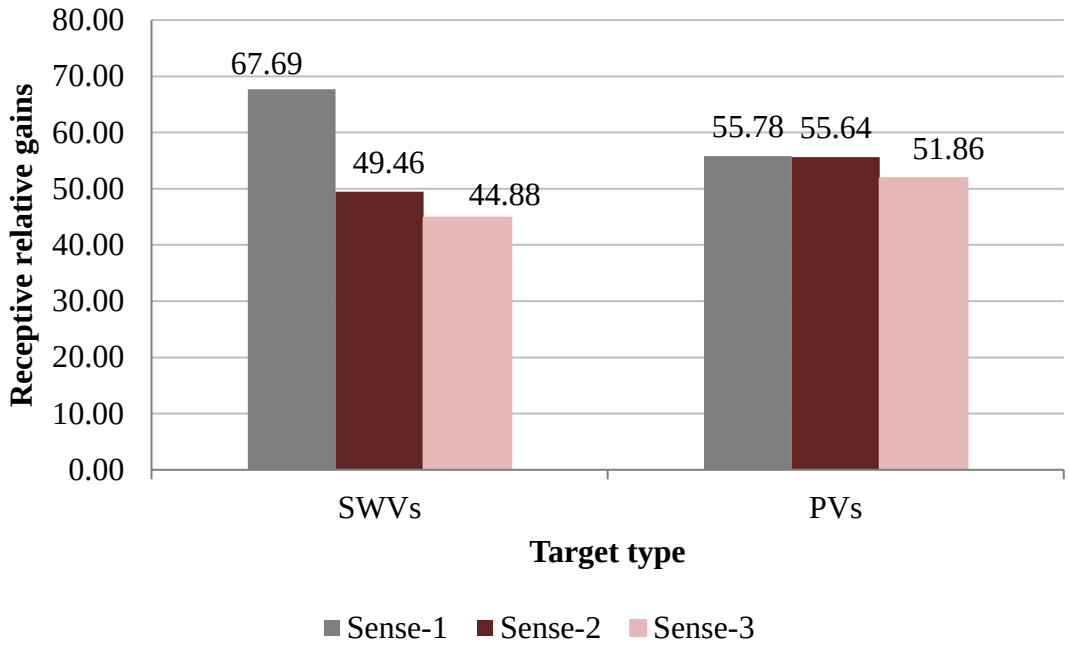


Figure 5.7. Means of receptive relative gains for the three meaning senses by target type (SWVs and PVs)

5.4.4. Productive learning gains

The results of the mixed-design ANOVA-2 showed that the main effect of the PO activity on productive relative gains was not statistically significant, F (2,46) = 0.538, p =.588, η_p^2= 0.023. These results suggest that, in terms of productive gains, a similar pattern of learning occurred across all the three activities. There was no significant main effect neither of target type, F (1, 46) = 0.275, p =.603, η_p^2 = .006, nor of meaning sense, F (2, 92) = 0.327, p =0.722, η_p^2 = .007. The interaction effect between target type and PO activity was not significant, F (2, 46) = 0.262, p =.177, η_p^2 = .011. Further, the meaning sense × PO activity interaction effect was also not significant, F (4, 92) = 2.151, p =.085, η_p^2 = .086. Interestingly, the meaning sense × target type interaction effect was significant, F (2, 92) = 4.645, p =.013, with a large effect size (η_p^2 = .092), indicating that productive learning of the three meaning senses was affected by the type of the target item, which is similar to what is found in the receptive gains. Figure 5.8 shows that the learning pattern of sense-1, the most frequent meaning sense, was the highest followed by sense-2 and sense-3 respectively whereas in PVs an opposite pattern occurred with sense-3, the least frequent meaning sense, being the highest followed by sense-1 and sense-2. Further, sense-1 and sense-2 scored higher in SWVs than PVs whereas for the sense-3, an opposite pattern was found in which PVs scored higher than SWVs. Further, there was no significant interaction effect of target type × meaning sense × PO activity, F (4, 142) = 1.834, p =.126, η_p^2 = .049.

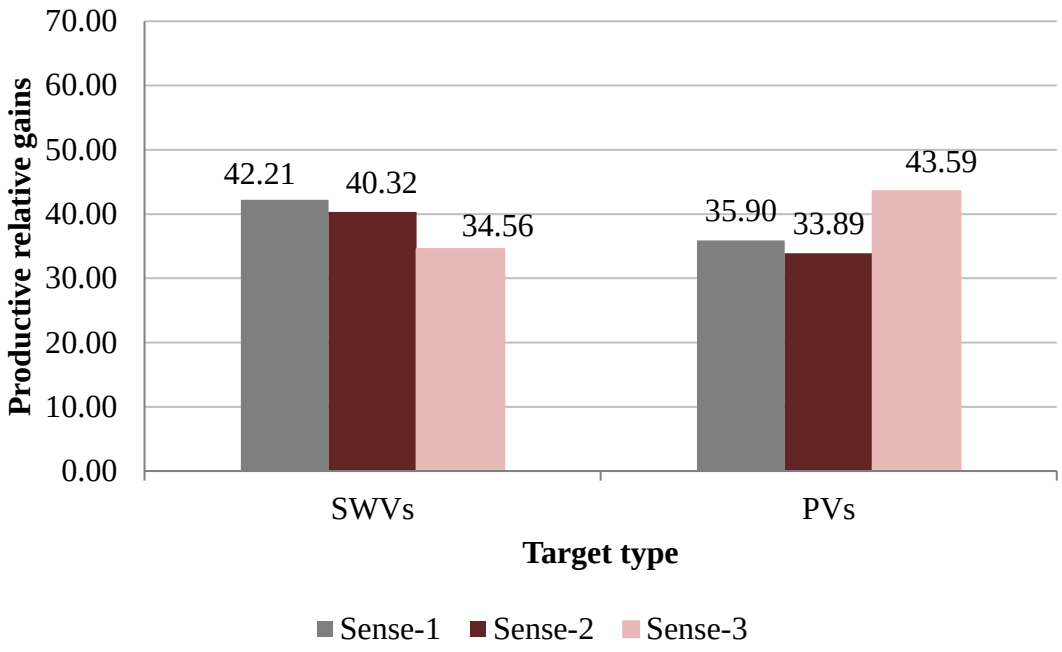


Figure 5.8. Means of productive relative gains for the three meaning senses by target type (SWVs and PVs)

5.4.5. Further analysis

The results shown in Section 5.4.1 revealed that there were no significant differences in VLT scores among the three treatment groups. However, similar to what was carried out in Study 1, a series of Spearman's rank-order correlations were carried out between the participants' level of vocabulary knowledge and their receptive and productive learning gains. The aim was to examine if the small differences in learners' VLT might have affected the receptive or the productive learning gains. The relationship between the VLT-2K, VLT-3K, the combined VLT scores (2K+ 3K), the SWVs, PVs and the combined (SWVs and PVs) receptive relative gains and the SWVs, PVs and the combined (SWVs and PVs) productive relative gains of the target items were examined. The results showed that only VLT-2K scores correlated strongly with the SWVs productive relative gains (r = .447, p = .001) and PVs productive relative gains (r = .447,

p = .005). This indicates that participants with more vocabulary knowledge at the 2K level might increase their productive vocabulary gains of SWVs and PVs to a greater extent than those with less vocabulary knowledge.

5.4.6. Summary of the results

The results of the two mixed-design ANOVAs are summarised in Table 5.14. The results revealed that the sentence reconstruction activity may be more effective than the other two activities in terms of receptive gains. The effect of target type was not significant in for receptive or productive gains. The target type × PO activity interaction effect was not significant in either ANOVA, which means that the learning of SWVs and PVs were similarly affected by the treatment they received. The effect of meaning sense was significant for receptive gains but not for productive gains; the most frequent meaning sense gains had some advantage over the other two less frequent meaning senses only at a receptive level. However, all three meaning senses were similar in terms of productive gains. The meaning sense × target type effect was significant for receptive and productive learning gains. These results indicate that learning the three meaning senses of SWVs was different from learning the three meaning senses of PVs at both the receptive level and the productive level. More precisely, on the receptive level, the gains for the three meaning senses for SWVs would lessen gradually from the most frequent meaning sense to the least frequent meaning sense, whereas for PVs, the scores of all the three meaning senses were relatively similar. On the productive level, an opposite pattern occurred in which the third, least frequent meaning sense was the highest for both SWVs and PVs; however, the difference in scores of the three meaning senses was more pronounced in PVs than SWVs. The effect of PO activity × target type × meaning sense was not significant for receptive or productive gains, which means that the three meaning

senses of SWVs and PVs were learned in a similar manner receptively and productively and that learning was not affected by the type of treatment they received.

Table 5.14 *Summary of the mixed-design ANOVAs results for receptive relative gains and productive relative gains*

	Learning gains	
	Receptive	Productive
PO activity	*** p <.001	n.s. p =.588
Target type	n.s. p =.842	n.s. p =.603
PO activity × Target type	n.s. p =.816	n.s. p =.771
Meaning sense	*** p <.001	n.s. p =.722
PO activity × Meaning sense	n.s. p =.718	n.s. p =.081
Target type × Meaning sense	*** p <.001	* p <.05
PO activity × Target type × Meaning sense	n.s. p =.177	n.s. p =.622

5.5 Discussion

This study explored the differences between the effects of three spoken PO activities on the learning of polysemous SWVs and PVs by examining the receptive and the productive relative gains. In this section, the main results of the study are discussed in response to each of the three main research questions and interpreted and discussed in light of previous research.

5.5.1. Receptive and productive learning gains in the three spoken PO activities

The first research question asked about the differences between the three spoken PO activities (sentence reconstruction, listen-and-retell meaning and picture description) on L2 learners' receptive and productive relative gains. The findings suggest that the sentence reconstruction activity was superior to the listen-and-retell meaning activity but not to the picture description activity for receptive relative gains. However, even though the results showed statistically significant difference between the activities, the results should be interpreted with caution, as the effect size was small. In terms of the productive relative gains, there were no significant differences among the three activities. It is interesting to note, however, that learners in the three PO activities were substantially able to recall the target items, even without the implementation of techniques that encourage recalling these target words (e.g., providing the first letter, giving the number of letters of the target items), as suggested by Nation and Newton (2008). Furthermore, the productive relative gains may have been higher than the ones reported in the present study but went undetected due to the study's use of the receptive pre-test scores in the calculations of the relative productive gains (see Section 3.3.7). These findings further support Swain's Output Hypothesis (1985) for the importance of output in advancing language learning in general and vocabulary learning in particular.

There is a very limited number of studies that empirically compare the effects of different spoken PO activities, as discussed in section 5.1.2. The findings of this study are in line with Ellis and He (1999) and De la Fuente's (2002) conclusion that it is generally beneficial to include spoken PO activities for receptive and productive vocabulary learning, though the two studies did not include any comparisons of different activities. Both studies used picture description activities and related the findings to the amount of dialogic interaction and negotiation encouraged by the activity. No analysis of instances

of negotiation was included in the present study, so it could not be speculated that the advantage of one activity over another is related to negotiation.

As discussed in Section 5.3.5, the three activities differed in their conceptualisation within TFA and ILH frameworks. The hypothesis was that the activity that induces higher IL according to the ILH and TFA frameworks would result in better learning gains on both the receptive and productive levels (Laufer & Hulstijn, 2001), which was evidenced in several previous studies examining written PO activities (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017). Accordingly, two patterns (conceptualising the activities within the ILH and TFA) for the same hypothesis were formed (see Figures 5.1 and 5.2). Both hypothesised patterns were partially confirmed for the receptive relative gains but were rejected for the productive relative gains. The pattern of the results was as follows:

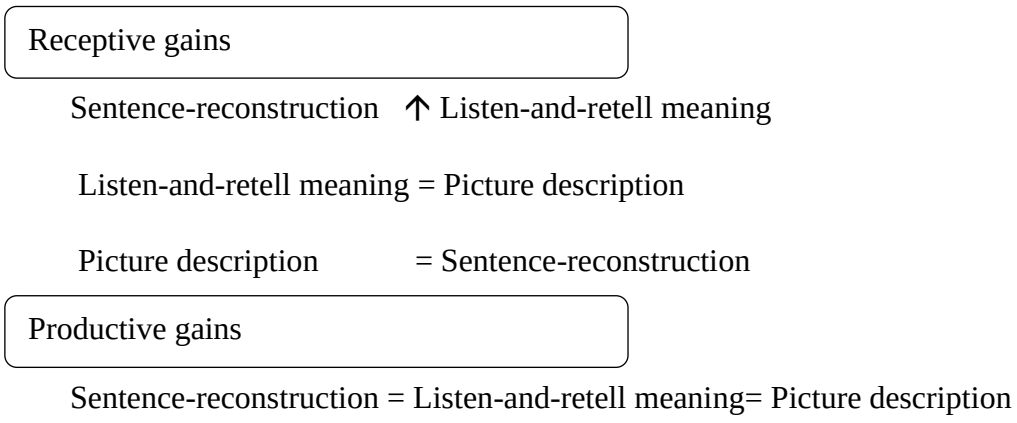


Figure 5.9. Results of pattern of the effectiveness for the three activities

In terms of the ILH, both sentence reconstruction and picture description induced similar IL indices (in total) and were expected to result in similar learning gains and the findings confirmed this pattern. However, it was also expected that the listen-and-retell

meaning would result in lower learning gains than both sentence reconstruction and picture description activities, since it induced lower ILs. The results confirmed the pattern only for the difference between the sentence reconstruction and the listen-and-retell meaning activities. The difference between the listen-and-retell meaning and the picture description was not statistically significant, which contradicts the hypothesis. One possible explanation for these variations is related to the differences within each component of the ILH between the two activities. With reference to Table 5.1, although the two activities induced a similar overall ILs, there were differences between the components. More precisely, the Search component was present only in the picture description activity, whereas the amount of Evaluation was stronger in the sentence reconstruction than in the other two activities. The sentence-reconstruction activity required another IL layer not present in the other two activities, as the participants were engaged in making comparisons between their own production and the ones given. This additional IL within the activity might have led to greater gains, though only at the receptive level.

In terms of TFA, which is considered to be a broader framework for conceptualising the depth of processing, the three activities differed in their levels. Referring to Table 5.2, the sentence reconstruction activity was expected to lead to the greatest gains according to the TFA framework, followed by picture description then listen-and-retell meaning. This pattern was partially confirmed, as the sentence reconstruction was higher than listen-and-retell meaning but not picture description. Further, there were no differences between listen-and-retell meaning and picture description activities. Again, the reason behind this variation could be related to the variations among the different components of TFA. The difference between the sentence reconstruction and listen-and-retell meaning activities can be identified in the ‘retrieval’

component. The sentence reconstruction activity required learners to retrieve the form and reconstruct the given sentence using their own words. The listen-and-retell meaning activity asked learners to retrieve the meaning and/or translation of a given form. For the picture description activity, no retrieval is required, as both form and meaning were provided, and learners had to work on linking each form to the appropriate meaning(s). Nation and Webb (2011) asserted that tasks involving productive retrievals (i.e., retrieval of forms) are more difficult than tasks with receptive retrievals (i.e., retrieval of meanings). Similarly, Laufer, Elder, Hill and Congdon (2004) argued that a task requiring the retrieval of the form of a L2 lexical item is more complicated than a task requiring the retrieving the meaning for a supplied L2 form. The reason for this difference is that L2 forms are different from L1 forms (except for cognates) and searching for an appropriate form therefore requires learners to retrieve new materials. On the other hand, semantic concepts are broadly similar across languages; retrieving the meaning of given forms requires retrieval of concepts already familiar from the L1 (Laufer et al., 2004). The results are in accordance with Laufer et al. (2004) and Nation and Webb's (2011) conclusions, as it seems that when working on retrieving the form, a more complicated retrieval process is involved. This is likely why the sentence-reconstruction activity resulted in higher receptive learning gains.

Further, the level of the TFA 'generation' (i.e., varied encounter and varied use) aspect is also different among the three activities. In both the sentence reconstruction and picture description activities, learners were required to use the target items in a way they had not encountered before, and hence both activities create higher level of generation than the listen-and-retell meaning activity. The listen-and-retell meaning activity required learners to recall the meaning of the target items they heard from their peers and is thus the only activity that did not require them to use the word in a new way, so that no generation was involved. Nation (2013) argued that encouraging the creative use of the

target items is an important condition that needs to be incorporated into the design of vocabulary activities. Listen-and-retell meaning activity was the only activity that did not incorporate any use of the target items, and, as a result, was the least effective, though only on the receptive level.

Further explanations for the differences between the three activities in the receptive learning gains may be related to the three functions of output: (1) noticing, (2) hypothesis-testing and (3) metalinguistic (reflective) talk (Swain, 1995). The third function, the metalinguistic talk, could not be anticipated, which is considered one of the limitations of the present study (see Section 5.6). However, the first two functions were operationalised differently within each of the three PO activities. Both noticing and hypothesis-testing differed in the component of word knowledge per activity. First, learners' noticing can be operated in two forms: noticing the input they receive or noticing a gap in their interlanguage that requires rectifying. These two forms might reinforce different components of word knowledge within each of the three activities. More precisely, learners in the sentence reconstruction activity may notice the meaning in the input they received and might have noticed the gap in their interlanguage to be filled by an appropriate form to better fit that meaning. On the other hand, learners in the listen-and-retell meaning group may pay selective attention to the form provided and notice the gap to be filled by the accurate meaning sense to represent that form. In the picture description activity, learners may notice both the form and the meaning provided and the gap in their interlanguage to be filled by the form-meaning(s) link between the choices provided. Hence, though the noticing function of output is incorporated within the three activities, the focus of these noticing processes varies depending on the activity. The findings support the conclusion that reconstruction activities, at the sentence or text level, lead to greater noticing of form – either noticing of the gap in their interlanguage (Swain, 1998)

or noticing of the discrepancy between their use and the target language (Doughty & Williams, 1998) – leading to more vocabulary learning than other activities. These findings are also in line with the conclusion that dictogloss has the potential to promote focus on form and thereby improve the accuracy of that form's production (e.g., Kowal & Swain, 1994, 1997; Malmqvist, 2005; Nabei, 1996; Swain, 1998; Swain & Lapkin, 2001). Dictogloss is one format of text reconstruction activity, and the present study extends this conclusion to the sentence reconstruction activity in which learners might have been pushed to produce language that was not their own and was perhaps beyond their current level.

The second function is the hypothesis testing, in which Swain (1995) advocated that learners have the opportunity to practise implementing a hypothesis about how a language works when pushed to use that language. This function, or process, differed according to the component of word knowledge learners tried out in each condition. Specifically, in the sentence reconstruction activity, learners needed to try out different forms to represent the meaning given in the replaceable items. Meanwhile, in the listen-and-retell meaning group, they were provided with the forms and needed to test their hypothesis about the most appropriate meaning in the given context. In the picture description activity, learners had the opportunity to use the given forms to describe the given pictures, a testing which might lead to forming different hypotheses about the form-meaning links. These differences in how the functions are operationalised and how each activity reinforces different components of word knowledge might have affected learning gains. When the focus was on noticing a gap in the mental lexicon involving the appropriate form and trying out different forms to represent a given meaning in a context that has not been used before, it affected the results only for receptive learning gains.

Overall, although the hypothesis that the activity which induces higher ILs and TFAs will result in greater learning gains was only partially confirmed in this study, the

results are nonetheless encouraging. Learners were able not only to manifest high levels of receptive gains in all three activities (despite the above differences), but also to use these words orally in the production test. The findings also suggest that while the ILH and TFA frameworks are useful for evaluating differences among different activities, their overall scores are not always reliable predictors of learning gains. Rather, the specific configuration of components and subcomponents appears to be more important.

5.5.2. Receptive and productive learning gains by target type (SWVs and PVs) within the three spoken PO activities

In answer to the second research question, the results displayed that, irrespective of the activity, the scores SWVs and PVs were similar in terms of receptive and productive vocabulary gains. The findings confirm the hypothesis that the learnability of SWVs and PVs is relatively similar, as well as confirming the findings of Study 1.1. The findings of the present study are in line with Alali and Schmitt's (2012) conclusion that idioms are learned at a similar rate to SWVs. However, as discussed in Chapter 3 (Section 3.5.2), the results do not support findings from previous studies showing the difficulty of FSs over SWs, such as Kasahara (2010, 2011) and Peters (2012, 2014). The reason behind this discrepancy might be attributed, again, the type of FSs examined. In both Peters (2014) and Kasahara (2010, 2011), collocations were examined (precisely known-and-unknown collocations in Kasahara studies and verb+ noun/ adj. + noun in Peters' study) whereas idioms were tested in Alali and Schmitt (2012) and high frequency phrasal verbs in the current study.

Wray (2019) questioned whether there is any significant difference between SWs and multiword units and concluded that at the level of form, FSs can be divided into two categories: ones that can be replaced by a single word and those that cannot. For all those

in the former category, the difference in learning gains may be minimal. The results of the present study are in line with this conclusion, as the FSs examined were PVs, and in most cases, those PVs could be replaced with SWVs.

As discussed in Chapter 2 (Section 2.2.3.4), previous studies have considered PVs to be problematic in EFL due to differences in their structural properties, syntactic peculiarity and semantical properties (Boers, 2000; Hulstijn & Marchena, 1989; Kao, 2001; Kurtyka, 2001; Littlemore & Low, 2006; Moon, 1998). There has also been some evidence suggesting that learners tend to avoid PVs in spoken contexts (e.g., Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004; Siyanova & Schmitt, 2007) and instead opt to use SWVs, especially for the figurative meaning senses of PVs. Though it may be argued that learners may have been conditioned to use PVs, as they were instructed to do so in the practice sessions. However, learners in the present study opted to use PVs in the oral test even when a SWV that could fit the given meaning was available. For instance, the following is one DCT from the oral test:

> *Your teacher gave the attendance sheet to the whole class to sign confirming their attendance. One of the students signed it and put it on her desk, so you did not receive it. What would you do?*

The answer in the given context would include the PV *pass on*, with the meaning of *give something to someone after receiving it from someone else*. Another possible answer with a high-frequency SWV would be *give*. However, 24 out of 49 learners in the present study used the PV in this context. The other participants (n=25) did not give any answer to the question. The possibility of using a SWV instead of PV was an option in some test items as the PVs used in the study are high frequent and could be easily substituted with a SWV. However, participants in these cases either used the target PV

or did not give an answer in most of these cases. These results suggest that learners produced a similar number of PVs and SWVs and that when PVs received the same amount of instruction as SWVs, the two types of verbs were learned at a similar rate (Adolphs & Durow, 2004; Schmitt, 2004).

Research has shown that EFL learners may become more proficient in English when they learn to use FSs (Boers et al., 2006). It is therefore crucial that the learning of FSs is fostered through instructional interventions (Nation, 2001). The present study adds to the growing body of evidence supporting this claim, which includes Dornyei's (2009) findings that teaching FSs is of crucial importance in L2 instruction. Spoken PO activities such as those investigated in this study are one appropriate and effective method to establish the form-meaning link not only on a receptive level but also on a productive level for both SWVs and PVs in classroom-based instruction, though these findings cannot be generalised to all types of FSs.

To conclude, the results of the vocabulary gains for both SWVs and PVs appears to point in one direction: the difficulty in learning PVs is not caused by their intrinsic features but rather more related to the type and amount of learners' exposure to them, as suggested by Pellicer-Sánchez (2020). Schmitt and Redwood (2011) have stated that the more frequently learners are exposed to a specific form, be it SWs or FSs, the more likely it is to be learned, whether that exposure was inside or outside the classroom. When the same type and amount of instruction is provided for both, similar results appear to be found (Pellicer-Sánchez, 2020).

5.5.3. Receptive and productive learning gains by meaning sense (1, 2 and 3) within the three spoken PO activities

The results showed that the most frequent meaning sense was better learned at the

receptive level in SWVs (SRG= 85% vs. 67% vs. 63%, LRMG= 51% vs. 32% vs. 34%, PDG=67% vs. 50% vs. 38%) and PVs (SRG= 75% vs. 67% vs. 70%, LRMG= 41% vs. 43% vs. 33%, PDG=51% vs. 58% vs. 53%). The difference was minimal at the productive level of mastery in SWVs (SRG= 48% vs. 33% vs. 24%, LRMG= 43% vs. 46% vs. 40%, PDG=36% vs. 42% vs. 40%) and PVs (SRG= 37% vs. 32% vs. 40%, LRMG= 37% vs. 38% vs. 48%, PDG=34% vs. 32% vs. 43%). The hypothesis that the most frequent meaning sense would be easier to learn than the other two meaning senses was only confirmed for the receptive learning gains, as it similarly was in Study 1.1. Though the most frequent meaning sense was the most effectively learned on the receptive level, the learning gains for the other two meaning senses can still be considered substantial.

The first, most frequent meaning senses being more readily recognised supports the "frequency effect" discussed in Chapter 3. The common observation of usage-based approaches is that language learning is mainly based on the learners' exposure to L2 in use and the input they receive. Though all meaning senses were controlled to be the same within the design of the study, it may be hypothesised that learners had been previously exposed to those more frequent meaning senses before the study. That previous exposure, although not sufficient to show in the pre-test, may have boosted the learning of the more frequent senses and shown an advantage in the receptive learning gains but not in the productive learning gains, perhaps because productive components of vocabulary knowledge tend to be more complex to acquire (Cheng & Matthews, 2018). Furthermore, the first meaning sense is typically the most literal one and therefore the easiest to learn, as in *hit* and *hold up* in Table 5.15. Among the three meaning senses for both *hit* and *hold up*, sense-1 scored higher than the other two.

Table 5.15 *Percentage of learners who responded correctly to the receptive post-test by meaning sense (1, 2 and 3) for sentence reconstruction, listen-and-retell meaning and picture description groups*

	Meaning sense	SRG	LRMG	PDG
Hit	1. to touch something quickly and with force	94%	50%	54%
	2. to affect something/someone badly	78%	39%	38%
	3. to reach a place, position, or state	83%	44%	23%
hold up	1. to raise something to a high position so it can be seen	72%	50%	77%
	2. to remain strong /in a good condition after a bad period	56%	44%	54%
	3. to delay/ prevent the progression of something/someone	61%	33%	54%

Webb and Nation (2017) suggested that the learning burden of multiple senses of a word can be gradually reduced through various methods. One method involves focusing on the core meaning sense first and then extrapolating to different meanings through some strategy or process, rather than attempting to learning each meaning sense separately. Guessing from context, for instance, is a strategy that can be applied when a familiar word is encountered in an unfamiliar sense. Further, both learners and teachers might try looking up all the senses of a word in a dictionary to work out its core meaning, consciously raising learners' awareness of the idea of having a core meaning and other meaning senses for a single lexical item. The present study suggests that the presentation and practice of multiple meaning senses of both SWVs and PVs in a single learning

session is not detrimental. In fact, there was a considerable amount of learning of the three meaning senses on both the receptive and productive levels under these conditions. These findings, with the exception of the first meaning senses being more effectively learned than the others on the receptive level, appear to support the all-at-once approach for teaching polysemous words (e.g., Csábi, 2004; Boers, 2000; Verspoor & Lowie, 2003).

Learning the multiple meanings of polysemous items in a single session is possible at least to some extent, and there appears to be no detrimental effect in presenting them all together. This conclusion is in line with what I claimed in Study 1.1 and supports findings from previous studies (e.g., Morimoto & Loewen, 2007; Khodadady & Khaghaninizhad, 2012). Though no comparisons were made between instructional styles presenting all meanings together and those presenting meanings one by one, it can be hypothesised that teaching multiple meaning senses of an item together is potentially beneficial for establishing a more comprehensive knowledge of that item. Presenting the different senses of a word together may also raise learners' awareness that any word might have multiple senses that they need to learn. Furthermore, in most cases, these meaning senses are all hinted in the most literal meaning sense of that word (Webb & Nation, 2017); for instance, the literal meaning sense of *clear* is to *remove all objects or people from a place*. The other two meaning senses, *to free of legal charges* and *to pass over without touching (as an aeroplane does)*, can be inferred from the core, literal meaning sense.

Some previous studies have reported that the learning burden of related words presented together can be greater than that of unrelated words (e.g., Higa, 1963, Tinkham, 1993, 1997; Waring, Erten, Tekin, 2008; Nation, 2000), as the interference of different meanings may confuse learners, who mix them up. However, most of these studies examined sets of words related in meaning, such as opposites or close synonyms, as opposed to several meanings for the same word. The findings of the present study imply

that learning multiple, interrelated meaning senses of one item, presented and practiced together, is plausible but do not imply that when teaching these meaning senses at different times, learning would be any different. However, as Webb and Nation (2017) recommended, raising the awareness of interference can be useful for learners, who might find it difficult to learn 'related' words together. This may also apply to meaning senses of a word. Simultaneously raising awareness of the several different meaning senses a single word may hold may be beneficial for learners who find it difficult to understand that one word can have different but related meanings.

To summarise, Study 2 empirically supports the effectiveness of the three spoken PO activities for the receptive and productive learning of polysemous SWV and PV target items. The first, most frequent meaning sense may be more readily learned than the other two meaning senses in receptive learning gains, but the productive learning gains were found to be similar for all three meaning senses.

5.6 Limitations and conclusions

The same limitations discussed for Study 1 in Chapter 3 apply to this study as well. First, it should be noted that the study did not include any immediate post-test; thus, it may be the case that some learning had been lost by the time of the delayed post-test. Future studies are recommended to employ both an immediate and delayed post-test to more accurately capture learning gains. Further, the results of the productive gains should be interpreted with caution, as they may be slightly underestimated due to the use of the receptive pre-test scores in their calculation in the absence of an oral pre-test. Further, FSs in the present study are represented through PVs only; in future studies, other types of FSs should be examined in PO instruction. Finally, it is important to note that the effect

of using static and dynamic clips as visual stimuli in the teaching materials was not investigated in the study. Future studies should look at the effect that the presentation of different types of visual stimuli may have on learning gains.

Further limitations of this study include the number of participants per group, which may have affected the pattern of results. It would be useful to replicate the study with a larger number of participants to provide more robust results. In addition, there was no control group included in the present study; hence, in future studies it would be good to have a control group completing only the pre- and post-tests with no intervention to serve as a baseline to compare and assess the effects of the three spoken PO activities.

In addition, the present study investigated only three spoken PO activities: sentence reconstruction, listen-and-retell meaning and picture description. The findings thus are limited to these three spoken PO activities and cannot be generalised to other spoken PO activities. Compared to written ones, few spoken PO activities have been examined in the literature, as discussed in Section 5.1.2. It has been found that writing exercises, such as composition writing and sentence writing, are the most effective activities for vocabulary learning (Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017). These writing exercises are characterised by their capability of inducing a higher degree of pre-planning and chunking, which results in improved memorisation of vocabulary (Zou, 2017). Therefore, it would be interesting to replicate Zou's study with speaking activities that simulate these exercises in its capability of incorporating some degree of pre-planning and chunking but in a spoken mode, such as storytelling, free conversation and sentence production.

Ultimately, the findings do not imply that one activity is clearly more effective than another. All three of the chosen spoken PO activities can be considered effective in increasing the learning of receptive and productive knowledge of vocabulary. If we look at productive learning gains, which have been shown to be more challenging for learners

(Cheng & Matthews, 2018; Henriksen, 1999; Nation, 2013), the results indicate a lack of differences among the three activities. Furthermore, the effectiveness of the chosen spoken PO activities was equal for both SWVs and PVs, which suggests that learning SWs and FSs can be improved through similar methods if the learning conditions for both are controlled to be similar. The findings also suggest that the acquiring multiple meaning senses of an item taught at the same time is possible to some extent. All of these findings come from the examination of receptive and productive learning gains in the present study. As discussed in Chapter 4, more in-depth examination of measures beyond learning gains is needed in vocabulary studies to more accurately reflect the effectiveness of vocabulary teaching approaches and techniques. This is the main focus of Study 2.2, reported in the following chapter.

Chapter 6

Study 2.2: Exploring differences between three spoken pushed-output activities: differences in lexical profiles

Study 1.2 examined the lexical profiles of utterances with SWVs and PVs produced by learners in two treatments: traditional and pushed-output. The results showed that learners in the pushed-output (PO) group performed better in all the measures of lexical profiling, i.e., overall text length, mean length of utterance (MLU) and lexical richness (including lexical diversity, lexical density and lexical sophistication), which might be expected due to a clear task effect. One question that merits further exploration, however, is whether different PO activities would lead to the same level in the lexical profiling measures (overall text length, MLU, diversity, density, sophistication) of the responses. Study 2.1 explored the effectiveness of three PO activities on the learning gains of polysemous SWVs and PVs. The results showed that the sentence reconstruction activity performed more effectively than the other two activities at the receptive level and all three activities led to similar gains at the productive level. However, we do not know whether the different types of spoken PO activities may lead to different levels of lexical profiling measures. It may be hypothesised that activities with higher ILs and TFA scores would lead to not only better gains but also to more complex use of the newly learned vocabulary. However, no empirical studies have investigated these potential differences. This chapter reports the results of Study 2.2, which investigated differences in the lexical profiling measures between three different PO activities. The data of the productive test from Study 2.1 were analysed using a range of measures of lexical profiling (i.e., overall

text length, MLU and lexical richness), similar to what was done in Study 1.2, reported in Chapter 4.

6.1. Background of the study

Several previous studies have reported that vocabulary can be learned from a range of written PO activities, such as composition writing and sentence writing (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017); cloze and fill-in-the-blank activities (e.g., Hulstijn & Laufer, 2001; Keating, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017); text reconstruction activities (e.g., Doughty & Williams, 1998; Kowal & Swain, 1997; Nassaji & Tian's, 2010; Swain, 1998; Thornbury, 2002); dictogloss (e.g., Kowal & Swain 1994; Malmqvist, 2005; Nabei, 1996; Swain, 1998; Swain & Lapkin, 2001); jigsaw (e.g., Pica, Holliday, Lewis & Morgenthaler 1989: Pica, Lincoln-Porter, Paninos & Linnell 1995; Swain & Lapkin, 2001); and sentence-combining and translation activities (Tahmasbi & Farvardin, 2017). Most of these studies used the ILH to distinguish between the PO activities and reported that the activities which induced higher ILs resulted in increased vocabulary learning and retention. However, far fewer spoken PO activities have been examined (e.g., Ellis & He, 1999; De la Fuente, 2002), and the relationship between ILH and these activities has not been explored.

Chapter 5 discussed the need for the investigation of differences between spoken PO activities on learning polysemous SWVs and PVs and presented the methodology conducted for that investigation. It also reported on a study that examined three PO activities: sentence reconstruction, listen-and-retell meaning, and picture description activities on receptive and productive learning gains of SWVs and PVs. The results partially supported the contentions of the ILH and TFA and showed that, in terms of

receptive relative gains, the sentence reconstruction activity outperformed the listen-and-retell meaning activity but was similar in effect to the picture description activity. Further, no differences were found between the listen-and-retell meaning and picture description activities in terms of receptive learning gains. On the other hand, all the three activities were similar in effect on productive learning gains. Yet, as discussed in Chapter 4, there is a clear need to investigate spoken PO activities using measures beyond learning gains. Through such examination, the development of vocabulary use could be captured more comprehensively to reflect not only the recognition and recall of form-meaning links but also the way the learned vocabulary is used in context. In view of the findings in Study 2.1, a question arose regarding the effect of these activities with regard to the lexical profile measures used in Study 1.2, i.e. text length, MLU and lexical richness (lexical density, lexical diversity and lexical sophistication). The assumption is that the different activities, with different their ILs and TFAs, might lead to different lexical profiles, since differences were found in the receptive gains. Differences in the specific features of each activity based on their conceptualisations within ILH and TFA frameworks may affect the way vocabulary are used in production. For instance, the sentence reconstruction activity was rated as inducing the highest amount of evaluation according to the ILH framework. Furthermore, the type of retrieval involved in the activities is considered productive (retrieval of the form) only in the sentence reconstruction activity, according to the TFA framework (refer to Sections 5.3.5.2 and 5.3.5.3 for more detail on the conceptualisation of the activities within the ILH and TFA frameworks). Both features could affect the way learners use vocabulary in context. However, no empirical evidence has been found to support this assumption. For this reason, Study 2.2 aims to analyse the data obtained from the oral test of the three groups through an examination of measures of lexical profiling.

Lexical profiling refers to the (indirect) examination the lexical features of spoken or written production (Kremmel & Pellicer-Sánchez, in press). There are several aspects of this examination that have been explored in the literature: lexical density, lexical diversity, lexical sophistication and proportion of errors (Laufer & Nation, 1995). Various measures of lexical features of production have been used, including the Lexical Frequency Profile (LFP; Laufer & Nation, 1995); Type-Token Ratio (TTR; Arnaud, 1984; Daller & Phelan, 2007; Daller & Xue, 2007; van Hout & Vermeer, 2007); D or VoC-D (Durán, Malvern, Richards & Chipere, 2004); MTLD (McCarthy & Jarvis, 2007); Guiraud's Index, Advanced Guiraud (Daller, van Hout & Treffers-Daller, 2003); Coh-Metrix (S. A. S. A. Crossley et al., 2011); and P-Lex (Meara & Bell, 2001). Other grammatical measures, such as MLU, have also used to examine their relationship with the aforementioned measures (e.g., Dethorne, Johnson & Loeb, 2005; J. Miller, 1991).

Among all these measures, Study 1.2 examined the following: overall text length, MLU and lexical richness represented by lexical diversity, lexical density and lexical sophistication. Overall text length refers to the overall length of all participants responses inputted together as one text. MLU refers to the average length of a learner's utterances, and it can be measured using either words (Nice, 1925b) or morphemes (Brown, 1973; De Villiers & De Villiers, 1973). Lexical richness is considered an umbrella term that refers to the lexical features of a learner's production: density, diversity and sophistication (Laufer & Nation, 1995). Lexical density (LD) is the ratio of the lexical words (i.e. verbs, nouns, adjectives, and adverbs) to the total number of tokens in a text (Ure, 1971). Lexical diversity (LDV) refers to the ratio of different words (or types) to the total amount of lexical items (or tokens) used in a text. Several measures of LDV have been used in previous studies, and among these measures, VoC-D and MTLD are the most commonly used, as the two seem to be the least affected by the length of a text (Durán et al., 2004; Malvern & Richards, 2002; P. M. McCarthy & Jarvis, 2007, 2010). Lexical sophistication

(LS) is the proportion of sophisticated or advanced words in a text (Read, 2000). Several approaches can be used to measure LS, as discussed in Section 4.1.1.3, and a number of tools for measuring LS have been developed, such as LFP, Coh-Metrix and *TAALES*. Refer to Chapter 4 for a comprehensive overview of the aforementioned measures of lexical profiling.

6.2. Research questions

Study 2.2 aimed to answer the following research questions:

1. Is there a difference between the sentence reconstruction, listen-and-retell meaning and picture description activities in terms of length measures (overall text length and MLU) and lexical richness (i.e., LDV, LD, and LS)?
2. Is there a difference between SWVs and PVs in terms of lexical profiling measures (i.e. overall text length, MLU, and LD,) within each activity?
3. Is there a difference between the three meaning senses of the target items in the overall text length and MLU scores within each activity?

No previous studies have looked at the differences between PO activities in terms of lexical profiles (i.e. overall text length, MLU and lexical richness). However, based on the superior effectiveness of text reconstruction activities in terms of quantity of vocabulary learned reported in previous studies (e.g. Doughty & Williams, 1998; Kowal & Swain, 1997; Nassaji & Tian, 2010; Swain, 1998; Thornbury, 2002) and the results presented in Chapter 5, it was hypothesised that this effectiveness advantage might be present in the lexical features of the utterances. With regards to differences by target type, the results reported in Chapter 4 showed that SWVs hold some advantage over PVs in overall text length, MLU, and LD. However, it was also observed that such variation in

the measures was somehow minimal in the PO group. Hence, given that the same amount and type of input has been shown to lead to the same gains in previous studies (e.g., Alali & Schmitt, 2012) and in the findings of Study 2.1, it was hypothesised that there might not be differences in lexical profile either. Further, it was predicted that there might be differences between the three meaning senses, in which the first meaning sense is used in longer sentences than the other two meaning senses, as in Study 1.2.

6.3. Methodology

The methodology section of this chapter briefly reviews the activities used in Study 2.1 then moves to the analysis tools and data scoring procedures. Refer to Chapter 5 (Section 5.3.5) for comprehensive explanations of the design of the activities.

6.3.1. Pushed-output activities

Activities of three different types were designed for the purpose of the study: sentence reconstruction, listen-and-retell meaning and picture description. These activities differed in several factors: the type of retrieval required to complete the activity, the input received, and their ILs and TFA scores. Table 6.1 summarises these differences.

Table 6.1 *Summary of the differences in design between the three PO activities (sentence reconstruction, listen-and-retell meaning, and picture description)*

	Sentence reconstruction	Listen-and-retell meaning	Picture description
Type of retrieval	Form	Meaning	Form-meaning link
Type of input	Include the meaning (synonym)	Include the form	Include the form (keys) and meaning (pictures)
Total IL	3 (+Need, - Search, ++ Evaluation)	2 (+Need, -Search, + Evaluation)	3 (+Need, + Search, + Evaluation)
Total TFA	13	10	11

6.3.2. Data preparation

The data for the current analysis are the transcriptions of the oral test used in Study 2.1. In the oral test, learners were asked to respond to 72 test items: 51 DCTs and 21 factual questions, as explained in Section 5.3.6.2. They spent 40 minutes completing each oral test (n=2). Their answers were recorded and transcribed in preparation for analysis. The duration of the recording was 80 minutes per learner, equal to 3,920 minutes and 20,231 words in total. The main goal of this analysis is to explore the lexical profiles of the utterances in which the target items were used correctly. Hence, prior to any analysis, all the incorrect responses and responses that did not include the target items were deleted. For more details about the design of the study, please refer to Section 5.3.

As discussed in Section 4.3.1, it must be acknowledged that the data included in the present study is short, fragmented responses inputted together as one text, which might be considered problematic for some of the measures, namely lexical diversity and lexical sophistication. Hence, a 100-words threshold was applied for the inclusion of the data in

the present study. When examining the spoken data, this threshold was only met for the combined test (SWVs and PVs together) in all the three groups. Thus, only differences in lexical diversity and sophistication between PO activities would be examined with no distinction neither between the target types nor the meaning senses. It was not possible to include analysis categorised by target type or by the meaning sense as all responses to all test items for these categories were less than 100 words, which is below the threshold. Furthermore, some participants were excluded from the listen-and-retell meaning activity due to not meeting this 100-word threshold (N=8). This approach ensures that, while the data is still coming from fragmented speech, all text included in the analyses are sufficiently long for the analyses reported.

6.3.3. Analysis tools

The tools that were used for the analysis are the same ones used in Study 1.2. These include:

- Kutools (2017) for scoring the overall text length and MLU
- Text Inspector (2017) for scoring LDV and other related measures, such as the total number of nouns, verbs, adjectives and adverbs
- TAALES software (Kyle & Crossley, 2015) for scoring LS-Frequency and LS-Range

6.3.4. Scoring and analysis

Similar procedures to those used in Study 1.2 were used in scoring and analysing each lexical profile measure: overall text length, MLU, LDV, LD, LS-Frequency, and LS-Range. The following is a brief overview of these procedures.

6.3.4.1. Length measures

Overall Text length. The length of the texts was first scored using Kutools by counting the total number of words by treatment mode, by target type, and by meaning sense. Afterwards, a mixed-design ANOVA with repeated measures was carried out with the PO activity (sentence reconstruction, listen-and-retell meaning and picture description) as the between-subjects variable, and the target type (SWVs and PVs) and meaning sense (1, 2, and 3) as the within-subjects repeated measures variable (mixed-design ANOVA-1). The assumption of sphericity was checked, and it indicated that none of the effects violated this assumption. The results of the Levene's test indicated that the assumption of homogeneity of variance was not violated for any of the meaning senses, (p > .05). Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of 0.025. Further, the effect sizes are reported according to Plonsky and Oswald (2014) guidelines: r > .25 small effect; r > .4 medium effect; r > .6 large effect.

MLU. MLU was counted in words, a standard introduced by Nice (1925) and further recommended by Hickey (1991) and Parker and Brorson (2005) for being faster, easier, more neutral and more reliable approach than other existing measures of MLU. A word in the present study is an orthographic word, with the exception of PVs, which were counted as one orthographic word to avoid bias in results by target type. MLU was calculated by dividing the total number of words by the total number of utterances. Next, a mixed-design ANOVA with repeated measures was carried out for MLU (ANOVA-2). The between-subjects variable was the PO activity (sentence reconstruction, listen-and-retell meaning, and picture description). The within-subject variables were the target item type (SWV or PV) and meaning sense (1, 2 and 3). The results of the sphericity tests indicated that this effect was violated; hence, the results of the Greenhouse-Geisser test are used in reporting these effects, as recommended by Field (2013). The results of the

Levene's test assumption of homogeneity of variance was not violated for MLU scores (*p.*>.05). Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of 0.025. The effect sizes in this chapter are estimated according to Plonsky and Oswald's (2014) guidelines as follows: r > .25 small effect, r > .4 medium effect, and r > .6 large effect.

6.3.4.2. Lexical richness

As explained in Chapter 4 (Section 4.3.3.1), lexical richness is represented by three measures: LD, LDV and LS (Laufer & Nation, 1995).

LD. LD is the proportion of lexical words (tokens) to the total number of tokens. The lexical words are nouns, adjectives, adverbs and verbs (excluding modal verbs and the auxiliary verbs *be*, *do* and *have*). The number of lexical words was calculated using *Text Inspector* (2017). Excel was then used for scoring LD. A mixed-design ANOVA with repeated measures was carried out for LD (ANOVA-3). The between-subjects variable was the PO activity (sentence reconstruction, listen-and-retell meaning, and picture description). The within-subject variables were the target item type (SWV or PV). The results of the sphericity tests indicated that this effect was not violated. The results of the Levene's test assumption of homogeneity of variance was significantly violated for LD scores (*p.*<.05). The Games-Howell post-hoc tests were then used, as suggested by Field (2013) for being the most accurate post-hoc test when equal variance is not assumed. Statistical significance was accepted for the two-way interaction effect and simple main effects at a Bonferroni-adjusted alpha level of 0.025. The effect sizes in this chapter are estimated according to Plonsky and Oswald's (2014) guidelines as follows: r > .25 small effect, r > .4 medium effect, and r > .6 large effect.

LDV. Different measures can be used for LDV, including NDW, TTR, VoC-D and MTLD. Three measures were calculated for the present analysis – TTR, VoC-D and

MTLD – but only VoC-D and MTLD were used in the inferential statistics, as these two are more frequently used and also expected to be less sensitive to the length of the text. *Text Inspector* (2017) was used for calculating the scores of the LDV measures, and the responses of each participant were inputted all together to allow individual scoring for the combined test (i.e., scores for SWVs and PVs together), with no distinction neither between the target types (SWVs and PVs) nor the three meaning senses (1, 2, an 3), as discussed in Section 6.3.2. The utterances included in the analysis for the combined responses had a minimum of 100 words. Prior to performing any test, the normality of the VoC-D and MLTD scores was checked using Kolmogorov-Smirnov tests, and the data turned out to be non-normally distributed. Hence, two Kruskal-Wallis Tests were used to investigate differences in each measure (i.e., VoC-D, and MTLD) scores between the three PO activities (sentence reconstruction, listen-and-retell meaning, and picture description). The pattern of the results was the same across the two measures; thus, only the results of the MTLD scores are reported since MTLD has been shown to be a more reliable measure to account for differences in text length (Fergadiotis et al., 2015; Treffers-Daller, 2013). Further, the effect sizes are reported according to Plonsky and Oswald (2014) guidelines: $r > .25$ small effect; $r > .4$ medium effect; $r > .6$ large effect.

LS. LS can be measured through different approaches: corpus-derived frequency counts, range, n-gram, frequency of multiword items, imageability, familiarity and concreteness (Read, 2000; Kyle & Crossley, 2015). In Study 1.2, the analysis of LS was examined using two indices: frequency and range of all words (AWs), content words (CWs) and functions words (FWs) in the utterances. The same approach to analysis is used in the present study. The chosen reference corpus is the BNC spoken corpus, and the data were analysed using the software *TAALES* (Kyle & Crossley, 2015).

Prior to performing any tests, the normality of the scores LS-Frequency and LS-Range was checked using Kolmogorov-Smirnov tests, and the data was found to be non-

normally distributed. Initially, three Kruskal-Wallis Tests were carried out for each of the categories: AWs, CWs and FWs for each measure: LS-Frequency and LS-Range. However, since the pattern of the results was the same for all the categories, only the results for AWs are reported for LS-Frequency and LS-Range. The effect sizes are reported according to Plonsky and Oswald (2014) guidelines: r > .25 small effect; r > .4 medium effect; r > .6 large effect.

6.4. Results

The presentation of the results is divided according to the two categories: length measures including overall text length and MLU, and lexical richness measures, divided into three sub-measures: LD, LDV and LS. In each of these sub-sections, the descriptive statistics are presented first, followed by the results of the inferential statistics.

6.4.1. Length measures

6.4.1.1. Overall text length

Tables 6.2 and 6.3 presents the descriptive statistics of the overall text length scores. It shows that all the three groups produced relatively similar scores within all the categories.

Table 6.2 *Descriptive statistics for the overall text length scores by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	299.39 (107.76)	265	203	616	156.28 (63.28)	144	101	344	143.11 (47.01)	129	101	272
LRMG	268.50 (94.47)	220	203	485	136.60 (52.48)	106	101	239	131.90 (46.03)	108	102	246
PDG	287.92 (180.87)	214	203	875	143.23 (89.80)	107	102	435	144.69 (91.47)	110	101	440

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Table 6.3 *Descriptive statistics for the overall text length scores of the three meaning senses (1, 2, and 3) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	SWVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	52.67 (26.06)	47	13	124	59.33 (23.87)	51	36	124	44.28 (20.39)	50	15	96
LRMG	47.60 (19.03)	49	17	77	45.30 (23.23)	37	15	88	43.70 (16.53)	38	23	74
PDG	49.62 (19.78)	46	27	111	45.85 (38.55)	36	17	167	47.77 (37.41)	39	17	157
	PVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	40.22 (18.11)	37	20	93	43.17 (22.05)	43	11	104	59.72 (19.47)	54	30	115
LRMG	45.10 (12.84)	41	27	69	34.90 (21.66)	33	11	83	51.90 (21.84)	45	26	94
PDG	43.54 (23.26)	36	19	108	47.31 (46.15)	37	10	192	53.85 (29.32)	51	24	140

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The results of the mixed-design ANOVA-1 on the overall text length scores showed that the main effect of PO activity was not statistically significant, F (2, 38) = .174, p = .84, with a small effect size (η_p^2 = 0.009). The main effect of target type was not significant, F (1, 38) = 1.911, p = .175, η_p^2 = 0.048. The interaction effect between PO activity and target type was also not significant: F (2, 38) = 1.367, p = .267, η_p^2 = 0.067. These results suggest that, regardless of the activity, there were no differences in the overall text length scores between SWV and PV utterances. Further, the main effect of meaning sense on the overall text length scores was not significant, F (2, 76) = 1.518, p = .226, η_p^2 = 0.038. Further, no statistically significant interactions were found neither between meaning sense and PO activity (F (4, 76) = 0.745, p = .565, η_p^2 = 0.038) nor between the meaning sense and the target type (F (2, 76) = 1.254, p = .248, η_p^2 = 0.018). This suggests that learners were able to produce utterances with similar lengths for any of the meaning senses regardless of the activity and type of target item. The three-way interaction effect between PO activity, target type and meaning sense was not statistically significant, F (4, 76) = 2.134, p = .085, η_p^2 = 0.101.

6.4.1.2. Mean length of utterances (MLU)

Table 6.4 presents samples of learners' responses to the oral test by PO activity (sentence reconstruction, listen-and-retell meaning and picture description), by target type (SWVs and PVs) and by meaning sense (Senses 1, 2, and 3).

Table 6.4 *Sample of learners' responses in the oral test for the three meaning senses of a SWV (train) and a PV (hold up) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description)*

	TRAIN (SWV)		
	SRG	LRMG	PDG
Sense-1: Teach someone how to do something	The university will train the students to get the new computer system	The college train the students how to use computers	The university train students and teachers
Sense-2: Practise a sport or exercise	The school in Jeddah will train the kids for the upcoming game or match	They will train the team to win the game	They train the team
Sense-3: Point or aim something	They will train the lights on the bride	Train lights on bride	Train the lights
	HOLD UP (PV)		
	SRG	LRMG	PDG
Sense-1: Hold something in a high position	I will raise or hold up my hand to answer the question	I will hold up my hand to answer	I will hold up my hand
Sense-2: Remain strong or in good condition	I wish the strong storm settle down and the tent will hold up	I hope the tent hold up	I want the tent to hold up
Sense-3: Delay	sorry I am late because accident hold up me	I will tell them there was accident hold me up	I will say I am late because the accident hold up the street

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The descriptive statistics of MLU are presented in Tables 6.5 and 6.6. The results show that higher means occurred for the sentence reconstruction group over the other two groups. For differences by target type, learners within each group produced utterances of relatively similar length for both SWVs and PVs. Little variation was found in MLU for each meaning sense. Referring to Table 6.2, the sample shows that one learner in the sentence reconstruction group produced relatively longer utterances than two learners in the other two groups. We can see also that same learner producing utterances with relatively similar length for the three meaning senses; however, utterances might be considered shorter for PVs compared to SWVs.

Table 6.5 *Descriptive statistics for MLU scores by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	7.69 (2.99)	6.82	4.70	15.00	7.56 (2.97)	6.56	4.61	14.88	7.81 (3.25)	7.01	4.56	16.00
LRMG	4.15 (2.55)	3.92	0.00	9.15	4.15 (2.16)	3.55	1.75	8.85	4.72 (2.29)	4.25	1.62	9.46
PDG	6.50 (3.31)	6.59	2.46	15.09	6.45 (3.18)	6.78	2.56	14.50	6.55 (3.48)	6.39	2.25	15.71

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

Table 6.6 *Descriptive statistics for MLU scores of the three meaning senses (1, 2 and 3) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Sense-1				Sense-2				Sense-3			
	SWVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	7.61 (2.90)	7.23	4.00	14.00	7.72 (3.79)	6.20	3.67	16.83	7.38 (3.07)	6.22	4.00	15.00
LRMG	3.71 (2.49)	3.67	0.00	8.56	3.76 (2.59)	3.00	0.00	8.80	3.87 (2.71)	3.50	0.00	9.25
PDG	6.50 (3.18)	6.67	2.57	13.88	6.22 (3.34)	5.67	2.50	15.18	6.55 (3.41)	6.27	2.60	14.27
	PVs											
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	7.02 (2.76)	6.20	4.00	13.29	7.75 (3.76)	7.23	3.00	17.33	8.41 (3.96)	6.90	4.44	18.75
LRMG	3.91 (2.82)	3.67	0.00	9.86	3.83 (3.17)	3.43	0.00	10.33	4.55 (3.15)	4.50	0.00	9.40
PDG	6.45 (3.00)	5.67	2.50	13.50	6.76 (4.08)	6.00	2.33	17.45	6.37 (3.66)	6.63	2.00	15.56

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The results of the mixed-design ANOVA-2 showed that the main effect of PO activity was significant, F (2, 46) = 7.306, $p < .01$, with a small effect size ($\eta_p^2 = 0.241$). The post-hoc tests revealed that the mean difference in MLU scores between the sentence reconstruction group and the listen-and-retell meaning group was significant (3.712, 95% CI [1.39, 6.03], $p = .001$). However, differences were found neither between the sentence reconstruction or the picture description groups (1.174, 95% CI [-1.72, 4.07], $p = .577$)

nor between the listen-and-retell meaning and the picture description groups, (2.538, 95% CI [-0.23, 5.30], p = .076), as shown in Figure 6.1.

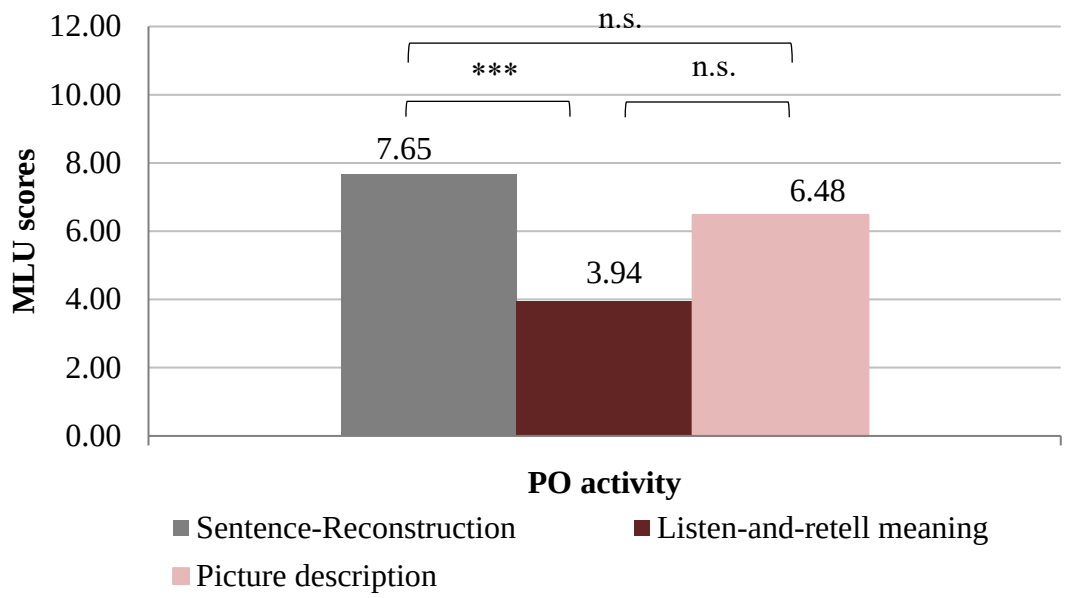


Figure 6.1 Means of MLU scores by PO activity (sentence reconstruction, listen-and-retell meaning and picture description)

The main effect of target type was not significant, F (1, 46) = 1.022, p = .317, η_p^2 = 0.022. The interaction effect between PO activity and target type was also not significant, F (2, 46) = .117, p = .890, η_p^2 = 0.005. These results suggest that, regardless of the activity, there were no differences in MLU between SWV and PV utterances. Further, the main effect of meaning sense on MLU scores was not significant, F (2, 92) = 1.580, p = .211, η_p^2 = 0.033. There were no statistically significant interactions neither between meaning sense and PO activity (F (4, 92) = 0.667, p = .616, η_p^2 = 0.028) nor between the meaning sense and the target type (F (2, 92) = 1.337, p = .268, η_p^2 = 0.028). This suggests that learners were able to produce utterances with similar lengths for any of the meaning senses regardless of the activity and type of target item. The three-way interaction effect

between PO activity, target type and meaning sense was not statistically significant: F (4, 92) = 1.159, p = .334, η_p^2 = 0.048.

6.4.2. Lexical richness

6.4.2.1. Lexical density (LD)

The descriptive statistics for LD are presented in Table 6.7. There are considerable variations in the scores by PO activity. The scores of the sentence reconstruction group were the highest, followed by the picture description group and then the listen-and-retell meaning group. For differences by target type, it appears that utterances with SWVs scored higher than utterances with PVs.

Table 6.7 *Descriptive statistics for LD of SWVs PVs and combined by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (%) (SD presented in brackets)*

	Combined				SWVs				PVs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	54.78 (12.7)	50.55	41.36	84.38	61.77 (13.1)	62.29	42.86	84.62	49.04 (15.4)	43.58	20.00	83.33
LRMG	43.79 (4.2)	43.62	37.07	54.81	46.39 (4.9)	45.99	37.40	57.41	40.65 (4.1)	41.04	33.33	52.00
PDG	51.05 (9.2)	49.26	36.50	67.16	54.43 (10.9)	52.03	38.72	73.77	47.23 (6.9)	49.57	34.45	57.69

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The results of the mixed-design ANOVA-3 revealed that the main effect of PO activity on LD scores was statistically significant, F (2, 46) = 7.604, p < .01, with a small

effect size (η_p^2 = 0.248). The post hoc analyses revealed that the difference between the listen-and-retell meaning and sentence reconstruction groups was significant (11.885, 95% CI [4.26, 19.51], p = .001), whereas for the other two group comparisons, the mean differences were not statistically significant; i.e., between the listen-and-retell meaning group and the picture description group (4.575, 95% CI [-3.75, 12.90], p = .537) and between the sentence reconstruction group and the picture description group (-7.310, 95% CI [-15.64, 1.02], p = .103), as shown in Figure 6.2. These results indicate that the sentence reconstruction group had some advantages over the listen-and-retell meaning group in terms of LD scores.

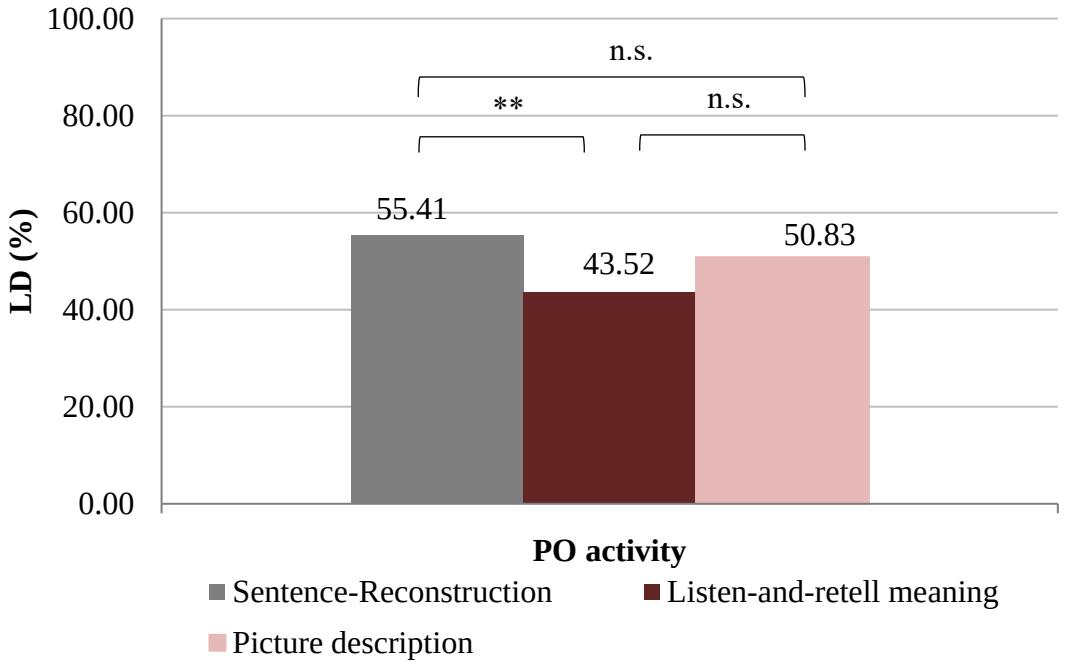


Figure 6.2 Mean LD scores by PO activity (sentence reconstruction, listen-and-retell meaning and picture description)

The main effect of target type on LD scores was significant, F (1, 46) = 44.084, p < .001, η_p^2 = 0.489, a medium effect size. The follow-up contrasts revealed that utterances with SWVs were lexically denser than utterances with PVs. The interaction between PO

activity and target type was not statistically significant, F (2, 46) = 3.003, p = .059, η_p^2 = 0.115. This suggests that utterances with SWVs were lexically denser than those with PVs in any of the treatment groups, as shown in Figure 6.3.

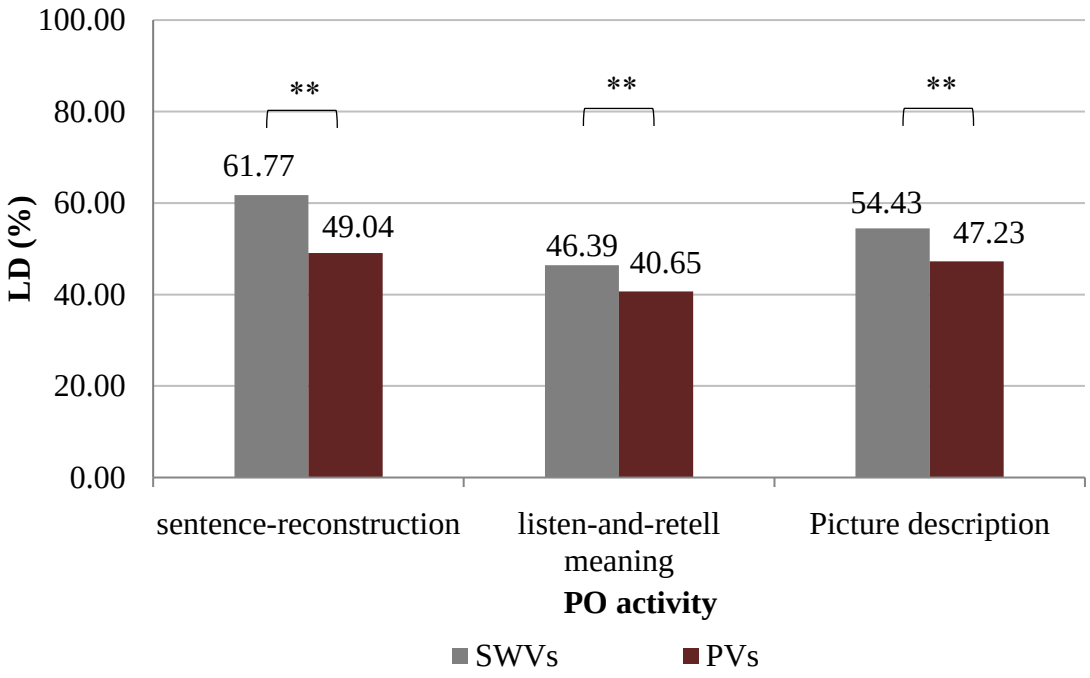


Figure 6.3 Means of LD scores for SWVs and PVs by PO activity (sentence reconstruction, listen-and-retell meaning and picture description)

6.4.2.2. Lexical diversity (LDV)

Table 6.8 presents the descriptive statistics for the three measures of LDV (TTR, VoC-D and MTLD). The results showed that the scores of LDV in all three measures were relatively similar across the groups (sentence reconstruction, listen-and-retell meaning and picture description).

Table 6.8 *Descriptive statistics for LDV measures by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	Combined											
	TTR				VoC-D				MTLD			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	1.36 (3.76)	0.46	0.36	16.43	54.05 (13.48)	53.9	27.98	85.24	54.09 (13.94)	54.16	25.82	82.80
LRMG	0.65 (0.19)	0.64	0.37	1.00	56.76 (20.70)	55.3	23.86	97.93	51.89 (16.27)	53.00	25.11	76.66
PDG	1.01 (1.80)	0.51	0.30	7.00	59.70 (23.60)	61.58	25.57	107.87	56.38 (20.88)	52.72	26.76	85.78

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The results of the Kruskal-Wallis Test showed that the difference in MTLD scores between the three PO activities was not statistically significant, X^2 (2, n=45) = .433, p = .89. These results indicate that learners in the three PO activities were able to produce utterances with a similar level of lexical diversity.

6.4.2.3. Lexical sophistication – Frequency (LS-Frequency)

The descriptive statistics for LS-Frequency are presented for AWs, CWs and FWs in Table 6.9. The differences between the three groups were very minimal.

Table 6.9 *The frequency scores of the utterances based on BNC spoken corpus (AWs, CWs and FWs) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	AWs				CWs				FWs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	9.12 (1.03)	9.15	6.60	10.85	1.13 (0.21)	1.15	0.81	1.63	19.19 (1.90)	19.08	16.01	22.63
LRMG	7.32 (2.55)	8.55	0.95	9.79	0.78 (0.39)	0.79	0.09	1.56	18.58 (5.84)	18.18	2.71	28.33
PDG	8.43 (2.33)	9.50	4.23	11.16	0.94 (0.45)	0.82	0.45	2.24	18.85 (3.65)	19.12	12.82	23.98

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The results of the Kruskal-Wallis Test showed that there was not a statistically significant difference in LS-Frequency scores among the three PO activities, X^2 (2, *n*=49) = 4.960, *p* = .084.These results indicate that learners completing any of the three PO activities produced utterances with similar LS-Frequency scores.

6.4.2.4. Lexical sophistication- Range (LS-Range)

The descriptive statistics for LS-Range scores are presented in Table 6.10 for AWs, CWs and FWs. The range scores of the sentence reconstruction group were higher than those of the other two groups.

Table 6.10 *The LS-Range scores of the utterances based on BNC spoken corpus (AWs, CWs and FWs) by PO activity (sentence reconstruction, listen-and-retell meaning and picture description) (SD presented in brackets)*

	AWs				CWs				FWs			
	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.	M (SD)	Md.	Min.	Max.
SRG	76.19 (2.16)	76.06	71.49	80.56	59.77 (2.63)	59.23	56.03	65.89	97.01 (1.14)	96.87	94.47	98.77
LRMG	70.44 (6.88)	71.83	54.29	80.39	55.57 (7.48)	57.23	38.89	66.26	96.32 (1.83)	96.53	91.33	99.18
PDG	72.89 (4.56)	74.11	64.41	80.19	56.41 (4.92)	55.38	50.58	68.05	96.86 (0.83)	96.76	95.44	98.23

Note: SRG = sentence-reconstruction group, LRMG= listen-and-retell meaning group, PDG = picture description group.

The results of the Kruskal-Wallis Test showed that there was a statistically significant difference in LS-Range scores among the three PO activities, X^2 (2, *n*=49) = 10.940, $p < .01$, with a small effect size (*r*. = 0.013). This finding was followed up with multiple Mann-Whitney U tests by group. The results revealed that the difference was significant between the sentence reconstruction and listen-and-retell meaning groups (U = 64.000, p = .002), also between the sentence reconstruction and picture description groups (U = 58.000, p = .018), but not between the listen-and-retell meaning and picture description groups (U = 101.00, p = .522), as shown in Figure 6.4.

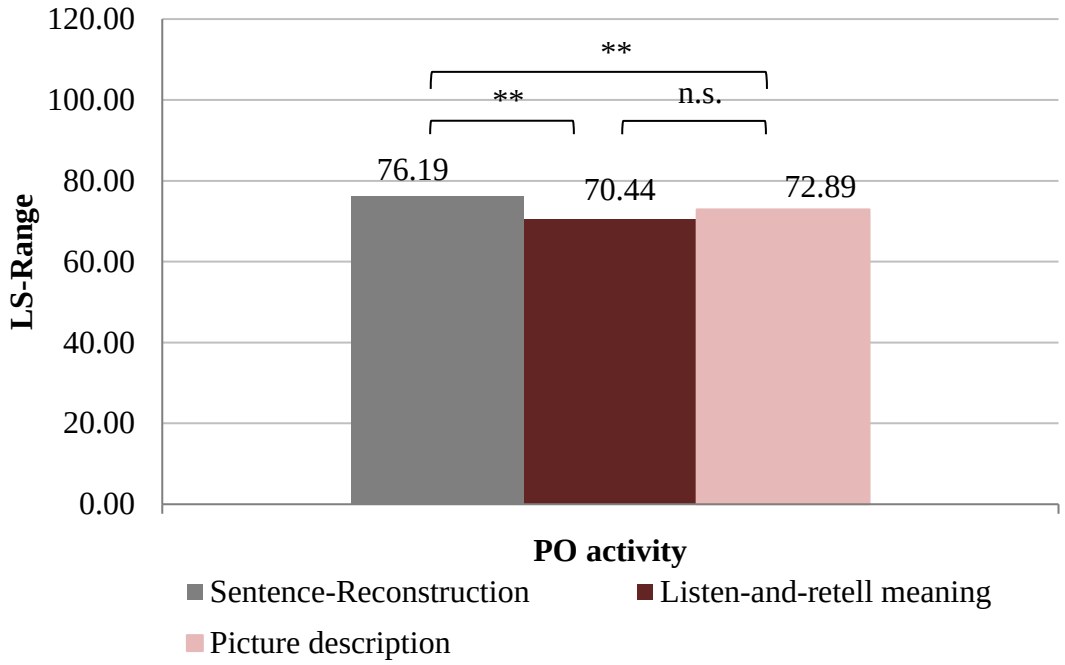


Figure 6.4 Means of LS-Range scores by PO activity (sentence reconstruction, listen-and-retell meaning and picture description)

6.4.3. Further analysis

Several studies reported strong correlations between MLU and some measures of lexical profiling (e.g. Dethorne, Johnson & Loeb, 2005; Miller, 1991; Ukrainetz & Blomquist, 2002), as well as between lexical richness and vocabulary size of learners (e.g. Daller, Van Hout & Treffers-Daller, 2003; Durán et al., 2004; Engber, 1995; Jarvis, 2002; Laufer et al., 1995; Yu, 2009). Study 1.2 reported strong correlations between TTR and LD and LS measures (both Frequency and Range); LD and LS (both Frequency and Range); MLU and LD, LS and correlates negatively with LDV (only TTR). Within LDV measures, VoC-D correlates highly with only MTLD. However, the vocabulary levels test (VLT) scores of the two levels, 2K and 3K, did not correlate with any of the measures of lexical profiling. Following the same approach in Study 1.2, a series of Spearman's

correlations were carried out, as the data were non-normally distributed, to examine the relationship between the measures of lexical profiling and VLT. Table 6.11 presents the results and shows that VoC-D correlates with all the measures of lexical profiling (LD, LS-Frequency, LS-Range and MLU); MTLD correlates only with LS-Frequency; LD correlates with LS (both Frequency and Range) and MLU; LS-Frequency correlates with LS-Range and MLU; and LS-Range correlates with MLU. Within the three measures of LDV, a strong correlation was found between VoC-D and MTLD, similar to the results reported in Study 1.2. However, none of the measures used in the present study correlates with either of the two levels of VLT, similar to what was found in Study 1.2. Learners, whether having high or low vocabulary knowledge levels, have no relationship with their ability to produce longer and more lexically-rich utterances. As discussed in Section 4.4.3, the fact that all the materials that learners were exposed to (short lessons, activities, and target items) were mostly designed to be of the 2K high frequent words, would explain the lack of correlation between the two. Besides, as reported by Uchihara and Clenton (2018), if the purpose of the conversation is to be comprehensible to listeners, then it is less likely learners would be using vocabulary beyond the most frequent 2,000 words.

Table 6.11 *Correlations between lexical richness measures (LDV, LD and LS), MLU and VLT (2K and 3K) for the three PO activities*

		LDV			LD	LS		VLT	
		TTR	VoC-D	MTLD		Freq.	Range	2K	3K
LDV	TTR								
	VoC-D	-0.134							
	MTLD	-0.030	**.887****						
LD		0.053	**.437****	0.288					
LS	Freq.	0.036	**-.624****	**-.420****	**-.695****				
	Range	0.038	**-.331***	-0.201	**-.826****	**.720****			
VLT	2K	-0.082	-0.009	-0.019	-0.058	0.135	0.021		
	3K	0.005	0.185	0.119	0.209	-0.141	-0.239	0.048	
MLU		-0.078	**-.315***	-0.257	**-.692****	**.613****	**.726****	-0.019	-0.027

6.4.4. Summary of the results

The results of the present study are summarised in Table 6.12. With regards to differences by PO activity, the results showed that the sentence reconstruction activity holds some advantages over the listen-and-retell meaning activity in terms of MLU, LD and LS-Range. The effect of target item type was only investigated in three measures: overall text length, MLU, and LD. The scores for SWVs were higher than PVs only in LD scores, but no significant differences were found neither in the overall text length nor the MLU scores. For none of the three measures was the target type × PO activity interaction effect found to be significant. The meaning sense effect was not significant for both the overall text length or the MLU scores, which indicates that all the three meaning senses were similar in terms of their overall text length and MLU scores. Also,

the meaning sense was not affected by the PO activity or the target type, nor the PO activity × Target type × Meaning sense interaction effect was significant in either measure: overall text length and MLU.

Table 6.12 *Summary of the results for overall text length, MLU, LDV, LD, LS-Frequency and LS-Range*

	Text length	**MLU**	**LD**	**LDV**		**LS**	
				VoC-D	**MTLD**	**Freq.**	**Range**
PO activity	n.s. p =.841	** p <.01	** p <.01	n.s. p = .552	n.s. p =.784	n.s. p =.039	** p <.01
Target type	n.s. p =.175	n.s. p =.317	*** p <.001				
PO activity × Target type	n.s. p =.267	n.s. p =.890	n.s. p =.059				
Meaning sense	n.s. p =.226	n.s. p =.213					
PO activity × Meaning sense	n.s. p =.565	n.s. p =.609					
Target type × Meaning sense	n.s. p =.248	n.s. p =.266					
PO activity × Target type × Meaning sense	n.s. p =.085	n.s. p =.333					

6.5. Discussion

This study has explored the differences between three spoken PO activities in the lexical profiling of utterances which included target polysemous SWVs and PVs, using several measures such as overall text length, MLU, LDV, LD and LS. The main results of the study are discussed in response to each of the three main research questions and interpreted in light of previous research in this section.

6.5.1. Lexical profiles in the three spoken PO activities (sentence reconstruction, listen-and-retell-meaning and picture description)

The findings of the present study suggest that the effect of the three PO activities is not consistent. The sentence reconstruction activity performed more strongly than the listen-and-retell activity in MLU, LD, and LS-Range. The sentence reconstruction group carried an effect size of η_p^2 =0.241 in MLU, η_p^2 =0.248 in LD, and r=0.013 in LS-Range, which are all considered small effects (Plonsky & Oswald, 2014). However, differences were found neither between the picture description and sentence-reconstruction nor between the picture description and listen-and-retell meaning activities. Furthermore, the three activities were all similar in effect in terms of LDV (both VoC-D and MTLD) and LS-Frequency. The advantage reported for the sentence reconstruction activity over the listen-and-retell meaning activity for the receptive relative gains in Study 2.1 (reported in Chapter 5) also was found in some of the measures of lexical profiling, namely MLU, LD and LS-Range.

The hypothesis was that the activity which induces more IL and TFA would result in better performance on all the measures of lexical profiling. Two patterns of activity were formed according to this hypothesis. First, according to the ILH framework, the expected pattern was that the sentence-reconstruction would be similar in effect to the

picture description activity and that both activities would result in better performance than the listen-and-retell meaning activities. This pattern was partially confirmed, as the sentence reconstruction resulted in similar scores to the picture description in all measures. However, only the sentence reconstruction performed better than the listen-and-retell meaning activity in three measures: MLU, LD LS-Range. The difference between the sentence-reconstruction and listen-and-retell meaning activities in terms of their conceptualisation within the ILH framework is mainly in the amount of evaluation (refer to Table 5.1). The former induced a higher amount of evaluation, as it not only required learners to use the target items but also encouraged comparisons between the learners' production and the given sentences. For the other measures (i.e. LDV and LS-Frequency), the pattern was not confirmed as all three activities resulted in similar performance, which is similar to the pattern found for the productive relative gains reported in Study 2.1.

Second, the pattern of the results according to the TFA framework was expected to be as follows: the sentence reconstruction activity would outperform the picture description and listen-and-retell meaning activities, respectively. However, the results showed that this pattern was partially confirmed between the sentence reconstruction and listen-and-retell meaning activities but not in all the other comparisons. The difference between the two activities is in the retrieval component, as the listen-and-retell meaning activity involves retrieval of the meaning (receptive retrieval), whereas the sentence reconstruction activity required the retrieval of the form (productive retrieval) (refer to Table 5.2). Given that productive retrieval is more demanding than receptive retrieval, as demonstrated by previous studies (e.g., Laufer, Elder, Hill & Congdon, 2004; Nation & Webb, 2011), this advantage of the sentence reconstruction activity may have led to improving the performance of learners in some of the measures. Further, the amount of generation was higher in the sentence reconstruction activity than the listen-and-retell

meaning activity, which also may be related to the pattern of the results. Learners in the sentence reconstruction activity had the opportunity to use the target items in sentences, which was not required in the listen-and-retell meaning activity; hence, their responses in the oral test were better on some of the measures, as they had had more practice using the target items in the learning phase. Accordingly, the TAP theory supports the idea that the process in which the learning occurs affects the retrieval of the information learned. The input-acquisition relationship proposed by Ellis and Collins (2009) similarly provides evidence for the same notion that better learning, or performance, would occur when the mode in which the target items ware learned might affect the testing of these target items. The findings are consistent with both, and this is represented in the present study by a stronger performance in terms of the lexical features of the utterances.

Additionally, Nation (2013) encouraged incorporating activities that require more creative use of the target items to enhance their acquisition. Both the sentence reconstruction and picture description activities required learners to use the target items in context, whereas the listen-and-retell meaning activity did not. It is possible that if the listen-and-retell meaning activity required leaners not only to recall the meaning but also to use the given target items in new sentences, there would then be no differences among the activities in measures of lexical profiling.

When checking the exact types produced by each group, it was found that learners in the sentence reconstruction group used a wider variety of vocabulary in their production, such as *annoying, beg, blamed, coach, compare, demand, embarrassing, pity, scream, staff, toys, chill, nap, needy, cheating, cheerful, crashed, failed, prefer, replace, roof, tricks, clarify, device, disagree, respond, absent, baseball, congratulate, perfumes* and *tidy.* Learners in the picture description group used words such as *encourage, gift, practice, released, upset, jail, resolve, correct, encourage, fool, fortune, ignore, repeat, confidence* and *solve*. On the other hand, learners in the listen-and-retell meaning group

produced a smaller variety compared to the other two groups, such as *connect, direct, exercise, flow, message, shade, decrease, claim, create, interrupted, search* and *delay*. It can be concluded that using the sentence reconstruction activity in the learning phase promoted the ability of learners to use a richer vocabulary in their L2 production.

Another possible explanation is related to the differences among the activities in terms of the degree of re-planning and the use of the chunking strategy discussed in Section 5.5.1. The sentence reconstruction activity allowed learners to practise using the items in their mental lexicon in the re-planning stage, whereas no such process would occur in the listen-and-retell meaning activity. Further, the learners doing sentence reconstruction might have been using their own strategies to reconstruct a given meaning and chunking the given information and words into original sentences to a higher degree than those doing the other two activities. This procedure resulted in stronger performance not in the actual recall of the target items but in the way these target items were used in the testing phase, which might be considered their third time practising the target items.

Previous studies reported that EFL learners' vocabulary use is linked to their vocabulary size (e.g., Bardel & Gudmundson, 2018; Daller et al., 2003; Durán et al., 2004; Engber, 1995; Jarvis, 2002; Laufer & Nation, 1995; Zareva et al., 2005; Yu, 2009). The findings of the correlation analyses contradict this conclusion, as no correlations were found between any of the lexical profiling measures and learners' VLT scores, which is similar to the findings of Study 1.2. It is possible that the lack of variation among learners in their VLT scores is related to the lack of relationship between VLT and all the measures of lexical profiling.

Overall, the results are in favour of the activities that require use of target items in original sentences, such as the sentence-reconstruction and the picture description activities, for improving the lexical profiles of utterances produced by learners for either SWVs or PVs. One final remark to be added is that the measures of lexical profiling serve

as a more systematic and objective marking scheme for the spoken and written production of learners and can help assess the vocabulary size that is available to them for use (James Milton, 2009). They can assess the learners' ability to use and produce a wider variety of vocabulary in their speaking or writing and how well these productions are. However, some of these measures seem to behave unpredictably and may be insensitive to changes caused by the type of instructions. This possibility will be further discussed in the general conclusion chapter (Chapter 7).

6.5.2. Lexical profiles by target type (SWVs and PVs) within the three spoken PO activities

The second research question in the present study is concerned with the differences in the lexical profile measures (only overall text length, MLU, and LD) by target types: SWVs and PVs. The results indicate that differences do exist between SWVs and PVs in terms of LD, with SWVs utterances scoring higher in LD. The findings also show that no significant differences between SWVs and PVs existed in terms of the overall text length and MLU scores.

The hypothesis that utterances with SWVs will show the same lexical profile levels as those with PVs was confirmed only for the length measures: overall text length and MLU. In Study 1.2, the results showed that the main effect of target item type was significant in all the three measures. The results of the present study are partially consistent with the results of Study 1.2. The main effect of target item type was significant in only LD scores. It must be acknowledged that these results are rather limited due to using only three measures, namely overall text length, MLU, and LD.

The results of the present study support the findings of previous studies reporting the difficulty of learning FSs compared to SWVs (e.g., Peters, 2014; Kasahara, 2010, 2011). This additional difficulty was not shown in the actual recall scores of PVs but in

the way these PVs were used in context when compared to SWVs. This further supports the need to employ such analysis, which moves beyond examining the learning gains to more fully capture the actual differences among target item types.

There are several explanations for the findings suggesting the superiority of SWVs over PVs in some of measures of lexical profiling. First, the frequency of exposure to the two target types may affect the results. Although the exposure in the study was controlled to be the same for the two target types, learners may have been exposed to SWVs more than PVs in their normal classroom materials. PVs are not frequently used or rarely taught in the Saudi context, and most English language learning materials do not include specific instruction on PVs (refer to Section 1.2 for more details about the Saudi context).

As suggested in previous studies, another explanation could be the tendency among learners to avoid using PVs (e.g., Dagut & Laufer, 1985; Hulstijn & Marchena, 1989; Laufer & Eliasson, 1993; Liao & Fukuya, 2004; Siyanova & Schmitt, 2007). This was not shown in the actual use of PVs in the present study, as no differences were found in the recall of PVs but was found in the context in which these PVs were used. It is possible that learners felt insecure when using PVs in context, leading them to use them in simpler sentences.

All in all, it can be hypothesised that it might take more time and exposures for learners to be able to use PVs in longer utterances. The findings imply that the differences between SWVs and PVs, and between SWs and multiword units in general, do exist in various aspects of vocabulary use. One final remark here is that the differences between SWVs and FSs should be further investigated through measures beyond the learning gains.

6.5.3. Lexical profiles by meaning sense (1, 2 and 3) within the three spoken PO activities

The results reported in Study 2.1 revealed that receptive relative gains were higher for the first and most frequent meaning sense, while productive relative gains were the same for all three meaning senses. The results of the present study were consistent with the results of the productive relative gains analysis, as no differences were found between the three meaning senses in terms of the overall text length and MLU scores. It was hypothesised that the first, most frequent meaning sense would be used in longer utterances, but this result was not confirmed in the present study. Study 1.2 showed that utterances using the target item in its first, most frequent meaning sense were significantly longer than utterances using the other two meaning senses. However, the difference was only significant for the traditional group, whereas learners in the pushed-output treatment produce utterances with the same length for all three meaning senses. The present study's findings support the same conclusion, as all the three treatment groups were PO activities.

These findings support the notion that teaching polysemous words at one exposure is plausible. Several teaching approaches have been investigated, such as the image-schema-based instruction and translation-based instruction (e.g., Morimoto & Loewen, 2007; Khodadady & Khaghaninizhad, 2012) and guessing from context (Verspoor & Lowie, 2003). All of these studies used learning gains in their examination. The present study confirms the same conclusion for spoken PO instruction, not only for the learning gains but also for the length of utterances learners produced for each meaning sense.

The findings further support the all-at-once approach for teaching polysemous words (e.g., Csábi, 2004; Boers, 2004; Verspoor & Lowie, 2003) and provide evidence for the benefits of such an approach in leading to the production of longer utterances. The all-at-once approach may help in consciously raising learners' awareness of the idea that one lexical item can have multiple meaning senses that they need to learn and use.

Learners benefit from creating associations between the core meaning of a word and its more figurative, secondary meaning senses. Boers (2004) stated that having a sense of the core literal meaning encourages learners to infer the other peripheral meaning senses on its basis. Webb and Nation (2017) demonstrated that in most cases, the multiple meaning senses of a word can be signalled by the literal meaning sense of that word; hence, their learning is most efficient and effective when done together. Accordingly, it can be inferred that when the three items were presented and practised together, some aspects of their vocabulary use were also improved. The present study provides evidence only for the MLU, but perhaps it would be the case for other measures of lexical profiles. This needs to be further explored in future research.

The findings contradict the conclusion that polysemous words should be taught in a piecemeal fashion, presenting the most literal meaning of these words first, followed by other. peripheral meaning senses taught individually over a period of time (Shortall, 2002; Webb & Nation, 2017). It has been claimed that the learning burden of word multiple senses can be reduced if the focus of the teaching and learning is on the core meaning sense first, then later dealing with the other senses as a matter of process than as matter of learning additional items, using strategies such as guessing from context, using dictionaries, etc. (Webb & Nation, 2017). However, if those meaning senses were sufficiently interrelated on semantic grounds that one sense could be inferred from the other, these meaning senses may be better learned together, as it allows for better connection in the mental lexicon and better memorisation. This was shown by the results of the productive learning gains in the present study, whose findings showed that these polysemous words, be they SWVs or PVs, with their multiple meaning senses, were used in utterances of more or less the same length. It is important to note that the study does not imply that the piecemeal fashion of learning would be less effective than the all-at-

once approach. Future research should compare the presentation and practice of the three meaning senses together and separately.

6.6. Limitations and conclusions

There are several limitations of the present study. First, as discussed in Chapter 4 (Section 4.6), the approach of inputting all the data together as one text overlooks the fact that this data set is comprised of responses to a number of different test items/stimuli and therefore not thematically well connected. Furthermore, since the data can be considered fairly short productions, the results need to be interpreted with caution. The aim was to investigate the overall improvement of lexical production after instruction using one of three activities, not only through the examination of learning gains but also through other measures of lexical profiling. Future research should consider examining more thematically connected texts of greater length, such as monologues.

Second, the analysis in the present study examined different measures of lexical profiling but did not account for the grammatical or syntactical accuracy of the full constructions. The reason behind this is that the focus of the present study (and of the entire thesis) is examining vocabulary learning and vocabulary use. In future PO studies, multidimensional analysis of the spoken data would provide a more comprehensive picture of the extent to which the instruction provided in the present study improved the learners' performance in areas beyond vocabulary.

Third, for reasons of feasibility, the present study did not investigate the amount of negotiation or the number of Language Related Episodes (LREs) produced by participants while completing the activities. One of the functions of output in language learning is the reflective (meta-talking) function, which can be measured through the analysis of instances of negotiation (De la Fuente, 2006; De la Fuente, 2002; Ellis & He,

1999) or the analysis of LREs (Swain & Lapkin, 2001; Swain & Lapkin, 1995). LREs are defined as instances in which learners may ask about the meaning, spelling or pronunciation, ask about the correctness of a grammatical form or correct their own or another's use of a word, form or structure (Leeser, 2004; Swain & Lapkin, 1995). It has been evidenced that learners' use of LREs not only represent language learning in process (Donato, 1994; Swain & Lapkin, 1998) but also correlate positively with L2 development (Basturkmen, Loewen & Ellis, 2002; Leeser, 2004; Williams, 1999). It would be interesting to further explore the amount and type of negotiation, as well as LREs of learners' interactions while completing the spoken PO activities.

Fourth, the study used only one-week delayed post-test, without an immediate test or another delayed one. It would be interesting to examine vocabulary use at several points in time to more accurately capture improvement as well as retention rates of the various measures in the present study. Finally, the sample size per group can be considered another limitation for the study. Replicating the study with a larger sample of participants would be helpful.

In conclusion, the present study highlights the potential advantages reported in Chapter 5 of the spoken sentence reconstruction activity over the listen-and-retell meaning activity for learning polysemous SWVs and PVs. This advantage is related to how the listen-and-retell meaning activity in the present study did not offer learners the opportunity to practise using of the target items in context. Furthermore, the findings are mixed regarding the differences between SWVs and PVs. Differences between the two target item types may be minimal when exploring measures such as text length and MLU, but PVs were found to be more difficult than SWVs when examining LD. These mixed results suggest that aspects of vocabulary use beyond learning gains could exhibit some differences between the two target item types. In the future, it would be interesting to explore the differences between the two types through examination of more appropriate

data in terms of text length and thematic saturation. Finally, the present study highlights the feasibility of learning multiple meanings of polysemous words at the same time, as such words were used in utterances of the same length. Although in Study 2.1 the receptive relative gains showed a pattern in which the first, most frequent meaning sense was better learned than the others, the productive learning gains, which is considered a more advanced component of vocabulary learning were sufficient to suggest that multiple meaning senses could be learned together.

Chapter 7

General Discussion and Conclusions

The present chapter aims to contextualise the findings of the two classroom intervention studies conducted for the present thesis within the previous EFL research on the pedagogy of vocabulary. The two studies provided a thorough examination of the effectiveness of spoken pushed-output (PO) instruction in teaching polysemous SWVs and PVs. This examination was conducted using two different approaches for analysis: receptive and productive (spoken) vocabulary learning gains and the lexical profiling of the utterances produced in the productive test. Three central aims were addressed in the two studies: differences between the treatment conditions, differences between SWVs and PVs, and differences between the three meaning senses. This chapter summarises the main findings and discusses their implications in relation to these three points. Further, some methodological concerns and limitations are discussed, and directions for future research are also provided.

7.1. Learning vocabulary through spoken PO instruction

The first central aim of this thesis was to investigate the effectiveness of spoken PO instruction for teaching vocabulary. In doing so, two classroom intervention studies were conducted; the first one compared spoken PO instruction to traditional instruction for teaching vocabulary. The findings show that, as much previous research has demonstrated (e.g., De la Fuente, 2006; De La Fuente, 2002; Ellis & He, 1999), the spoken PO treatment outperformed the other treatment in terms of the receptive and productive learning gains of both polysemous SWVs and polysemous PVs (see Chapter 3). The present thesis goes

further and shows that this is also true for all measures of lexical profiling, namely, mean length of utterance and the lexical richness represented in density, diversity, and sophistication (see Chapter 4). As discussed in Chapters 3 and 4, although it might be attributed to a clear task effect, this finding is encouraging insofar that it indicates that L2 learners can achieve reasonably good learning of the multiple meanings even for PVs within the spoken PO instruction.

The second study examined the differences between three spoken PO activities in learning polysemous SWVs and PVs: sentence reconstruction, listen-and-retell meaning, and picture description activities. The three activities differed in terms of their ILs and TFAs as well as the type of retrieval required to complete the activity, as illustrated in Section 5.3.5. The findings showed that learners in the sentence reconstruction activity outperformed learners completing the listen-and-retell meaning activity in some of the measures, namely, receptive learning gains, mean length of utterance, lexical density, and lexical sophistication (range). However, learners completing the same two activities reached the same level in terms of their productive learning gains as well as in two measures of lexical profiling: lexical diversity and lexical sophistication (frequency). The picture description activity was similar in effect to both the sentence reconstruction activity and the listen-and-retell meaning activity concerning all measures: learning gains and lexical profiling. The superiority of the sentence reconstruction activity over the listen-and-retell meaning activity was explained in relation to the type of retrieval required to complete the activities. In both the learning and testing stages, learners were required to retrieve the form in spoken mode. This process of retrieving the form is more complicated than retrieving the meaning, as asserted by Nation and Webb (2011) as well as Laufer, Elder, Hill, and Congdon (2004). However, all three activities resulted in similar and reasonably good productive learning gains, which are more complex than the receptive gains (Henriksen, 1999; Nation, 2013). Thus, it seems reasonable to suggest

that the present thesis' findings imply that no activity had a clear advantage over the others. As stated above, there were some differences among the three activities, but this depended on the type of measure. The value and the choice of these activities would, therefore, depend on the goals that teachers envision for their learners and the learners' needs and preferences. That is, teachers should carefully consider what levels of knowledge they are trying to enhance or what learners prefer or need to improve regarding the different components of the vocabulary knowledge; they should then use teaching methodologies that address those components.

Nation (2007) conceived that language learning should be situated within an integrated, four-tiered approach called "the four strands", consisting of meaning-focused input, meaning-focused output, language-focused learning, and fluency development (p. 2). Meaning-focused input implies that language can be learned through listening and reading, i.e., using the language receptively. Further, meaning-focused output refers to learning language through speaking and writing, i.e., using language productively. Language-focused learning involves "the deliberate learning of language features such as pronunciation, spelling, vocabulary, grammar and discourse" (Nation, 2007, p. 6). Fluency development involves all of the four skills: listening, reading, writing, and speaking. Learners should make use of all they learned within this strand. Any well-designed coursework should be "an even balance" of these strands with the same amount of time given to each strand (Nation, 2007, p. 2). Spoken PO activities can be situated within the meaning-focused output in L2 learning. Nation (2007) stated that "Swain's (1985) output hypothesis has been influential in clarifying the role of speaking and writing in second language learning" (p. 3). Swain (2000), in the discussion of collaborative dialogue, provided evidences for the success of output in L2 learning, which is derived from activities situated within communicative meaning-based interactions. These activities include conversations, giving a speech, writing a letter, telling a story, and so

on. It is apparent that these activities allow for more creative use of language features in a free production process. The activities in the present thesis were different in the sense that they were output activities with a specific language focus, i.e., improving vocabulary learning. The activities were designed for practicing a specific and limited set of lexical items. The output in these activities can be seen as the means for deliberately focusing on vocabulary. Putting these altogether, the spoken PO vocabulary activities in the present thesis combined meaning-focused output and language-focused learning.

Apart from the functions of output presented by Swain (2005), which were discussed comprehensively within previous chapters (see Chapters 2, 3, and 4), Nation (2007) listed certain conditions for an approach to be considered language-focused. These conditions included:

> (1) The learners give deliberate attention to language features, (2) The learners should process the language features in deep and thoughtful ways, (3) There should be opportunities to give spaced, repeated attention to the same features, (4) The features that are focused on should be simple and not dependent on developmental knowledge that the learners do not have, and (5) Features that are studied in the language-focused learning strand should also occur often in the other three strands of the course. (Nation, 2007, pp. 6-7)

The pushed-output instruction examined in the present thesis met four of these conditions with the exception being the last one, that is, ensuring the occurrence of the target features within the other three strands. Direct attention was paid to the taught vocabulary not only in the practice stage but also in the presentation stage, i.e., the mini-lesson that was given at the beginning of each session. Second, all the activities in the present thesis required learners to orally use the items, implying that these lexical items might have been processed deeply to be used in speaking. Further, learners were working on vocabulary throughout the time of the study, though the taught vocabulary was

different every week. Initially, I wanted to examine spoken PO instruction in a longitudinal study involving spaced spoken practice of the same taught vocabulary within a six-month period. However, the practicalities of the teaching context and access to participants did not allow this; hence, a shorter treatment period had to be elected.

As discussed in Section 2.2.4.2, there is substantial evidence that learning vocabulary deliberately results in durable knowledge. It is also evident that deliberate learning is effective for learning multiword units. The productive learning gains as well as the improvements in the lexical profiling measures reported in the present thesis could reflect the three major roles that Williams (2005) advocates for language-focused learning: (1) noticing a form/word for the first time in the input, potentially leading to a conversion to intake; (2) noticing that an interlanguage form is at odds with the target language input, leading to destabilization of that form; or (3) incorporating a new form into the developing interlanguage.

Further, the present thesis revealed some major features of spoken PO activities that may have contributed to the generation of learning opportunities, including the instances of collaborative completion of the activities. Learners were working in dyads on joint problem-solving processes. These processes were sometimes at the level of completing only the language form, which requires a very minimal level of interaction and collaboration, while other instances involved collaborating on the meaning they were intending to express. Learners were engaged in different forms of collaborative behaviours, ranging from simple repetition to testing their own hypothesis about the appropriateness of words and different meanings. Language, in this case, as advanced by Swain (1995), served as a mediating tool that allowed the learners to lead and initiate problem-solving processes, they may face during the process of L2 acquisition, which is called collaborative dialogue, as advocated by Swain (2000). It would have been interesting to record and analyse these instances of collaborative dialogue; however, this

was not possible in the teaching context due to a lack of recording equipment. However, based on my own observations during data collection, this active collaboration, as well as engagement with the activities, seemed to lead to learners feeling more confident and comfortable in language learning.

Additionally, another interesting feature of the interaction occurring in these spoken activities was the provision of feedback, which forms a critical part of the language learning process. Learners mutually corrected their own and each other's oral production, e.g., pronunciation and appropriacy. The two elements of feedback that are relevant to the Output Hypothesis were the recasts and scaffolding. In recasts, one learner corrected the speaker by rephrasing their utterances and changing one or more components of the sentence while still referring to the intended meaning (Long, 1996). This way, the environment became conducive as well as relevant for creating the learning process. In scaffolding, one learner assisted their peer in three possible ways: graduation, in which help gradually improves from implicit to explicit up until an appropriate level; contingency, in which help is given only when it is needed and stopped when the learner is able to tackle the problem independently; and ongoing assessment (Ellis & Barkhuizen, 2005). The importance of feedback in this context also provided the learners with the relevant resources concerning the target language in the production process.

It might also be possible that the input that learners produced for their peers could have been incorrect in terms of either the pronunciation or the appropriacy of meaning in the specified context, which could have affected their learning. It would be interesting to implement immediate feedback techniques in future studies. One example would be to implement any interactive learning tools that are available nowadays, such as *Sounds: The Pronunciation App*, which allows learners to not only listen to how a word is pronounced but also to record and playback their own pronunciation. Another example would be to use the *howjsay Pronunciation Dictionary* in which learners can listen to the

correct pronunciation of words. While completing the activities together, learners might have access to such tools that could help them to improve their own and their peers' pronunciation.

Spoken PO activities could also encourage more learner autonomy in learning vocabulary (Ellis & Wulff, 2014). Learner autonomy, as defined by Benson (2013), is the capacity to take control of one's learning, which is broadly recognised as a desirable goal of L2 learning. Several proponents posit that the language classroom is where the autonomy of learners begins (e.g., Nunan, 1997; Sinclair & Ellis, 1984). Little (1995) stated that autonomy is "not a steady state achieved by learners once and for all" (p. 3), which means learners' readiness to engage in autonomy fluctuates considerably from time to time and from one activity to another. Engaging learners in the two-way spoken PO activities could have been the very first step to train learners in how to be autonomous in their learning. That is because all responsibility and control were transferred to learners to complete the activities together. To do so, they needed to produce language, and that production was the input for their peers. Learners might have felt more responsible for that production and aware of the importance of its accuracy for their peers.

To conclude, it should be acknowledged that pushed output instruction is not without its limitations. First, the instances of pushed-output could be difficult to implement in an authentic classroom. It is often the case that teachers have specific coursework to finish in a tight schedule, which does not allow for many opportunities to implement such instruction. Further, as Krashen (1998) posits, pushing learners to produce output could provoke stress and anxiety. While there may be some merit to this argument, think-aloud protocol data from output and feedback studies (e.g., Mackey, 2002; Mackey, Gass, & McDonough, 2000) suggest that many learners want to receive and learn from feedback concerning their output. It also could be argued that the influence of output on acquisition is doubtful, as many studies have shown that language learning

can take place based on input in the absence of output (S. Krashen, 1998). It is important to note that, given that Swain (1985, 1995) did not negate the essential presence of input and merely stated that output facilitates acquisition in ways that differ from input, this criticism does not seem relevant. Many spoken PO vocabulary activities include a mixture of input and output, as the ones used in the present thesis, and, in this situation, one person's output can be another person's input.

7.2. SWs vs. FSs: similar or different?

The second central aim of this thesis involved comparing the learning of SWVs to the learning of PVs within the two studies. Learning PVs proved to be similar to that of SWVs in terms of receptive and productive learning gains. These findings align to the body of research demonstrating that the learnability of both SWs and FSs will be similar if other factors are controlled for, such as the amount and type of exposure and practice (Pellicer-Sánchez, 2020).

With reference to Nation's (1990) four strand approach, FSs need to be learned across the four strands: meaning-focused input, meaning-focused output, language-focused learning, and fluency development. Exercises that deliberately focus on multiword units are very effective in yielding results in both the receptive and productive knowledge of multiword units (Nation, 1990). Five main foci for the deliberate learning of FSs were identified by Nation (1990): encouraging noticing, rote learning, the use mnemonics, seeing patterns, and strategy development. FSs are rarely included in any principled manner in authentic classrooms (Alali & Schmitt, 2012) despite their importance in language learning and for achieving a native-like proficiency (Gouverneur, 2008). The present thesis shows that FSs can indeed be taught in a principled manner, even in a context like the Saudi context where FSs are often marginalised. Furthermore,

the spoken PO instruction that was implemented in the two studies in the present thesis could have also helped to develop these five foci of deliberate learning. As a result, considerable learning gains occurred on both the receptive and productive levels.

These findings provide more evidence for the learnability of FSs in the context of the EFL classroom. Gairns and Redman (1986) advised that multiword units should be treated as regular items and taught particularly if these multiword units were more frequent in the language. The present study supports this, as both types of items were learned in a similar manner in terms of the learning gains. FSs are important for improving communicative competence, and the instruction of lexical phrases should be more systematically included in language teaching materials.

However, it should be acknowledged that SWVs still hold an advantage over PVs when examining measures beyond the learning gains, such as the overall text length, MLU and lexical density. Although the improvements were similar at the level of gains, learners were better at using SWs and still found it more difficult to use FSs in context. This could probably be due to the fact that they have more experience with using SWs and might feel more comfortable using them. This finding is in line with Schmitt and Carter's (2004) conclusion that the use of FSs typically lags behind other aspects of language learning. Teachers can complement spoken PO activities for PVs or others with more overt instructional strategy to enhance the use of PVs in context. This might help learners in developing strategies to use PVs on their own in contexts beyond the classroom. Notwithstanding these explanations, the findings suggest that FSs need far more exposure than SWs in the classroom instruction and materials to further improve the quality of their usage in context.

7.3. Learning polysemous words

The third central aim of this thesis involved investigating student learning of the multiple meaning senses of polysemous SWVs and PVs. The findings indicated that some differences occurred within receptive gains in which the first meaning sense was learned better. However, the overall performance in productive gains indicated that no sense was different than another.

As discussed in Chapter 2 (Section 2.1), the knowledge of a word is multidimensional in nature and entails much more than knowing the literal form-meaning link. Few studies have examined the development of vocabulary knowledge depth, but many studies have focused on examining the development of vocabulary knowledge size, as discussed in Section 2.1.2. It has been suggested that different aspects of word knowledge develop in parallel to a certain extent (e.g., González-Fernández & Schmitt, 2019; Laufer, 1997; Schmitt, 1998, 2010; Webb, 2007). Polysemy, for instance, is one aspect that reflects the depth of vocabulary knowledge. Studies have shown that the recall knowledge of multiple meaning senses is the strongest contributor to vocabulary knowledge depth (e.g., González-Fernández & Schmitt, 2019; Tannenbaum, Torgesen, & Wagner, 2006). Polysemy is not a fringe component of language (M. McCarthy, 1990), rather, "it is at the very heart of word meaning" (p. 26). However, learners usually lack knowledge of the different meaning senses that a word can hold. Such a gap in the learners' knowledge may stem from different factors that characterise teaching practices in the EFL context: the inadequacy of input that allows multiple exposures to a range of senses to permit incidental learning, the misconception of teaching words as discrete items with a single meaning sense at one exposure despite the syntagmatic relations between the different senses; and the negligence of explicit, or deliberate, teaching of polysemy, just to name a few. The present thesis suggests that learning a word's range of

senses is feasible if other factors are controlled for, such as exposure and practice. Even less frequent meaning senses do not seem to be particularly more difficult to acquire than high frequent meaning senses, given that all meaning senses are provided with enough exposure and practice.

The present thesis reflects how this aspect could be improved through only one instruction approach: spoken PO instruction. Future studies should examine other instructional techniques for learning polysemous words. Further, the present thesis did not compare whether learning multiple meaning senses in one exposure is better than teaching the meaning senses separately. The findings, at least, provide evidence that learning them altogether is actually possible at both the receptive and productive levels.

7.4. Methodological implication

Two approaches to data analysis were used in the present thesis to examine the effectiveness of spoken pushed-output instruction. The first approach involved analysing receptive and productive learning gains, which are commonly used in most research on vocabulary teaching and learning. In doing so, two tests were used: a form-meaning recognition test (written) and a form-meaning recall test (oral). The aim of the form-meaning recognition test is to measure the receptive form-meaning link. Given the large number of target items in the study (n=144), it deemed more appropriate to use the written format to allow testing multiple target items within the same test item and reduced the number of distractors. Further, considering the learners' low proficiency level as well as their lack of familiarity with oral tests in general, I decided to go only for the productive test to be oral and not have the two tests which would be unfamiliar to learners. Now, if it is the case in which the test modality could have affected the learners' performance; then, this would be in favour for the traditional group. The results did not confirm that

but actually showed that learners in the spoken pushed-output treatment outperformed learners in traditional group in both tests. Further, the format of the form-meaning recall test used in the present study is less common in in vocabulary research. It would have been interesting if a different format that allows spontaneous aspects of L2 vocabulary use was utilised; however, as explained in Section 3.3.5.3, utilising such tests are considered beyond the learners' proficiency level and might have resulted in no knowledge being demonstrated. It would be interesting if a more spontaneous-use format would be used in future studies examining oral vocabulary knowledge within oral PO instruction.

The second approach looked beyond learning gains and explored the lexical profile of students' responses, utilising several measures, namely, mean length of utterances, lexical diversity, lexical density, and lexical sophistication. Examining the lexical profiles of the responses that learners produce orally provided interesting data and allowed for the differences between SWVs and PVs to emerge. This showed that PVs are indeed more difficult to use creatively in a context than SWVs, something that was not captured through the receptive or productive learning gains. The focus of the body of research examining these measures is not usually vocabulary pedagogy. Instead, the focus is typically examining the degree of variability in lexical profiles in relation to language proficiency. The present thesis suggests that studies examining vocabulary teaching in a classroom context would benefit from utilising measures beyond learning gains to obtain a richer picture of the vocabulary learning process.

It must be noted that, though this analysis allowed an examination of the overall lexical production of a learner from a relatively small piece of speech, it is unclear whether these measures would change with the test type. The productive data analysed in the present thesis were responses provided in a testing environment, lacking the textual homogeny that is usually found in other free production activities. It remains to be proven

whether the same results would be obtained with other types of productive data. It is further important to note that the lexical profiling measures also seem to behave unpredictably in some occasions. For instance, in Study 1, differences were found between SWVs and PVs with SWVs having an advantage over PVs in almost all measures. The only exception was lexical sophistication (in terms of range) for which PVs' scores were greater than SWVs. This was unexpected and, hence, unexplainable. My expectation was that if differences were found between SWVs and PVs, then SWVs would have the advantage over PVs based on previous research findings (Kasahara, 2011; Kasahara, 2010; Peters, 2014). The same pattern also occurred in Study 2, which makes it difficult to predict whether it is related to the type of data used in both studies (testing data), the analysis tool (TAALES), or the choice of the index (spoken BNC range counts). More research is needed in vocabulary pedagogy utilising such measures to reach a better understanding of how these measures behave in different contexts.

7.5. Limitations and directions for future studies

The limitations of the studies presented in the present thesis have already been addressed at the end of Chapters 3, 4, 5, and 6. This section covers more general limitations and directions for future studies. In my view, there are five main lines of inquiry that directly derive from the findings of the present thesis and are worth pursuing in future research.

The first line of research relates to the development of overall speaking skills under the different learning conditions. Previous research has shown that learners' overall speaking performance can be improved by means of written pushed-output instruction (Swain & Lapkin, 2001). However, the present thesis has not looked at how the overall speaking ability might have developed in the different treatment conditions. Thus, future

research could explore the effect that spoken PO activities have on the overall speaking performance by examining other non-lexical measures, such as grammar accuracy, morphology, appropriateness, contextual structure, etc., or by examining some fluency measures such as speech rate, articulation rate, number of silent pauses, number of repetitions, number of repairs, etc.

Another possible avenue for research relates to the interaction among learners while completing the activities. Many researchers have examined not only the final production of learners but also the language-related episodes (LREs) in which learners are talking about language. It would be interesting to analyse LREs within the process of completing the activities to further examine how these LREs are related in interaction to learning gains and to the development of both lexical and overall speaking development.

Third, the findings of the present thesis examined the polysemy of high frequency words. There are two main limitations that could be identified here. First, the study involved only one type of polysemy in which meaning senses were interrelated to some extent. There are different types of polysemy (Cruse, 2011), such as linear and non-linear polysemy. Linear polysemy accounts for a specialisation-generalisation relationship between the meaning senses. One type of linear polysemy is autohyponomy, which refers to the occasion "whereby a polysemic or polysemous lexical item functions in one of its senses as the superordinate to itself in another sense" (Huang, 2012., p. 36). An example of autohyponomy is *dog* which can refer to a *member of the canine race* in a more general way or to a *male member of the canine race* in a more specific way (Cruse, 2011). Another type of linear polysemy is automeronymy in which "a part and immediate whole have the same name (but distinct senses, cf. autohyponomy)" (Cruse, 2011, p. 173). One example would be the human body meronymy in which the term *body* can refer to *the whole body* or can also be used as *a close equivalent to "trunk"* (Cruse, 2011, p. 173). Non-linear polysemy refers to metaphorical and metonymous polysemy. For instance, the word

"mouth" in the sentence *"There are too many mouths to feed"* refers to people and is used metonymically, whereas, in *"Don't talk with your mouth full"*, it is used literally (Cruse, 2011, p. 173). The present thesis did not account for these differences among the meaning senses. Future research should examine the learning of different types of polysemous words. The second limitation relates to the fact that all target items in the present thesis, SWVs or PVs, were high frequency ones; hence, the findings cannot be generalised to polysemous low frequency lexical items. Future research could examine the knowledge of the multiple meaning senses of low frequency lexical items.

The fourth avenue of future research concerns the measures of lexical profiling of utterances used in the present thesis. The present thesis utilised four measures of lexical profiling: lexical diversity, lexical density, lexical sophistication, and mean length of utterances. As discussed in Chapter 4 (Section 4.1), there are many measurement tools that could have been explored. The rationale behind the choice of these specific measures is explained in Section 4.3.3. It could be argued that lexical profiling involves more than these measures; for instance, lexical sophistication involves more than looking at the frequency and range of the vocabulary used (Kyle & Crossley, 2015). Lexical sophistication could be measured using certain lexical measures related to the word, such as concreteness, contextual diversity, lexical access, etc., or lexicogrammatical measures, such as n-grams, dependency relations, and verb-verb argument construction strength of association (S. A. Crossley & Kyle, 2018). However, as with other aspects of language assessment, there is no established way of measuring aspects of lexical profiling (James Milton, 2009). It would be interesting to utilise different measures when examining the effect of spoken PO treatment or when investigating the differences between the learnability of SWs and FSs and the differences in the learnability of the multiple meaning senses of a word.

The final line of research is related to the perceptions of EFL teachers concerning the implementation of spoken PO instruction in the language classroom. Spoken PO instruction might be relatively new in the Saudi context; hence, both learners and teachers might not be familiar with it. The findings have shown that learners who are not familiar with such instruction at all can benefit from it. The spoken PO instruction was implemented by me and did not involve the teachers. The present thesis did not look at teachers' opinions and attitudes toward such an approach. In reality, it depends on the teachers' abilities to implement such instruction in the classroom context. It is also dependent on teachers' understanding of how the approach works and whether they have the resources necessary to make the implementation possible. It would be interesting in future studies to examine the teachers' perceptions of the implementation of spoken PO instruction in the EFL classroom.

7.6. Concluding remarks

While the written PO approaches have been proven to be effective for improving vocabulary knowledge, research has not provided sufficient evidence about the effectiveness of spoken PO approaches. Even more so, the spoken PO instruction was not investigated either for teaching FSs or for teaching polysemous individual words. The present thesis investigated the effectiveness of spoken PO activities for learning polysemous SWVs and PVs. The results showed that learning polysemous vocabulary, SWs and FSs, seems to be better improved through spoken PO instruction than through traditional instruction. Furthermore, the learners' ability to use the items in context was also better improved when examining several lexical profiling measures, such as mean length of utterances, lexical diversity, lexical density, and lexical sophistication. Moreover, when providing the same type and amount of exposure and practice, FSs seem

to be learned in a similar way to SWs, and the multiple meaning senses of words seemed to be learned similarly. Overall, the findings of the present thesis have provided evidence for the positive effect that spoken PO activities have on the development of polysemous SWVs and PVs, even in a context where this is not a familiar approach. The reported positive effects confirm the usefulness of the approach for learning different types of lexical items and warrant further exploration.

References

Abdalla, A. H. (2014). Investigating the Effect of Consciousness-raising Techniques on Learning the English Tense System by EFL Students. *Journal of Language Teaching & Research*, *5*(4), 810–818.

Adolphs, S., & Durow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use* (pp. 107–126). Amsterdam; Philadelphia: John Benjamins Publishing.

Al Khateeb, A. A. (2015). Hypothesis of interaction: reflections on its theoretical and practical contributions for Second Language Acquisition (SLA). Studies in English Language Teaching, 2(3), 294.

Al-Akloby, S. (2001). *Teaching and learning English vocabulary in Saudi Arabian public schools.* (Unpublished doctoral thesis). University of Essex, Colchester, England.

Alali, F. A. & Schmitt, N. (2012). Teaching formulaic sequences: The same as or different from teaching single words? *TESOL Journal*, *3*(2), 153–180.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409.

Al-Motairi, A. D. M. (2005). *Towards improving the teaching of English as a foreign language in the Saudi context: A case study of five Saudi boys secondary schools.* (Unpublished doctoral thesis). University of East Anglia, Norwich, England.

Al-Nafisah, K. I. (2000). *A study of the curriculum and methodology for the teaching of English in Saudi Arabia with particular reference to learning difficulties*

encountered by students. (Unpublished doctoral thesis). University of Wales, Swansea, Wales.

Al-Nujaidi, A. H. (2003). *Relationship between vocabulary size, reading strategies, and reading comprehension of efl learners in Saudi Arabia.* (Unpublished doctoral thesis). Oklahoma State University, Stillwater, Oklahoma.

Al-Seghayer, K. (2011). *English teaching in Saudi Arabia: Status, issues, and challenges*. Riyadh: Hala Print CO.

Aluthman, E. S. (2017). Compiling an OPEC word list: A corpus-informed lexical analysis. *International Journal of Applied Linguistics & English Literature*, *6*(2), 78–91.

Anderson, R., & Freebody, P. (1981). Vocabulary Knowledge. In J. Guthrie (Ed.), *Comprehension and Teaching: Research Reviews* (pp. 77–177). Newark, DE: International Reading Association.

Arnaud, P. J. L. (1984). The Lexical Richness of L2 Written Productions and the Validity of Vocabulary Tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and Problems in Language Testing* (pp. 14–28). Colchester, England: University of Essex.

Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. *Psychological Review*, 85, 139–152.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*(1), 101–114.

Baleghizadeh, S., & Derakhshesh, A. (2012). The Effect of Task Repetition and Noticing on EFL Learners' Oral Output. *International Journal of Instruction*, *5*(1), 141–152.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. *Language Learning*, *52*(2), 323–363.

Barcroft, J. (2006). Can writing a new word detract from learning it? More negative effects of forced output during vocabulary learning. *Second Language Research*, *22*(4), 487–497.

Bardel, C., & Gudmundson, A. (2018). Developing Lexical Complexity in Oral Production: limitations and possibilities of the advanced L2 learner. In K. Hyltenstam, K., Bartning, I. & Fant, L. (Eds.), *High-Level Language Proficiency in Second Language and Multilingual Contexts* (pp. 120–145). Cambridge: Cambridge University Press.

Barfield, A., & Gyllstad, H. (2009). *Researching collocations in another language: Multiple interpretations*. New York: Palgrave Macmillan.

Bassiouney, R., & Katz, E. G. (2012). *Arabic language and linguistics*. Washington, DC: Georgetown University Press.

Basturkmen, H., Loewen, S., & Ellis, R. (2002). Metalanguage in focus on form in the communicative classroom. *Language Awareness*, *11*(1), 1–13.

Bates, E., & Goodman, J. C. (1997). On the Inseparability of Grammar and the Lexicon: Evidence from Acquisition, Aphasia and Real-time Processing. *Language and Cognitive Processes*, 12 (5–6), 507–584.

Beal, V. (2007). *The weight of involvement load in college level reading and vocabulary tasks*. Doctoral dissertation, Concordia University, Portland, OR, USA.

Benson, P. (2013). *Teaching and researching: Autonomy in language learning* (Applied linguistics in action). Harlow: Longman.

Bensoussan, M. & Laufer, B. (1984). Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading*, *7*(1), 15–32.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes*, *43*(2), 79–120.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*(1), 3–44.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 657.

Boers, F. & Demecheleer, M. (1998). A cognitive semantic approach to teaching prepositions. *ELT Journal*, *52*(3), 197–204.

Boers, F. (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics*, *21*(4), 553–571.

Boers, F. (2013). Cognitive linguistic approaches to teaching vocabulary: Assessment and integration. *Language Teaching*, *46*(2), 208–224.

Boers, F., & Lindstromberg, S. (2008). How cognitive linguistics can foster effective vocabulary teaching. In F. Boers, & S. Lindstromberg, (Eds.), *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology* (pp. 1–64). Walter de Gruyter.

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., Demecheleer, M., Coxhead, A. & Webb, S. A. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research*, *18*(1), 54–74.

Boers, F., Eyckmans, J. & Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: more than mere mnemonics? *Language Teaching Research*, *11*(1), 43–62.

Boyle, R. (2009). The legacy of diglossia in English vocabulary: what learners need to know. *Language Awareness*, *18*(1), 19–30.

Bransford, J. D., Franks, J. J., Morris, C. D., & Stein, B. S. (1979). Some general constraints on learning and memory research. In L.S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 331–354). Hillsdale, NJ: Erlbaum.

Brodzinski, K. (2009). Zooming in on phrasal verbs. *Modern English Teacher*, *18*(3), 12–13.

Brown, R. (1973). *A first language: The early stages.* Oxford, England: Harvard University Press.

Catalán, R. M. J., & Llach, M. P. A. (2017). CLIL or time? Lexical profiles of CLIL and non-CLIL EFL learners. *System*, *66*, 87–99.

Čermák, F. (2001). Substance of idioms: Perennial problems, lack of data or theory? *International Journal of Lexicography*, *14*(1), 1–20.

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

Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research*, *10*(2), 157–187.

Chaudron, C. (1988). *Second language classrooms: Research on teaching and learning*. Cambridge: Cambridge University Press.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics*, *31*(5), 693–713.

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25.

Choi, Y., & Kilpatrick, C. (2014). Hypothesis Testing in Task-based Interaction. *Applied Language Learning*, *23-24*(1), 1-18.

Conklin, K. & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72–89.

Conklin, K. & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423.

Cowie, A. P. (1994). Phraseology. In R.E. Asher, & J. M. Y. Simpson (Eds.), *The Encyclopedia of Language and Linguistics* (Vol. 6, pp. 3168–3171). Oxford: Pergamon.

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

Craik, F. I. M., & Rose, N. S. (2012). Memory encoding and aging: a neurocognitive perspective. *Neuroscience & Biobehavioral Reviews*, *36*(7), 1729–1739.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268.

Crookes, G., & Rulon, K. A. (1988). Topic and feedback in native-speaker/nonnative-speaker conversation. *Tesol Quarterly*, *22*(4), 675–681.

Crossley, S. A., & Kyle, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*, *38*, 46–50.

Crossley, S. A., Salsbury, T. & Mcnamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*(5), 570–590.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580.

Crossley, S., Salsbury, T. & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*(3), 573–605.

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

Cruse, A. (2011). *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford University Press.

Csábi, S. (2004). A cognitive linguistic view of polysemy in English and its implications for teaching. In M. Achard, & S. Niemeier, (Eds.), *Cognitive Linguistics, Second*

Language Acquisition, and Foreign Language Teaching (pp. 233–256). Berlin, New York: Mouton de Gruyter.

D'Anna, C. A., Zechmeister, E. B., & Hall, J. W. (1991). Toward a meaningful definition of vocabulary size. *Journal of Reading Behavior, 23*(1), 109–122.

Dagut, M., & Laufer, B. (1985). Avoidance of phrasal verbs – A case for contrastive analysis. *Studies in Second Language Acquisition*, *7*(1), 73–79.

Daller, H., & Phelan, D. (2007). What is in a teacher's mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 234–244). Cambridge: Cambridge University Press.

Daller, H., & Xue, H. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 150–164). Cambridge: Cambridge University Press.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222.

Darwin, C. M., & Gray, L. S. (1999). Going after the phrasal verb: An alternative approach to classification. *Tesol Quarterly*, *33*(1), 65–83.

De Bot, K. (1996). The psycholinguistics of the output hypothesis. *Language Learning*, *46*(3), 529–555.

De la Fuente, M. J. (2002). Negotiation and oral acquisition of L2 vocabulary. *Studies in Second Language Acquisition*, *24*(1), 81–112.

De la Fuente, M. J. (2006). Classroom L2 vocabulary acquisition: Investigating the role of pedagogical tasks and form-focused instruction. *Language Teaching Research*, *10*(3), 263–295.

De Villiers, J. G., & De Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research*, *2*(3), 267–278.

DeKeyser, R. (1998). Beyond focus on form: Cognitive perspectives on learning and practicing second language grammar. In C. Doughty & J. Williams (Eds.), Focus on Form in Classroom Second Language Acquisition. Cambridge: Cambridge University Press.

DeThorne, L. S. (2002). *Language and IQ in children with language impairment*. Dissertation. University of Illinois, Urbana-Champaign.

Dethorne, L. S., Johnson, B. W., & Loeb, J. W. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics & Phonetics*, *19*(8), 635–648.

Donato, R. (1988). *Beyond group: A psycholinguistic rationale for collective activity in second-language learning*. Dissertation. Newark: University of Delaware.

Donato, R. (1994). Collective scaffolding in second language learning. In J. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33-56). Norwood, NJ: Ablex.

Dornyei, Z. (2009). *The Psychology of Second Language Acquisition.* Oxford: Oxford University Press.

Dupuy, H. J. (1974). *The rationale, development, and standardization of a basic word vocabulary test*. Washington, DC: US Government Printing Office.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics*, 25(2), 220–242.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL– International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

Eckerth, J., & Tavakoli, P. (2012). The effects of word exposure frequency and elaboration of word processing on incidental L2 vocabulary acquisition through reading. *Language Teaching Research*, *16*(2), 227–252.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C., & Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function Introduction to the special issue. *The Modern Language Journal*, *93*(3), 329–335.

Ellis, N. C., & Wulff, S. (2014). Usage–based approaches to SLA[1]. In B.Vanpatten & J. Williams, (Eds.). *Theories in Second Language Acquisition: An Introduction* (pp. 75-93). London: Routledge.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Ellis, R., & He, X. (1999). The roles of modified input and output in the incidental acquisition of word meanings. *Studies in Second Language Acquisition*, *21*(2), 285–301.

Ellis, R., Tanaka, Y. & Yamazaki, A. (1994). Classroom interaction, comprehension, and the acquisition of L2 word meanings. *Language Learning*, *44*(3), 449–491.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155.

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.

Evans, V. & Tyler, A. (2004). Rethinking English 'prepositions of movement': The case of to and through. *Belgian Journal of Linguistics*, *18*(1), 247–270.

Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology*, *69*(2), 157–169.

Fan, M. (2009). An exploratory study of collocational use by ESL students – A task based approach. *System*, *37*(1), 110–123.

Fan, M., & Nyikos, M. (2007). A review of vocabulary learning strategies: Focus on language proficiency and learner voice. In A. S. Cohen & E Macaro (Eds.), *Language learner strategies: thirty years of research and practice* (pp. 251–274). Oxford: Oxford University Press.

Fellbaum, C. (2000). Autotroponymy. In, Y. Ravin & C. Leacock (Eds.), *Polysemy: Theoretical and Computational Approaches* (pp. 52–67). Oxford: Oxford University Press.

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research*, *58*(3), 840–852.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Los Angeles : SAGE.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378.

Folse, K. S. (2006a). *The art of teaching speaking: Research and pedagogy for the ESL/EFL classroom*. Ann Arbor, MI: Michigan University Press.

Folse, K. S. (2006b). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273–293.

Freebody, P., & Anderson, R. C. (1981). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Center for the Study of Reading Technical* (Report; No. 225).

Gairns, R., & Redman, S. (1986). *Working with words: A guide to teaching and learning vocabulary*. Cambridge: Cambridge University Press.

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645–666.

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, *59*(1), 29–44.

Gass, S. M. (1988). Integrating Research Areas: A Framework for Second Language Studies. *Applied Linguistics*, *9*(2), 198–217.

Gass, S. M. (2003). Input and interaction. In C. J. Doughty and M. H. Long (Eds.), *The Handbook of Second Language Acquisition*. Malden, MA ; Oxford: Blackwell.

Gass, S. M. (2017). *Input, interaction, and the second language learner*. New York: Routledge.

Gass, S. M., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, *16*(3), 283–302.

Gass, S. M., Behney, J., & Plonsky, L. (2013). Looking at inter-language processing. In S. M. Gass (Eds.) *Second Language Acquisition: An Introductory Course* (pp. 252–292). London: Routledge.

Gass, S. M. (2017). *Input, interaction, and the second language learner*. London: Routledge.

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.

Geeraerts, D. (2009). Vagueness's puzzles, polysemy's vagaries. *Cognitive Linguistics (includes Cognitive Linguistic Bibliography), 4*(3), 223–272.

Gilquin, G. (2008). Taking a new look at lexical networks. *Lexis, 1,* 23–39.

Gilquin, G. (2010). *Corpus, cognition and causative constructions*. Amsterdam, Philadelphia: John Benjamins Publishing.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology, 66*(3), 325–331.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, London: University of Chicago Press.

González-Fernández, B., & Schmitt, N. (2019). Word Knowledge: Exploring the Relationships and Order of Acquisition of Vocabulary Knowledge Components. *Applied Linguistics*,

Gouverneur, C. (2008). The phraseological patterns of high-frequency verbs in advanced English for general purposes: A corpus-driven approach to EFL textbook analysis.

In F. Meunier, & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 223–243). Amsterdam: John Benjamins.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, Analysis, and Applications*, *145160*.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. *Phraseology: An Interdisciplinary Perspective*, *28*, 27–49.

Grant, L. E., & Nation, I. S. P. (2006). How many idioms are there in English? *ITL-International Journal of Applied Linguistics*, *151*(1), 1–14.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437.

Gries, S. T. (2015). Polysemy. In E. Dabrowska & D. S. Divjak (Eds.), *Handbook of cognitive linguistics* (pp. 472–490). Berlin: De Gruyter Mouton.

Groot, P. J. M. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, *4*(1), 56–76.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL – International Journal of Applied Linguistics*, *166*(2), 278–306.

Gyori, G. (2002). Semantic change and cognition. *Cognitive Linguistics*, *13*(2), 123–166.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics*, *10*(2), 221–240.

Halliday, M. (1989). *Spoken and written language.* (2nd ed.). Oxford: Oxford University Press.

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word*, *17*(2), 241–292.

Harley, B., & King, M. Lou. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, *11*(4), 415–439.

Hasan, A. S. (1988). *Variation in spoken discourse in and beyond the EFL classroom*. (Unpublished doctoral thesis). *University of Aston*.

Hazenberg, S., & Hulstun, J. H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied Linguistics*, *17*(2), 145–163.

Henriksen, B. (1999). Three dimensions of vocabulary development. *Studies in Second Language Acquisition*, *21*(2), 303–317.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development – A progress report. In C. Bardel, C. Lindquist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (Vol. 2, pp. 29–56). Amsterdam: John Benjamin.

Hickey, T. (1991). Mean length of utterance and the acquisition of Irish. *Journal of Child Language*, *18*(3), 553–569.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55–88.

Holster, T. A., & DeLint, D. F. (2012). Output tasks and vocabulary gains. *The Language Teacher*, *36*(4), 5.

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

Hu, H. M., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System*, *56*, 28–39.

Huang, S., Willson, V., & Eslami, Z. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal*, *96*(4), 544–557.

Huang, Y. (2012). *The Oxford dictionary of pragmatics*. Oxford: Oxford University Press.

Hulstijn, J. H. (2001). *Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity*. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–287). Cambridge: Cambridge University Press.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning*, *51*(3), 539–558.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance. *Studies in Second Language Acquisition*, *11*(3), 241–255.

Hulstijn, J. H., & Marchena, E. (1989). Avoidance: Grammatical or semantic causes? *Studies in Second Language Acquisition*, *11*(3), 241–255.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual & Multicultural Development*, *9*, 67–84.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in Second Language Acquisition*, *24*(4), 541–577.

Izumi, S. (2013). Noticing and L2 development: Theoretical, empirical, and pedagogical issues. In J.M. Bergsleithner, S., N. Frota, & J.K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 25-38). Honolulu: University of Hawaii, National Foreign Language Resource Center.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *Tesol Quarterly*, *34*(2), 239–278.

Izumi, S., Bigelow, M., Fujiwara, M. & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition*, *21*(3), 421–452.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84.

Jiang, N. (2004). Semantic transfer and its implications for vocabulary teaching in a second language. *The Modern Language Journal*, *88*(3), 416–432.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working Papers in Linguistics*, *53*, 61–79.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use* (pp. 269–301). Amsterdam, Philadelphia: John Benjamins Publishing.

Kaderavek, J. N., & Sulzby, E. (2000). Narrative production by children with and without specific language impairment: Oral narratives and emergent readings. *Journal of Speech, Language, and Hearing Research*, *43*(1), 34–49.

Kao, R.-R. (2001). Where have the prepositions gone? A study of English prepositional verbs and input enhancement in instructed SLA. *IRAL – International Review of Applied Linguistics in Language Teaching*, *39*(3), 195–216.

Kasahara, K. (2010). Are two words better than one for intentional vocabulary learning? *ARELE: Annual Review of English Language Education in Japan*, *21*, 91–100.

Kasahara, K. (2011). The effect of known-and-unknown word combinations on intentional vocabulary learning. *System, 39*(4), 491–499.

Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365–386.

Kellerman, E. (1986). An eye for an eye: Crosslinguistic constraints on the development of the L2 lexicon. In E. Kellerman and M. S. Smith (Eds.), *Cross-linguistic Influence in Second Language Acquisition* (pp. 35–48). Oxford: Pergamon.

Khodadady, E., & Khaghaninizhad, M. S. (2012). Acquisition of French polysemous vocabularies: Schema-based instruction versus translation-based instruction. *Porta Linguarum*, 17, 29–46.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91–113.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.

Kim, Y. (2011). The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition. *Language Learning, 61*, 100–140.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*(2), 28–41.

Klee, T., Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research, 47*(6), 1396–1410.

Krashen, S. D. (1981). *Second language acquisition and second language learning*. Oxford: Oxford University Press.

Krashen, S. D. (1984). *Writing, research, theory, and applications*. Oxford: Pergamon Press.

Krashen, S. D. (1998). Comprehensible output? *System*, *26*(2), 175–182.

Kremmel, B., & Pellicer-Sánchez, A. (under review). Measuring Vocabulary Development. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of SLA and Language Testing*. London: Routledge.

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., Columbus, G., & Schmitt, N. (2009). The acquisition of phrasal vocabulary. In S. Foster-Cohen (Eds.), *Language acquisition* (pp. 216–240). Palgrave Advances in Linguistics. Palgrave Macmillan, London.

Kurtyka, A. (2001). Teaching English phrasal verbs: A cognitive approach. In M. Putz, S. Niemeier, & R. Dirven (Eds.), *Applied Cognitive Linguistics II: Language Pedagogy* (pp. 29–54). Berlin: Walter de Gruyter.

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786.

Lado, R., Baldwin, B., & Lobo, F. (1967). *Massive vocabulary expansion in a foreign language beyond the basic course: the effects of stimuli, timing, and order of presentation*. (Project No. 5-1095). Washington, DC: U.S. Department of Health, Education, and Welfare.

Lapkin, S., Swain, M., & Smith, M. (2002). Reformulation and the learning of French pronominal verbs in a Canadian French immersion context. *The Modern Language Journal*, *86*(4), 485–507.

Laufer, B. (1990). Ease and difficulty in vocabulary learning: Some teaching implications. *Foreign Language Annals*, *23*(2), 147–155.

Laufer, B. (1997). The lexical plight in second language reading: words you don't know, words you think you know and words you can't guess (pp. 20–34). In J. Coady and T. Huckin (Eds.): *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*. New York: Cambridge University Press.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271.

Laufer, B. (2005). Focus on Form in Second Language Vocabulary Learning. *EUROSLA Yearbook*, 5, 223–250.

Laufer, B. (2011). The Contribution of Dictionary Use to the Production and Retention of Collocations in a Second Language. *International Journal of Lexicography*, *24*(1), 29–49.

Laufer, B. (2015). How Much Lexis is Necessary for Reading Comprehension? In P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (pp. 126–132). London: Macmillan.

Laufer, B., & Eliasson, S. (1993). What causes avoidance in L2 learning: L1-L2 difference, L1-L2 similarity, or L2 complexity? *Studies in Second Language Acquisition*, *15*(1), 35–48.

Laufer, B., & Girsai, N. (2008). Form-focused instruction in second language vocabulary learning: A case for contrastive analysis and translation. *Applied Linguistics*, *29*(4), 694–716.

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

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

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.

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

Laufer, B., & Rozovski-Roitblat, B. (2011). Incidental vocabulary acquisition: The effects of task type, word occurrence and their combination. *Language Teaching Research*, *15*(4), 391–411.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89–108.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202–226.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, *43*(4), 557–570.

Lee, T. D. (1988). Transfer-appropriate processing: A framework for conceptualizing practice effects in motor learning. In *Advances in psychology* (Vol. 50, pp. 201–215). Elsevier.

Leeser, M. J. (2004). Learner proficiency and focus on form during collaborative dialogue. *Language Teaching Research*, *8*(1), 55–81.

Lemmouh, Z. (2008). The relationship between grades and the lexical richness of student essays. *Nordic Journal of English Studies*, *7*(3), 163–180.

Le-Thi, D., Rodgers, M. P. H. & Pellicer-Sánchez, A. (2018). Teaching formulaic sequences in an English-language class: The effects of explicit instruction versus coursebook instruction. *TESL Canada Journal*, *34*(3), 111–139.

Lewandowska-Tomaszczyk, B. (2007). Polysemy, prototypes, and radial categories. In D. Geeraerts & H. Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp. 139–169). Oxford: Oxford University Press.

Li, M., & Kirby, J. R. (2014). The effects of vocabulary breadth and depth on English reading. *Applied Linguistics*, *36*(5), 611–634.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, *54*(2), 193–226.

Lightbown, P. M. (2007). Fair trade: Two-way bilingual education. *ELIA, 7, 9-34.*

Lightbown, P. M., & Spada, N. (2013). *How languages are learned*. Oxford: Oxford University Press.

Lin, L. H. F., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes*, *9*(4), 255–266.

Lindstromberg, S. (1996). Prepositions: Meaning and method. *English Language Teaching Journal, 50*(3), 225–236.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Malmö: Liber Förlag.

Little, D. (1995). Learning as dialogue: The dependence of learner autonomy on teacher autonomy. *System*, *23*(2), 175–181.

Littlemore, J., & Low, G. (2006). Metaphoric competence, second language learning, and communicative language ability. *Applied Linguistics*, *27*(2), 268–294.

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, *45*(4), 661–688.

Liu, D. (2013). *Describing and explaining grammar and vocabulary in ELT: Key theories and effective practices*. London: Routledge.

Loaiza, V. M., McCabe, D. P., Youngblood, J. L., Rose, N. S., & Myerson, J. (2011). The influence of levels of processing on recall from working memory and delayed recall tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1258.

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory*, *10*(5–6), 397–403.

Loewen, S. (2002) . *The occurrence and effectiveness of incidental focus on form in meaning-focused ESL lessons*. (Unpublished doctoral thesis). University of Auckland, New Zealand.

Loewen, S., & Reinders, H. (2011). *Key concepts in second language acquisition*. New York: Palgrave.

Long, M. H. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. Applied Linguistics, 4, 126–141.

Long, M. H. (1985). Input and second language acquisition theory. In S. Gass & C. Madden (Eds.) *Input in Second Language Acquisition* (pp. 377–393). Rowley, MA: Newbury House.

Long, M. H. (1990). Task, groups, and task-group interactions. In S. Anivan (Ed.), Language teaching methodology for the nineties. Singapore: RELC.

Long, M. H. (1991). Focus on Form: A Design Feature in Language Teaching Methodology. In K. De Bot, R. Ginsberg, & C. Kramsch (Eds.), Foreign Language Research in Cross-Cultural Perspectives (pp. 39-52). Amsterdam: John Benjamins.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413–468). New York: Academic Press.

Loschky, L. (1994). Comprehensible input and second language acquisition: What is the relationship? *Studies in Second Language Acquisition*, *16*(3), 303–323.

Macis, M. (2018). Incidental learning of duplex collocations from reading: Three case studies. *Reading in a Foreign Language*, *30*(1), 48.

Macis, M., & Schmitt, N. (2017). The figurative and polysemous nature of collocations and their place in ELT. *Elt Journal*, *71*(1), 50–59.

Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research*, *37*, 379–394.

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics*, *27*(3), 405–430.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition*, *22*(4), 471–497.

MacLennan, C. H. G. (1994). Metaphors and prototypes in the learning teaching of grammar and vocabulary. *IRAL-International Review of Applied Linguistics in Language Teaching*, *32*(2), 97–110.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104.

McCaig, I. R., Manser, M. H., & Warren, H. (1994). *Oxford Learner's Dictionary of English idioms*. Oxford: Oxford University Press.

McCarthy, M. (1990). *Vocabulary*. Oxford: Oxford University Press.

McCarthy, M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.

McCarthy, M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

McGregor, K. K., Oleson, J., Bahnsen, A., & Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language & Communication Disorders*, *48*(3), 307–319.

McLaughlin, B. (1987). *Theories of second-language learning*. London: Edward Arnold.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. In J. Tommola (Ed.), *Foreign language comprehension and production* (pp. 103–113). Turku: AFinLa Yearbook.

Meara, P. (1996). The Dimensions of Lexical Competence. In G. Brown, K. Malmkjaer, & W. J (Eds.), *Performance and Competence in Second Language Acquisition* (pp. 35–53). Cambridge: Cambridge University Press.

Meara, P. (2005). Lexical Frequency Profiles: A Monte Carlo Analysis. *Applied Linguistics*, *26*(1), 32–47.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing lexical characteristics of short L2 texts. *Prospect*, *16*(3), 5–19.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142–154.

Melka, F. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 84–102). Cambridge: Cambridge University Press.

Mennim, P. (2007). Long-term effects of noticing on oral output. *Language Teaching Research*, *11*(3), 265–280.

Miller, J. F. (1991). *Research on child language disorders: A decade of progress*. Austin, TX: Pro-ed.

Milton, J. & Fitzpatrick, T. (2014). *Dimensions of Vocabulary Knowledge*. Basingstoke: Palgrave Macmillan.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Mirzaee, A., & Ranjbar, M. (2012). The role of noticing in l2 learners' production of intonation patterns. *Journal of Teaching Language Skills*, *31*(1), 141–170.

Mokhtar, A. A. (2017). English lexical acquisition of adult learners in instructional settings: Issue on lexical input. *Proceedings of the ICECRS*, *1*(1).

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: the effect of two enhancement techniques. *Computer Assisted Language Learning*, *31*(1–2), 1–26.

Moon, R. (1998). Vocabulary connections: multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40–63). Cambridge: Cambridge University Press.

Moonen, M., de Graaff, R., Westhoff, G., & Brekelmans, M. (2014). The multi-feature hypothesis: Connectionist guidelines for L2 task design. *Language Teaching Research*, *18*(4), 474–496.

Morgan-Short, K., & Bowden, H. W. (2006). Processing instruction and meaningful output-based instruction: Effects on second language development. *Studies in Second Language Acquisition*, *28*(1), 31–65.

Morimoto, S., & Loewen, S. (2007). A comparison of the effects of image-schema-based instruction and translation-based instruction on the acquisition of L2 polysemous words. *Language Teaching Research*, *11*(3), 347–372.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Memory and Language*, *16*(5), 519.

Murphy, G. (2004). *The big book of concepts*. Cambridge, MA: MIT Press.

Nagy, W. (1995). On the role of context in first-and second-language vocabulary learning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 64–83). Cambridge: Cambridge University Press.

Nakata, T., & Webb, S. A. (2016). Vocabulary learning exercises: Evaluating a selection of exercises commonly featured in language learning materials. In B. Tomlinson (Ed.), *SLA research and materials development for language learning* (pp. 123–138). Oxon, England: Routledge.

Nassaji, H., & Hu, H. M. (2012). The relationship between task-induced involvement load and learning new words from context. *International Review of Applied Linguistics in Language Teaching*, *50*(1), 69–86.

Nassaji, H., & Tian, J. (2010). Collaborative and individual output tasks and their effects on learning English phrasal verbs. *Language Teaching Research*, *14*(4), 397–419.

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

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

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2005). Teaching and learning vocabulary. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 581–595). Mahwah, N.J.: Lawrence Erlbaum

Nation, I. S. P. (2007). The Four Strands. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 2–13.

Nation, I. S. P. (2013). *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.

Nation, I. S. P., & Meara, P. (2010). Vocabulary. In N. Schmitt (Ed.) *An introduction to applied linguistics* (pp. 34–52). London: Arnold.

Nation, I. S. P., & Newton, J. (2008). *Teaching ESL/EFL Listening and Speaking*. London: Taylor & Francis Group.

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

Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(2), 151–171.

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*. John Benjamins Amsterdam.

Neuman, W. L. (2003). *Social Research Methods: Qualitative and Quantitative Approaches* (5th Edition). London: Allyn and Bacon.

Neumann, S. (2013). *Contrastive register variation: A quantitative approach to the comparison of English and German* (Vol. 251). Berlin/Boston: Walter de Gruyter.

Newton, J., & Nation, I. S. P. (1997). Vocabulary and teaching. In J. Coady and T. Huckin (Eds.), Second Language Vocabulary Acquisition: A Rationale for Pedagogy (pp. 238-254). Cambridge: Cambridge University Press

Nice, M. M. (1925a). A child who would not talk. *The Pedagogical Seminary and Journal of Genetic Psychology*, *32*(1), 105–142.

Nice, M. M. (1925b). Length of sentences as a criterion of a child's progress in speech. *Journal of Educational Psychology*, *16*, 370–379.

Niu, R., & Helms-Park, R. (2014). Interaction, modality, and word engagement as factors in lexical learning in a Chinese context. *Language Teaching Research*, *18*(3), 345–372.

Nobuyoshi, J. & Ellis, R. (1993). Focused communication tasks and second language acquisition. *ELT Journal*, *47*(3), 203–210.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528.

Nunan, D. (1997). Designing and adapting materials to encourage learner autonomy. In P. Benson & P. Voller (Eds.), *Autonomy and Independence in Language Learning* (pp. 192–203). London: Longman.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*(2), 161–175.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217–237.

Omazic, M. (2008). Processing of idioms and idiom modifications: A view from cognitive linguistics. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 67–79). Amsterdam: John Benjamins.

Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech, Language, and Hearing Research*, *45*(5), 927–937.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 174–199). New York: Cambridge University Press.

Parker, M. D., & Brorson, K. (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language*, *25*(3), 365–376.

Pellicer-Sánchez, A. & Boers, F. (2019). Pedagogical approaches to the teaching and learning of formulaic language. In A. Siyanova-Chanturia, & A. Pellicer-Sánchez (Ed.), *Understanding Formulaic Language* (pp. 153–173). London: Routledge.

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.

Pellicer-Sánchez, A. (2020). Learning single words vs. multiword items. In S. A., Webb (Ed.), The Routledge handbook of vocabulary studies. London: Routledge.

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental Vocabulary Acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language*, *22*(1), 31–55.

Perdek, M. (2010). Getting through to phrasal verbs: A cognitive organization of phrasal verb entries in monolingual pedagogical dictionaries of English. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex international congress* (pp. 1390–1398). Ljouwert: Afûk.

Peters, E. & Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes, 20*, 28–39.

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention-drawing techniques. *The Language Learning Journal*, *40*(1), 65–79.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, *18*(1), 75–94.

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

Petersen, D. B., Gillam, S. L., Spencer, T., & Gillam, R. B. (2010). The effects of literate narrative intervention on children with neurologically based language impairments: An early stage study. *Journal of Speech, Language, and Hearing Research*, *53*(4), 961–981.

Philip, G. (2008). Reassessing the canon: 'Fixed'phrases in general reference corpora. In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective* (pp. 95–108). Amsterdam, Philadelphia: John Benjamins Publishing.

Philp, J., & Iwashita, N. (2013). Talking, tuning in and noticing: Exploring the benefits of output in task-based peer interaction. *Language Awareness*, *22*(4), 353–370.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9), 3526–3529.

Pica, T. (1988). Interlanguage adjustments as an outcome of NS-NNS negotiated interaction. *Language Learning*, *38*(1), 45–73.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes? *Language Learning*, *44*(3), 493–527.

Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, *11*(1), 63–90.

Pica, T., Holliday, L., Lewis, N., Berducci, D., & Newman, J. (1991). Language learning through interaction: What role does gender play? *Studies in Second Language Acquisition*, *13*(3), 343–376.

Pica, T., Lincoln-Porter, F., Paninos, D., & Linnell, J. (1996). Language learners' interaction: How does it address the input, output, and feedback needs of L2 learners? *TESOL Quarterly*, *30*(1), 59–84.

Pichette, F., De Serres, L., & Lafontaine, M. (2011). Sentence reading and writing for second language vocabulary acquisition. *Applied Linguistics*, *33*(1), 66–82.

Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition from extensive reading: A case study. *Reading in a Foreign Language*, *18*(1), 1–28.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912.

Polio, C., & Gass, S. M. (1998). The role of interaction in native speaker comprehension of nonnative speaker speech. *The Modern Language Journal*, *82*(3), 308–319.

Prabhu, N. S. (1987). *Second language pedagogy*. Oxford: Oxford University Press.

Prevoo, M. J. L., Malda, M., Mesman, J., Emmen, R. A. G., Yeniad, N., van Ijzendoorn, M. H., & Linting, M. (2014). Predicting ethnic minority children's vocabulary from socioeconomic status, maternal language and home reading input: different pathways for host and ethnic language. *Journal of Child Language*, *41*(5), 963–984.

Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, *17*(4), 409–441.

Qian, D. (1998). Depth of vocabulary knowledge, assessing its role in adults' reading comprehension in English as a second language. ProQuest Dissertations and Theses.

Qian, D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review*, *56*(2), 282–308.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536.

Qian, D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52.

Rassaei, E. (2017). Effects of three forms of reading-based output activity on L2 vocabulary learning. *Language Teaching Research*, *21*(1), 76–95.

Ratner, N. B., & Silverman, S. (2000). Parental perceptions of children's communicative development at stuttering onset. *Journal of Speech, Language, and Hearing Research*, *43*(5), 1252–1263.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing*, *10*(3), 355–371.

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

Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 209–227). Amsterdam: John Benjamins.

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, *53*(2), 333–349.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly*,10, 77–89.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning*, *61*, 1–36.

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

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to Learn: Conversation in Second Language Acquisition* (pp. 237–326): Rowley, MA: Newbury House.

Schmitt, N. (1998). Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language Learning*, *48*(2), 281–317.

Schmitt, N. (2000). *Vocabulary in language teaching*. Stuttgart, Germany: Ernst Klett Sprachen.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing, and use*. Amsterdam, Philadelphia: John Benjamins Publishing.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Houndmills, Basingstoke, Hampshire; New York: Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–951.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use* (pp. 1–22). Amsterdam; Philadelphia: John Benjamins Publishing.

Schmitt, N., & McCarthy, M. (1997). Vocabulary: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*(1), 17–36.

Schmitt, N., & Redwood, S. (2011). Learner knowledge of phrasal verbs: A corpus-informed study. In F. Meunier, S. de Cock, G. Gilquin & M. Paquot (Eds.), *A Taste for Corpora*. Amsterdam, Netherlands: John Benjamins.

Schmitt, N., Jiang, X. & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Seashore, R. H., & Eckerson, L. D. (1940). The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology*, *31*(1), 14.

Segalowitz, N., & Lightbown, P. M. (1999). Psycholinguistic approaches to SLA. *Annual Review of Applied Linguistics*, *19*, 43–63.

Shefelbine, J. L. (1990). Student factors related to variability in learning word meanings from context. *Journal of Reading Behavior*, *22*(1), 71–97.

Shehadeh, A. (2002). Comprehensible output, from occurrence to acquisition: An agenda for acquisitional research. *Language Learning*, *52*(3), 597–647.

Shehadeh, A. (2003). Learner output, hypothesis testing, and internalizing linguistic knowledge. *System*, *31*(2), 155–171.

Shekary, M., & Tahririan, M. H. (2006). Negotiation of meaning and noticing in text-based online chat. *The Modern Language Journal*, *90*(4), 557–573.

Shepherd, N. (2009). A reply to Konrad Brodzinski. *Modern English Teacher*, *18*(3), 14–15.

Shintani, N. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research*, *15*(2), 137–158.

Shortall, T. (2002). Teaching Grammar. *Developing Language Professionals in Higher Education Institutions (DELPHI) Project*. Available online at http:// www.delphi.bham.ac.uk/modules.htm.

Sidtis, D. V. L. (2012). Two-track mind: Formulaic and novel language support a dual-process model. In M. Faust (Eds.) *Advances in the neural substrates of language: Toward a synthesis of basic science and clinical research* (pp. 342–367). London: Blackwell Publishing.

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly*, *37*(3), 419–441.

Sinclair, B., & Ellis, G. (1984). Autonomy begins in the classroom. *Modern English Teacher*, *11*(4), 45–47.

Sinclair, J. M. (1996). The empty lexicon. *International Journal of Corpus Linguistics*, *1*(1), 99–119.

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (2019). Formulaic language: Setting the scene. In A. Siyanova-Chanturia, & A. Pellicer-Sánchez (Ed.), *Understanding Formulaic Language* (pp. 1–15). London: Routledge.

Siyanova-Chanturia, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(2), 119–139.

Siyanova-Chanturia, A., & Sidtis, D. V. L. (2019). What online processing tells us about formulaic language. In A. Siyanova-Chanturia, & A. Pellicer-Sánchez (Ed.), *Understanding Formulaic Language* (pp. 38–61). London: Routledge.

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

Soars, L., & Soars, J. (2003). *New Headway – Pre-Intermediate*. Oxford: Oxford University Press.

Sökmen, A. J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 237–258). Cambridge: Cambridge University Press.

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

Spada, N., & Lightbown, P. M. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition*, *15*(2), 205–224.

Stahl, K. A. D., & Bravo, M. A. (2010). Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher*, *63*(7), 566–578.

Stengers, H., Boers, F., Housen, A. & Eyckmans, J. (2011). Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? *IRAL-International Review of Applied Linguistics in Language Teaching*, *49*(4), 321–343.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly*, *11*(3), 271–282.

Storch, N. (2008). Metatalk in a pair work activity: Level of engagement and implications for language development. *Language Awareness*, *17*(2), 95–114.

Stubbs, M. (1986). Lexical density: A technique and some findings. In M. Coultard (Ed.), *Talking About Text: Studies presented to David Brazil on his retirement, Discourse Analysis Monograph* (pp. 27–48). Birmingham: English Language Research, University of Birmingham

Sun, C. H. (2017). The value of picture-book reading-based collaborative output activities for vocabulary retention. *Language Teaching Research*, *21*(1), 96–117.

Sun, Y., & Wang, L. (2003). Concordances in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer-Assisted Language Learning*, 16, 83–94.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in Second Language Acquisition*, *15*, 165–179.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford University Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471–483). Mahwah, N.J.: Lawrence Erlbaum.

Swain, M., & Lapkin, S. (1982). *Evaluating Bilingual Education: A Canadian Case Study*. Clevedon, UK: Multilingual Matters..

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(3), 371–391.

Swain, M., & Lapkin, S. (1998). Interaction and second language learning: Two adolescent French immersion students working together. *The Modern Language Journal*, *82*(3), 320–337.

Swain, M., & Lapkin, S. (2001). Focus on form through collaborative dialogue: Exploring task effects. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks, second language learning and testing* (pp. 99- 118). Harlow: Longman.

Szudarski, P. (2012). Effects of meaning-and form-focused instruction on the acquisition of verb-noun collocations in L2 English. *Journal of Second Language Teaching & Research*, *1*(2), 3–37.

Szudarski, P., & Carter, R. (2016). The role of input flood and input enhancement in EFL learners' acquisition of collocations. *International Journal of Applied Linguistics, 26*(2), 245–265.

Tahmasbi, M. & Farvardin, M. T. (2017). Probing the Effects of Task Types on EFL Learners' Receptive and Productive Vocabulary Knowledge: The Case of Involvement Load Hypothesis. *SAGE Open*, *7*(3), 2158244017730596.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science*, *12*(1), 49–100.

Tannenbaum, K. R., Torgesen, J. K., & Wagner, R. K. (2006). Relationships between word knowledge and reading comprehension in third-grade children. *Scientific Studies of Reading*, *10*(4), 381–398.

Text Inspector (2015, 2017) *Online lexis analysis tool* at textinspector.com

Thordardottir, E. T., & Namazi, M. (2007). Specific language impairment in French-speaking children: Beyond grammatical morphology. *Journal of Speech, Language, and Hearing Research*, *50*(3), 698–715.

Thornbury, S. (2002). *How to teach vocabulary*. London: Longman.

Tomlin, R. S., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition*, *16*(2), 183–203.

Toomer, M., & Elgort, I. (2019). The Development of Implicit and Explicit Knowledge of Collocations: A Conceptual Replication and Extension of Sonbul and Schmitt (2013). *Language Learning*, *69*(2), 405–439.

Treffers-Daller, J. (2013). *Measuring lexical diversity among L2 learners of French*. Amsterdam, Philadelphia: John Benjamins.

Tuggy, D. (1993). Ambiguity, polysemy, and vagueness. *Cognitive Linguistics (Includes Cognitive Linguistic Bibliography)*, *4*(3), 273–290.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*(5), 323–352.

Uchihara, T., & Clenton, J. (2018). Investigating the role of vocabulary size in second language speaking ability. *Language Teaching Research*.

Ure, J. (1971). Lexical density and register differentiation. In G. Perren & J. L. Trim (Eds.), *Applications of linguistics* (Vol. 452, pp. 443–452). London: Cambridge University Press.

Van den Branden, K. (1997). Effects of negotiation on language learners' output. *Language Learning*, *47*(4), 589–636.

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

Vande Kopple, W. J. (2003). MAK Halliday's continuum of prose styles and the stylistic analysis of scientific texts. *Style*, *37*(4), 367–381.

VanPatten, B. (2003). *From input to output: A teacher's guide to second language acquisition*. New York: McGraw-Hill.

VanPatten, B., & Williams, J. (2014). *Theories in second language acquisition: An introduction*. Routledge.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, *22*(2), 217–234.

Verspoor, M. & Lowie, W. (2003). Making sense of polysemous words. *Language Learning*, *53*(3), 547–586.

Waring, R. , & Nation, I. S. P. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge: Cambridge University Press.

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research*, *11*(2), 121–142.

Webb, S. A. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52.

Webb, S. A. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.

Webb, S. A. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95.

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

Webb, S. A., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.

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

Weinreich, U., & Webster. (1964). Webster's third: A critique of its semantics. *International Journal of American Linguistics*, *30*(4), 405–409.

Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review*, *53*(1), 13–40.

Williams, J. (1999). Learner-generated attention to form. *Language Learning*, *49*(4), 583–625.

Williams, J. (2005). Form-focused instruction. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 695–716). Mahwah, N.J.: Lawrence Erlbaum.

Wilson, R. A., & Beinke, L. W. (1979). *Applications of graph theory*. London: Academic Press.

Wode, H. (1999). Incidental vocabulary acquisition in the foreign language classroom. *Studies in Second Language Acquisition*, *21*(2), 243–258.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review*, *63*(1), 13–33.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2019). Concluding Question: Why Don't Second Language Learners More Proactively Target Formulaic Sequences? In A. Siyanova-Chanturia, & A. Pellicer-Sánchez (Ed.), *Understanding Formulaic Language* (pp. 260–281). London: Routledge.

Xiao, R., & Yue, M. (2009). Using corpora in Translation Studies: the state of the art. In P. Baker (ed.), *Contemporary Corpus Linguistics* (pp. 237–262). London: Continuum.

Yarowsky, D. (1993). One sense per collocation. In *Proceedings of the workshop on Human Language Technology* (pp. 266–271). Association for Computational Linguistics.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. *Pragmatics and beyond New Series*, 29–46.

Yee, Ning, & Hua. (2017). Chapter 14- the interaction hypothesis: Why you shouldn't learn languages alone | Second Language Acquisition.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*(2), 236–259.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595.

Zeng, G., & Takatsuka, S. (2009). Text-based peer–peer collaborative dialogue in a computer-mediated learning environment in the EFL context. *System*, *37*(3), 434–446.

Zhang, X. (2017). Effects of receptive-productive integration tasks and prior knowledge of component words on L2 collocation development. *System*, *66*, 156–167.

Zora, S., & Johns-Lewis, C. (1989). Lexical density in interview and conversation. *York Papers in Linguistics*, 14, 89-100.

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, *21*(1), 54–75.

Appendices

Appendix 1

Study 1

List of target SWVs used in the study

	Target item	BNC Freq.	Freq. COCA	English definition	Arabic definition	Coverage %
Week 1	1. play	20702	286559	To amuse oneself	يتسلى	31
				To engage in sport	يشترك بالرياضة	28
				To produce music	يعزف	17
	2. charge	11664	75880	To pay some money	يحاسب	25.5
				To command someone with a duty	يكفل شخصاً بمهمة	17.5
	3. serve	5247	77318	To accuse formally	يتهم	10.5
				To work for	يخدم	42.2
				To provide food or drink	(نادل) يقدم الطعام	18.5
				To fulfil an obligation	يؤدي خدمة	15.5
	4. cover	10763	104236	To wrap something	يغلف	36
				To include	يشمل	33
				To be enough	يغطي التكاليف	27
	5. divide	1732	15914	To split something	يوزع	42
				To find how many times a number contains another (calculation)	(حساب) يقسم	28
				To disagree	يعارض	25
	6. commit	1318	19840	To do something illegal or wrong	يرتكب	44.5
				To obligate oneself to do something	يتعهد	22
				To put something to (a state or place) for keeping	يحتفظ ب	14
Week 2	7. run	21170	264022	To move quickly	يركض	31
				To manage a business	يدير شغلاً	29.5
				To flow(Liquid)	يصب	19
	8. blow	3093	37559	To spend carelessly	يبدد المال	45
				To send an out air (of a person's lips)	يزفر	22
				To lose the chance to do something	يفسد أمراً ما	10
	9. train	7782	58921	To teach someone how to do something	يدرب /يعلم	41
				To exercise	(رياضة) يتدرب	28.5
				To point something	يصوب	14
	10. hit	9441	175122	To touch quickly and with force	يصدم	37.5

			To affect badly	يسبب عنأة	37
			To reach a place, position or state	يصل لمكان	12.5
11. relate	2555	16430	To show a connection	يربط	24
			To feel sympathy with	يستوعب /يتفهم	20
			To tell the story of something	يروي /يحكي	17.5
12. order	33646	222669	To request food, things, etc.	يطلب	48
			To give instructions to someone	يأمر	21.5
			To arrange in a list	ترتيب	14
13. fall	1513	137328	To move down quickly	يسقط	27
			To become less	يصبح اقل	24.5
			To pass into a state or situation	يصاب بشئ	23.5
14. clear	24762	195798	To remove people or things from a place	ينظف /يصفي	27
			To pass without touching	يمر بدون لمس	22.5
			To free from guilt	يتحرر من الأتهام	20.5
15. address	6679	93061	To write details	يكتب عنوان	29
			To deal with a problem	يعالج مشكلة	25.5
			To speak to someone formally	يخاطب برسمية	21
16. lead	14192	143108	To be in control of a group	يكون مسئول	26
			To show someone where to go	يرشد الى الطريق	21
			To be an access means to a place	يقود الى	20.5
17. introduce	3422	19911	To put something into use for the first time	يقدم شيئ للمرة الأولى	54
			To present a person to others	يقدم شخص للاخرين	14
			To inject /push something	يُدخل	11.5
18. lift	4116	27558	To put in a higher position or level	يرفع	44
			To formally end a rule	ينهي قاعدة	28
			To take something to a different position	يغير مكان	15.5

Week 3

Week 4	19. save	1429	102509	To remove from danger	ينقذ	23
				To avoid the need to spend (money, time)	يوفر	19
				To keep and store for future use	يدخر	18
	20. treat	3632	44047	To handle in a certain way	عامل	40
				To give medical care	يعالج	33.5
				To put material to something to protect it	يغطي لحمايته	11.5
	21. enter	5052	46732	To come into a place	يدخل مكان	23.5
				To be involved in a group/join a group	يلتحق بمجموعة	21
				To write information into a computer	يدخل للكمبيوتر /يكتب	16
	22. collect	2738	22385	To bring together (as hobby)	يجمع	30.5
				To fetch someone/something from a place	يحضر شيء او شخص من مكان ما	27.5
				To get control of oneself after a shock	يلملم نفسه	25.5
	23. count	4049	57617	To see how many people/things are there	يعدُ	24.5
				To consider someone/something	يعتبر	17.5
				To be important	يتخذ أهمية	10
	24. settle	2455	23235	To resolve a problem	جزم/حسم	54
				To start living in a place	يستوطن/يستقر	12.5
				To sit in a comfortable way	يجلس مرتاحاً	12

List of target PVs used in the study

	Target item	BNC Freq.	Freq. COCA	English definition	Arabic definition	Coverage %
Week 1	1. lay down	622	4996	To put something away/down	يضع جانباً	31
				To lie flat on a surface	أرْقد	28
				To officially make new rules	يفرض القوانين	17
	2. take up	1757	8637	To use space or time	يستغل، يستغرق	25.5
				To discuss something (issue, idea, matter) with someone	يناقش موضوعاً	17.5
				To start doing a job or activity for pleasure	يبدأ بعمل ما للمتعة	10.5
	3. make up	2330	23158	To form the whole of an amount	يرّكب/يجّمع	42.2
				To compensate for something bad or lost (+for)	يعوّض	18.5
				To decide something	يقرّر	15.5
	4. bring out	389	3764	To make noticeable	يبرز/يثير الانتباه ل	36
				to make available to people to see, know or buy	يعُلن/ ينشر	33
				To take something out of a container	يُخرج	27
	5. move on	1131	20157	To start doing something new	يلتحق بعمل جديد	42
				To change physical location (spot, room, country)	يتنقل (يغير مكانه)	28
				To forget about a difficult experience and move forward	يواصل بعد تجربة صعبة	25
	6. go off	668	6336	To go somewhere for a particular purpose	انطلق	44.5
				To start making loud noise	يرّن (يصدر صوت مزعج)	22
				To explode or fires	انفجر	14
Week 2	7. fill in	763	5063	To substitute for someone (+ for)	يحل محل	31
				To give extra information *(+ on)*	*يعطي معلومات اضافية*	29.5
				To put material into something to make it full	يملأ	19

8.	come around	61	3405	To come near	يعرّج على	45
				To change one's opinion to agree with another's	يعيد النظر في أمر	22
				To happen again as a regular event	يتكرر، يعاد	10
9.	take off	673	10785	To remove clothes, shoes, jewelry…etc.	يخلع ملابسه	41
				To suddenly leave without telling	يغادر فجأة	28.5
				To leave the ground and fly	تقلع (الطائرة)	14
10.	pass on	567	4476	To give STH to SB after receiving it	تناقل الشيء	37.5
				To transmit between generations	يورث	37
				To die	يتوفى	12.5
11.	break down	627	5977	To stop working (machines)	يتعطل، يتوقف عن التشغيل	24
				To separate into smaller parts	يقّسم/يجّزأ	20
				To become very upset and start crying	ينهارلدرجة البكاء	17.5
12.	turn up	1071	4144	To find something	يجد/ يعثر	48
				To increase the output of something	يزيد انتاج/يرفع	21.5
				To arrive somewhere	جاء/حضر	14
13.	cut off	1113	12733	Remove a part of something by cutting it	يقطع بشيء حاد)سكين(	27
				To interrupt someone while speaking	يقاطع	24.5
				To stop providing something	يوقّف امداده	23.5
14.	get through	816	12228	To succeed in reaching a physical destination or stage (+ to)	يجتاز	27
				To successfully make someone understand something	يتفاهم مع	22.5

Week 3

			To reach someone by telephone	يتواصل هاتفياً مع	20.5
	15. throw out	99	2624	To refuse to accept	يرفض	29
				To get rid of something	يتخلص من/يرمي	25.5
				To force someone to leave	يطرد	21
	16. back up	953	18664	To move or drive backward	يتقهقر)يعود للوراء(	26
				To support or help someone	يساعد/يساند	21
				To prove that something is true	يدعم/يناصر	20.5
	17. hold up	228	6775	To raise something in a high position	يرفع	54
				To remain strong or successful	يتحمل/يصبر	14
				To delay	يؤجل	11.5
	18. go along	750	6393	To make progress with something	يتقدم، يحرز تقدماً	44
				To agree with someone's opinion	يوافق/ يقبل/ يساير	28
				To go to a place without planning	يمر بمكان بدون تخطيط	15.5
	19. put up	1909	16272	To attach something to a wall or ceiling	يُعَلِّق/ يَنْصب	23
				To accept something unpleasant (+ *with*)	يرضى/يتحمل	19
				To build something	يبني	18
	20. give out	134	2020	To distribute something to people	يوزع، ينشر	40
				To make something public	يُصدر/يُصرح	33.5
				To stop working (knees, heart...etc.)	يتعطّل :عضو جسدي	11.5
Week 4	21. hold back	201	2459	To decide not to do or say something	يتحفظ	23.5
				To prevent from making progress	يمنعه من التقدم/ يقيّد	21
				To stop showing an emotion (tears, laughter)	يكبت نفسه/يكبح مشاعره	16
	22. set off	1584	6969	To start a journey	ينطلق في رحلة	30.5
				To cause something to explode	يُفجّر القنبلة	27.5
				To make something begin or happen	يبدأ	25.5
	23. take in	708	5998	To let someone stay in your house	يسمح له بالبقاء	24.5
				To fully understand something	يستوعب، يفهم	17.5

			To deceive (cheat) someone	يخدع/يغش/يحتال	10
24. get off	851	18175	To leave (train, bus, airplane, lift)	ينزل من القطار أو السيارة إلخ	54
			To begin something successfully	يبدا بداية جيدة	12.5
			To avoid being punished for something wrong	يتجنب عقوبة	12

Appendix 2

Study 1

Practice worksheets for traditional group **Session 1**

1- Match the verb on the left column with its complement on the right column:

_____	1. to lay down	A. to a new place
_____	2. to serve	B. in the garden
_____	3. to move on	C. him with theft
_____	4. to play	D. your pen
_____	5. to charge	E. to France
_____	6. to go off	F. lunch

_____	7. to lay down	A. football
_____	8. to serve	B. with cleaning the dishes
_____	9. to move on	C. in the army
_____	10. to play	D. and wake me up
_____	11. to charge	E. on the sofa
_____	12. to go off	F. and forget about it

_____	13. to lay down	A. and kill people
_____	14. to serve	B. 200 Riyals
_____	15. to move on	C. the law
_____	16. to play	D. to the next question
_____	17. to charge	E. as an office as well
_____	18. to go off	F. the piano

2- Fill in blanks with a word from the box below. Remember that each word in the box can be used 3 times (*3):

lay down (*3)	serve (*3)	move on (*3)
play (*3)	go off (*3)	charge (*3)

1. You should …….. some plastic on the ground when you paint a house.
2. You need to forget about your past; just …….. and don't look back anymore.
3. When the alarm ……..., it was hard to get up because I went to bed so late last night.
4. They will …….. to new job where the salary is better.
5. The police may ……...the man with a crime.
6. The office may also …….. as a guest bedroom.
7. The gun …….. accidentally during a struggle and the bullet hit his heart.
8. She likes …….. with her dolls and talking with them a lot.
9. On radio FM they …….. Egyptian music as well as some talk radio.
10. My parents …….. us with looking after the house while they go out.
11. My dad …….. as a general for 5 years.
12. It's time to …….. to another town where we can get a fresh start.
13. If you ordered a take-away, they will …….. you for delivery.
14. He may leave Rabigh and …….. to Jeddah to look for work.
15. He …….. on the bed and tried to sleep.
16. He had already clearly …….. his principles in his opening speech.
17. Barcelona is …….. against Real Madrid tonight.
18. All dishes in this restaurant are …….. with green salad.

1. Match the verb on the left column with its complement on the right column

_____	1. to cover	A. to a state prison
_____	2. to make up	B. photography
_____	3. to divide	C. his face
_____	4. to commit	D. my mind whether to go
_____	5. to bring out	E. the new album
_____	6. to take up	F. into two semesters

_____	7. to cover	A. the fancy dishes
_____	8. to make up	B. a broad range of subjects
_____	9. to divide	C. for my behaviour
_____	10. to commit	D. the matter with the police
_____	11. to bring out	E. himself to charity
_____	12. to take up	F. the country

_____	13. to cover	A. a murder
_____	14. to make up	B. the colour of your eyes
_____	15. to divide	C. 60 per cent of the sales
_____	16. to commit	D. half of the room
_____	17. to bring out	E. by six is five
_____	18. to take up	F. your tuition fees

2. Fill in blanks with a word from the box below. Remember that each word in the box can be used 3 times (*3):

make up (*3)	cover (*3)	bring out (*3)
divide (*3)	take up (*3)	commit (*3)

1. 6000 Riyals may …….... the cost of the repairs that are needed to the car after the accident.
2. Apple may …….... a new iPhone in March 2016.
3. I …….... myself to working this project for the next two months
4. I forget Sara's birthday yesterday, so I'll try to …….... to her by taking her to a nice place.
5. Let's …….... this problem with the manager.
6. Now, can you …….... this sum by five?
7. You should …….... your mind about the dinner on Thursday.
8. She was …….... to a mental hospital
9. My sister has recently …….... sewing as a hobby.
10. New Headway book is …….... into 12 units.
11. Salt can help to …….... the flavour of food.
12. The study aims to find out what makes people …….... violent acts and kill innocents.
13. They …….... another plate from the kitchen for the new guests.
14. They …….... him with a blanket.
15. This issue may …….... the government into small conflicting groups.
16. Those big desks …….... too much space.
17. Those books …….... Saudi history since king Abdul-Aziz era.
18. Women …….... nearly 50% of law school students.

Note. Only samples are included here. The full materials are available in the CD.

Practice worksheets for pushed-output group **Session 1**

Student A (Worksheet)

1. Match the verb on the left column with its complement on the right column and say it to your partner:

	Verb	Complement
.....	1. to lay down	A. to a new place
.....	2. to serve	B. in the garden
.....	3. to move on	C. him with theft
.....	4. to play	D. your pen
.....	5. to charge	E. to France
.....	6. to go off	F. lunch

.....	7. to lay down	A. football
.....	8. to serve	B. with cleaning the dishes
.....	9. to move on	C. in the army
.....	10.to play	D. and wake me up
.....	11.to charge	E. on the sofa
.....	12.to go off	F. and forget about it

.....	13.to lay down	A. and kill people
.....	14.to serve	B. 200 Riyals
.....	15.to move on	C. the law
.....	16.to play	D. to the next question
.....	17.to charge	E. as an office as well
.....	18.to go off	F. the piano

1

Student A (Worksheet)

2. Here are some examples for the verb and its complement in question (1) but scrambled. Please arrange the words in the correct order to make sentences and say it to your partner:

1. your pen/ stop writing and / , please / lay down
2. after 6 pm/ most hotels/ lunch/ don't serve
3. moved on/ just today/ all the family/ to a new place
4. in the garden/ the children/ for an hour/ were playing
5. charged/ and sent him to prison/ him with theft/ the court
6. to France/ Ahmad/ for the summer/ may go off
7. lay down/ I'm going to/ because I'm not feeling well/ on the sofa
8. in the army/ my dad/ for 20 years/ served
9. move on/ I just want to/ about this nightmare/ and forget
10. did you/ football/ in high school? / play
11. charged/ me with cleaning the dishes/ my mom/ after dinner
12. and wake me up/ I was laying in bed waiting for/ go off/ the alarm to
13. the law/ the teacher/ laid down/ on the first day
14. can serve/ as well for my dad/ as an office/ the living room
15. let's/ to the next question/ move on/ , please
16. the piano/ he can/ very well/ play
17. charged/ the restaurant/ for dinner/ 200 Riyals
18. the bomb/ in Paris/ and killed many people/ went off

2

Student A (Worksheet)

3. Now, use the verb with a new complement and say it to your partner:

1. Lay down: .....
2. Lay down: .....
3. Lay down: .....
4. Play: .....
5. Play: .....
6. Play: .....
7. Go off: .....
8. Go off: .....
9. Go off: .....
10. Serve: .....
11. Serve: .....
12. Serve: .....
13. Move on: .....
14. Move on: .....
15. Move on: .....
16. Charge: .....
17. Charge: .....
18. Charge: .....

3

Student A (Answers for B)

1. Matching activity

	Verb
B	1. to lay down
F	2. to serve
A	3. to move on
C	4. to play
D	5. to charge
E	6. to go off

D	7. to lay down
A	8. to serve
F	9. to move on
E	10.to play
C	11.to charge
B	12.to go off

E	13.to lay down
D	14.to serve
A	15.to move on
F	16.to play
B	17.to charge
C	18.to go off

4

Student A (Answers for B)

2. Sentences reordering

1. He had already laid down his principles in his opening speech.
2. All dishes in this restaurant are served with green salad.
3. It's time to move on to another town where we can get a fresh start.
4. They play Egyptian music on radio FM.
5. If you ordered a take-away, they will charge you for delivery.
6. When the alarm went off, it was hard to get up.

7. My uncle laid down on the bed and tried to sleep.
8. The office may also serve as a guest bedroom.
9. You need to forget about your past; move on and don't look back anymore.
10. Barcelona is playing against Real Madrid tonight.
11. My parents charge us with looking after the house while they go out.
12. He may leave Rabigh and go off to Jeddah to look for work.

13. You should lay down some plastic on the ground when you paint a house.
14. My dad served as a general for 5 years.
15. The employees will move on to new jobs where the salary is better.
16. She likes playing with her dolls and talking with them a lot.
17. The police may charge the man with a crime.
18. The gun went off accidentally during a struggle.

5

Student A (Answers for B)

3. Sentences (verbs with new complements)

1. Lay down	☐ Yes	☐ No
2. Lay down	☐ Yes	☐ No
3. Lay down	☐ Yes	☐ No
4. Play	☐ Yes	☐ No
5. Play	☐ Yes	☐ No
6. Play	☐ Yes	☐ No
7. Go off	☐ Yes	☐ No
8. Go off	☐ Yes	☐ No
9. Go off	☐ Yes	☐ No
10.Serve	☐ Yes	☐ No
11.Serve	☐ Yes	☐ No
12.Serve	☐ Yes	☐ No
13.Move on	☐ Yes	☐ No
14.Move on	☐ Yes	☐ No
15.Move on	☐ Yes	☐ No
16.Charge	☐ Yes	☐ No
17.Charge	☐ Yes	☐ No
18.Charge	☐ Yes	☐ No

6

Student B (Worksheet)

1. Match the verb on the left column with its complement on the right column and say it to your partner:

	Verb	Complement
.....	1. to lay down	A. to another town where we can get fresh start
.....	2. to serve	B. his principles
.....	3. to move on	C. Egyptian music
.....	4. to play	D. for delivery
.....	5. to charge	E. this morning, it was so hard to get up
.....	6. to go off	F. with salad

.....	7. to lay down	A. as a guest bedroom
.....	8. to serve	B. to Jeddah to work
.....	9. to move on	C. with looking after the house
.....	10.to play	D. on the bed
.....	11.to charge	E. against Real Madrid
.....	12.to go off	F. and don't look back

.....	13.to lay down	A. to a new job
.....	14.to serve	B. the man with a crime
.....	15.to move on	C. accidentally during a struggle
.....	16.to play	D. as a General
.....	17.to charge	E. some plastic on the ground
.....	18.to go off	F. with her dolls

1

Student B (Worksheet)

2. Here are some examples for the verb and its complement in question (1) but scrambled. Please arrange the words in the correct order to make sentences and say it to your partner:

1. in his opening speech/ laid down/ he had already/ his principles
2. are served/ in this restaurant/ all dishes/ with green salad
3. where we can get a fresh start/ it's time to/ to another town/ move on
4. on radio FM/ play/ they/ some Egyptian music
5. you for delivery/ they/ if you ordered a take-away,/ will charge
6. went off,/ to get up/ when the alarm/ it was hard
7. my uncle/ and tried to sleep/ on the bed/ laid down
8. may also/ the office/ as a guest bedroom/ serve
9. move on/ about your past;/ you need to forget/ and don't look back anymore
10. against Real Madrid/ Barcelona/ tonight/ is playing
11. charge/ my parents/ while they go out/ us with looking after the house
12. and go off/ to look for work/ he may leave Rabigh/ to Jeddah
13. some plastic on the ground/ you should/ when you paint a house/ lay down
14. served/ as a general/ my dad/ for 5 years
15. to new jobs/ will move on/ where the salary is better/ the employees
16. playing/ and talking with them a lot/ with her dolls/ she likes
17. the man/ may charge/ the police/ with a crime
18. accidentally/ the gun/ during a struggle/ went off

2

Student B (Worksheet)

3. Now, use the verb with a new complement and say it to your partner:

1. Lay down: .....
2. Lay down: .....
3. Lay down: .....
4. Play: .....
5. Play: .....
6. Play: .....
7. Go off: .....
8. Go off: .....
9. Go off: .....
10. Serve: .....
11. Serve: .....
12. Serve: .....
13. Move on: .....
14. Move on: .....
15. Move on: .....
16. Charge: .....
17. Charge: .....
18. Charge: .....

3

4. Matching activity

	Verb
D	1. to lay down
F	2. to serve
A	3. to move on
B	4. to play
C	5. to charge
E	6. to go off

E	7. to lay down
C	8. to serve
F	9. to move on
A	10.to play
B	11.to charge
D	12.to go off

C	13.to lay down
E	14.to serve
D	15.to move on
F	16.to play
B	17.to charge
A	18.to go off

5. Sentences Reording

1. Stop writing and lay down your pen, please.
2. Most hotels don't serve lunch after 6 pm.
3. All the family moved on to a new place just today.
4. The children were playing in the garden for an hour.
5. The court charged him with theft and sent him to prison.
6. Ahmad may go off to France for the summer.

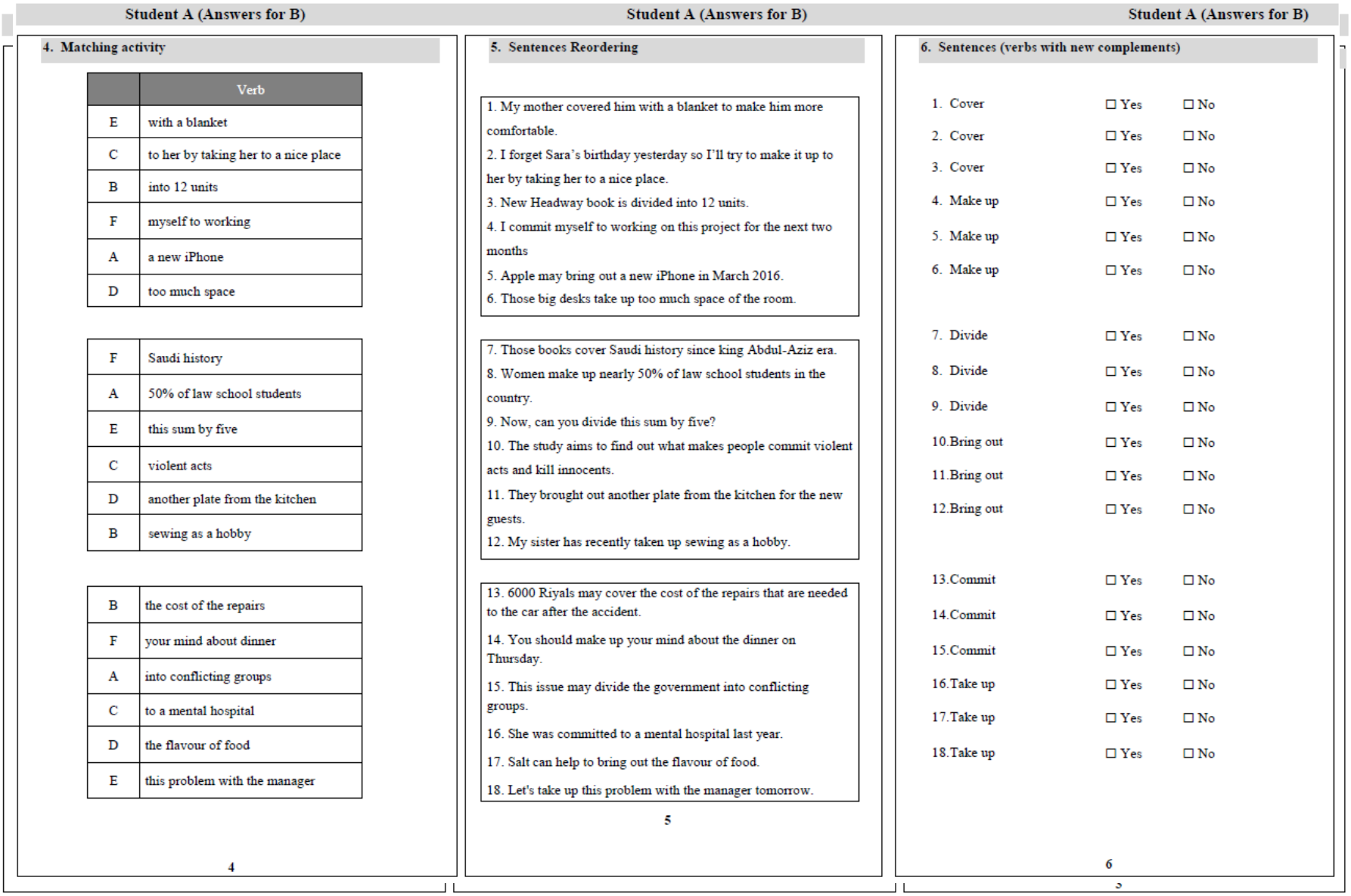
7. I'm going to lay down on the sofa because I'm not feeling well.
8. My dad served in the army for 20 years.
9. I just want to move on and forget about this nightmare.
10. Did you play football in high school?
11. My mom charged me with cleaning the dishes after dinner.
12. I was laying in bed waiting for the alarm to go off and wake me up.

13. The teacher laid down the law on the first day.
14. The living room can serve as an office as well for my dad.
15. Let's move on to the next question, please.
16. He can play the piano very well.
17. The restaurant charged 200 Riyals for dinner.
18. The bomb went off and killed many people in Paris.

6. Sentences (verbs with new complements)

1. Lay down	☐ Yes	☐ No
2. Lay down	☐ Yes	☐ No
3. Lay down	☐ Yes	☐ No
4. Play	☐ Yes	☐ No
5. Play	☐ Yes	☐ No
6. Play	☐ Yes	☐ No
7. Go off	☐ Yes	☐ No
8. Go off	☐ Yes	☐ No
9. Go off	☐ Yes	☐ No
10.Serve	☐ Yes	☐ No
11.Serve	☐ Yes	☐ No
12.Serve	☐ Yes	☐ No
13.Move on	☐ Yes	☐ No
14.Move on	☐ Yes	☐ No
15.Move on	☐ Yes	☐ No
16.Charge	☐ Yes	☐ No
17.Charge	☐ Yes	☐ No
18.Charge	☐ Yes	☐ No

Practice worksheets for pushed-output group **Session 2**



Student A (Answers for B)

4. Matching activity

	Verb
E	with a blanket
C	to her by taking her to a nice place
B	into 12 units
F	myself to working
A	a new iPhone
D	too much space

F	Saudi history
A	50% of law school students
E	this sum by five
C	violent acts
D	another plate from the kitchen
B	sewing as a hobby

B	the cost of the repairs
F	your mind about dinner
A	into conflicting groups
C	to a mental hospital
D	the flavour of food
E	this problem with the manager

4

Student A (Answers for B)

5. Sentences Reordering

1. My mother covered him with a blanket to make him more comfortable.
2. I forget Sara's birthday yesterday so I'll try to make it up to her by taking her to a nice place.
3. New Headway book is divided into 12 units.
4. I commit myself to working on this project for the next two months
5. Apple may bring out a new iPhone in March 2016.
6. Those big desks take up too much space of the room.

7. Those books cover Saudi history since king Abdul-Aziz era.
8. Women make up nearly 50% of law school students in the country.
9. Now, can you divide this sum by five?
10. The study aims to find out what makes people commit violent acts and kill innocents.
11. They brought out another plate from the kitchen for the new guests.
12. My sister has recently taken up sewing as a hobby.

13. 6000 Riyals may cover the cost of the repairs that are needed to the car after the accident.
14. You should make up your mind about the dinner on Thursday.
15. This issue may divide the government into conflicting groups.
16. She was committed to a mental hospital last year.
17. Salt can help to bring out the flavour of food.
18. Let's take up this problem with the manager tomorrow.

5

Student A (Answers for B)

6. Sentences (verbs with new complements)

1. Cover	☐ Yes	☐ No
2. Cover	☐ Yes	☐ No
3. Cover	☐ Yes	☐ No
4. Make up	☐ Yes	☐ No
5. Make up	☐ Yes	☐ No
6. Make up	☐ Yes	☐ No
7. Divide	☐ Yes	☐ No
8. Divide	☐ Yes	☐ No
9. Divide	☐ Yes	☐ No
10.Bring out	☐ Yes	☐ No
11.Bring out	☐ Yes	☐ No
12.Bring out	☐ Yes	☐ No
13.Commit	☐ Yes	☐ No
14.Commit	☐ Yes	☐ No
15.Commit	☐ Yes	☐ No
16.Take up	☐ Yes	☐ No
17.Take up	☐ Yes	☐ No
18.Take up	☐ Yes	☐ No

6

Student B (Worksheet)

1. Match the verb on the left column with its complement on the right column and say it to your partner:

	Verb	Complement
.....	1. to cover	A. a new iPhone
.....	2. to make up	B. into 12 units
.....	3. to divide	C. to her by taking her to a nice place
.....	4. to commit	D. too much space
.....	5. to bring out	E. with a blanket
.....	6. to take up	F. myself to working

.....	7. to cover	A. 50% of law school students
.....	8. to make up	B. sewing as a hobby
.....	9. to divide	C. violent acts
.....	10.to commit	D. another plate from the kitchen
.....	11.to bring out	E. this sum by five
.....	12.to take up	F. Saudi history

.....	13.to cover	A. government into conflicting groups
.....	14.to make up	B. the cost of the repairs
.....	15.to divide	C. to a mental hospital
.....	16.to commit	D. the flavour of food
.....	17.to bring out	E. this problem with the manager
.....	18.to take up	F. your mind about dinner

1

Student B (Worksheet)

2. Here are some examples for the verb and its complement in question (1) but scrambled. Please arrange the words in the correct order to make sentences and say it to your partner:

1. covered/ him with a blanket / to make him more comfortable/ my mother
2. so I'll try to/ to her by taking her to a nice place/ make it up/ I forget Sara's birthday yesterday
3. is divided/ New Headway book/ into 12 units
4. myself to working on this project/ I/ for the next two months/ commit
5. may bring out/ Apple/ in March 2016/ a new iPhone
6. of the room/ those big desks / too much space/ take up
7. Saudi history/ cover/ those books/ since king Abdul-Aziz era
8. 50% of law school students/ women/ make up nearly/ in the country
9. divide/ by five?/ now, can you/ this sum
10. to find out what makes people/ commit/ the study aims/ violent acts and kill innocents
11. brought out/ they/ for the new guests/ another plate from the kitchen
12. sewing as a hobby/ my sister recently/ has taken up
13. may cover/ 6000 Riyals/ that are needed to the car after the accident/ the cost of the repairs
14. on Thursday/ you should/ your mind about the dinner/ make up
15. the government into conflicting groups/ this issue/ may divide
16. committed/ last year/ she was/ to a mental hospital
17. can help to/ the flavour of food/ salt/ bring out
18. tomorrow/ let's/ this problem with the manager/ take up

2

Student B (Worksheet)

3. Now, use the verb with a new complement and say it to your partner:

1. Cover: .....
2. Cover: .....
3. Cover: .....
4. Make up: .....
5. Make up: .....
6. Make up: .....
7. Divide: .....
8. Divide: .....
9. Divide: .....
10. Bring out: .....
11. Bring out: .....
12. Bring out: .....
13. Commit: .....
14. Commit: .....
15. Commit: .....
16. Take up: .....
17. Take up: .....
18. Take up: .....

3

Student B (Answers for A) **Student B (Answers for A)** **Student B (Answers for A)**

4. Matching activity

	Verb
C	his face
D	my mind whether to go
F	into two semester
A	to a state prison
E	the new album
B	photography

B	a broad range of subjects
C	for my behaviour
F	the country
E	himself to charity
A	the fancy dishes
D	the matter with the police

F	your tuition fees
C	60 per cent of the sales
E	by six is five
A	a murder
B	the colour of your eyes
D	half of the room

4

5. Sentences Reordering

1. He covered his face with his hands.
2. I haven't made up my mind whether to go yet.
3. Each school year is divided into two semesters.
4. He was committed to a state prison.
5. Miami band may bring out their new album next month.
6. I was thinking about taking up photography.

7. The magazine covers a broad range of subjects from sewing to psychology.
8. I sent her flowers to make up for my behaviour.
9. The war divided the country into conflicting groups.
10. He committed himself to charity works.
11. My mother usually brings out the fancy dishes for special guests.
12. The school took the matter up with the police.

13. This money should cover your tuition fees.
14. Company cars make up 60 per cent of the sales
15. Thirty divided by six is five.
16. Terrorists committed some murders in Saudi.
17. That dress really brings out the colour of your eyes.
18. The new sofas take up half of the living room.

5

6. Sentences (verbs with new complements)

1. Cover	☐ Yes	☐ No
2. Cover	☐ Yes	☐ No
3. Cover	☐ Yes	☐ No
4. Make up	☐ Yes	☐ No
5. Make up	☐ Yes	☐ No
6. Make up	☐ Yes	☐ No
7. Divide	☐ Yes	☐ No
8. Divide	☐ Yes	☐ No
9. Divide	☐ Yes	☐ No
10.Bring out	☐ Yes	☐ No
11.Bring out	☐ Yes	☐ No
12.Bring out	☐ Yes	☐ No
13.Commit	☐ Yes	☐ No
14.Commit	☐ Yes	☐ No
15.Commit	☐ Yes	☐ No
16.Take up	☐ Yes	☐ No
17.Take up	☐ Yes	☐ No
18.Take up	☐ Yes	☐ No

6

Note. Only samples are included here. The full materials are available in the CD.

Appendix 3

VLT

Name: _____ الإسم:
Class: _____ الشعبة:

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

l	business	
2	clock	_____ part of a house
3	horse	_____ animal with four legs
4	pencil	_____ something used for writing
5	shoe	
6	wall	

You answer it in the following way.

l	business	
2	clock	___*6*__ part of a house
3	horse	___*3*__ animal with four legs
4	pencil	___*4*__ something used for writing
5	shoe	
6	wall	

Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are business, clock, and shoe.

If you have no idea about the meaning of a word, do not guess. But if you think you might know the meaning, then you should try to find the answer.

Version 1 The 2,000 word level

1 birth	
2 dust	_____ game
3 operation	_____ winning
4 row	_____ being born
5 sport	
6 victory	

1 adopt	
2 climb	_____ go up
3 examine	_____ look at closely
4 pour	_____ be on every side
5 satisfy	
6 surround	

1 choice	
2 crop	_____ heat
3 flesh	_____ meat
4 salary	_____ money paid regularly for
5 secret	doing a job
6 temperature	

1 bake	
2 connect	_____ join together
3 inquire	_____ walk without purpose
4 limit	_____ keep within a certain size
5 recognize	
6 wander	

1 cap	
2 education	_____ teaching and learning
3 journey	_____ numbers to measure with
4 parent	_____ going to a far place
5 scale	
6 trick	

1 burst	
2 concern	_____ break open
3 deliver	_____ make better
4 fold	_____ take something to someone
5 improve	
6 urge	

1 attack	
2 charm	_____ gold and silver
3 lack	_____ pleasing quality
4 pen	_____ not having something
5 shadow	
6 treasure	

1 original	
2 private	_____ first
3 royal	_____ not public
4 slow	_____ all added together
5 sorry	
6 total	

1 cream	
2 factory	_____ part of milk
3 nail	_____ a lot of money
4 pupil	_____ person who is studying
5 sacrifice	
6 wealth	

1 brave	
2 electric	_____ commonly done
3 firm	_____ wanting food
4 hungry	_____ having no fear
5 local	
6 usual	

Version 1 The 3,000 word level

1 belt	
2 climate	_____ idea
3 executive	_____ inner surface of your hand
4 notion	_____ strip of leather worn
5 palm	around the waist
6 victim	

1 betray	
2 dispose	_____ frighten
3 embrace	_____ say publicly
4 injure	_____ hurt seriously
5 proclaim	
6 scare	

1 acid	
2 bishop	_____ cold feeling
3 chill	_____ farm animal
4 ox	_____ organization or framework
5 ridge	
6 structure	

1 encounter	
2 illustrate	_____ meet
3 inspire	_____ beg for help
4 plead	_____ close completely
5 seal	
6 shift	

1 bench	
2 charity	_____ long seat
3 jar	_____ help to the poor
4 mate	_____ part of a country
5 mirror	
6 province	

1 assist	
2 bother	_____ help
3 condemn	_____ cut neatly
4 erect	_____ spin around quickly
5 trim	
6 whirl	

1 boot	
2 device	_____ army officer
3 lieutenant	_____ a kind of stone
4 marble	_____ tube through which blood
5 phrase	flows
6 vein	

1 annual	
2 concealed	_____ wild
3 definite	_____ clear and certain
4 mental	_____ happening once a year
5 previous	
6 savage	

1 apartment	
2 candle	_____ a place to live
3 draft	_____ chance of something
4 horror	happening
5 prospect	_____ first rough form of
6 timber	something written

1 dim	
2 junior	_____ strange
3 magnificent	_____ wonderful
4 maternal	_____ not clearly lit
5 odd	
6 weary	

Version 2 The 2,000 word level

1 copy		
2 event	_____	end or highest point
3 motor	_____	this moves a car
4 pity	_____	thing made to be like another
5 profit		
6 tip		

1 admire		
2 complain	_____	make wider or longer
3 fix	_____	bring in for the first time
4 hire	_____	have a high opinion of someone
5 introduce		
6 stretch		

1 accident		
2 debt	_____	loud deep sound
3 fortune	_____	something you must pay
4 pride	_____	having a high opinion of yourself
5 roar		
6 thread		

1 arrange		
2 develop	_____	grow
3 lean	_____	put in order
4 owe	_____	like more than something else
5 prefer		
6 seize		

1 coffee		
2 disease	_____	money for work
3 justice	_____	a piece of clothing
4 skirt	_____	using the law in the right way
5 stage		
6 wage		

1 blame		
2 elect	_____	make
3 jump	_____	choose by voting
4 manufacture	_____	become like water
5 melt		
6 threaten		

1 clerk		
2 frame	_____	a drink
3 noise	_____	office worker
4 respect	_____	unwanted sound
5 theater		
6 wine		

1 ancient		
2 curious	_____	not easy
3 difficult	_____	very old
4 entire	_____	related to God
5 holy		
6 social		

1 dozen		
2 empire	_____	chance
3 gift	_____	twelve
4 opportunity	_____	money paid to the government
5 relief		
6 tax		

1 bitter		
2 independent	_____	beautiful
3 lovely	_____	small
4 merry	_____	liked by many people
5 popular		
6 slight		

Version 2 The 3,000 word level

1 bull	
2 champion	_____ formal and serious manner
3 dignity	_____ winner of a sporting event
4 hell	_____ building where valuable
5 museum	objects are shown
6 solution	

1 abandon	
2 dwell	_____ live in a place
3 oblige	_____ follow in order to catch
4 pursue	_____ leave something
5 quote	permanently
6 resolve	

1 blanket	
2 contest	_____ holiday
3 generation	_____ good quality
4 merit	_____ wool covering used on
5 plot	beds
6 vacation	

1 assemble	
2 attach	_____ look closely
3 peer	_____ stop doing something
4 quit	_____ cry out loudly in fear
5 scream	
6 toss	

1 comment	
2 gown	_____ long formal dress
3 import	_____ goods from a foreign
4 nerve	country
5 pasture	_____ part of the body which
6 tradition	carries feeling

1 drift	
2 endure	_____ suffer patiently
3 grasp	_____ join wool threads together
4 knit	_____ hold firmly with your hands
5 register	
6 tumble	

1 administration	
2 angel	_____ group of animals
3 frost	_____ spirit who serves God
4 herd	_____ managing business and
5 fort	affairs
6 pond	

1 brilliant	
2 distinct	_____ thin
3 magic	_____ steady
4 naked	_____ without clothes
5 slender	
6 stable	

1 atmosphere	
2 counsel	_____ advice
3 factor	_____ a place covered with grass
4 hen	_____ female chicken
5 lawn	
6 muscle	

1 aware	
2 blank	_____ usual
3 desperate	_____ best or most important
4 normal	_____ knowing what is happening
5 striking	
6 supreme	

Appendix 4

Study 1

The receptive test (form-meaning recognition)

Name _____ **Section**_____

Instructions

- This is a vocabulary test. You must choose the right word to go with each meaning.
- هذا اختبار للمفردات لذا يجب عليكِ اختيار الكلمة الصحيحة التي تتماشى مع كل معنى
- Match the word in the left column with the meaning in the centre column. Shade the letter of the correct answer.
- صلي الكلمات في العمود الأيسر بمعناها في العمود الوسط ثم ظللي حرف الإجابة الصحيحة
- You answer it in the following way
- تستطيعِ الإجابة كالنحو التالي

A. feed **B.** claim **C.** leave **D.** come down **E.** make out **F.** break off	1. ينفصل عن شيء 2. يُشجّع شخصاً ما 3. يُغادر	1. Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ ● 2. ● Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ 3. Ⓐ Ⓑ ● Ⓓ Ⓔ Ⓕ

- Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are **claim**, **come down**, **make out**
- هناك بعض الكلمات موجودة في الإختبار لجعله أكثر صعوبة لذا لا يجب عليكِ البحث عن معاني هذه الكلمات الكلمات في المثال أعلاه **claim**, **come down**, **make out** تعتبر مثال لهذه الكلمات
- Try to do every part of the test.
- حاولي الإجابة على كل أجزاء الإختبار

A. play **B.** charge **C.** cover **D.** move on **E.** lay down **F.** take up	1. يضع شيئاً ما جانباً 2. يُحاسب/ يطلب مبلغ 3. يغُلف/يُغطي	1. A B C D ● F 2. A ● C D E F 3. A B ● D E F

A. divide **B.** commit **C.** serve **D.** make up **E.** bring out **F.** take up	4. يخدم 5. يستحوذ على شيئاً ما (مساحة/وقت) 6. يرتكب خطأ او جريمة	4. A B ● D E F 5. A B C D E ● 6. A ● C D E F

A. play **B.** charge **C.** cover **D.** make up **E.** bring out **F.** go off	7. يلعب/ يتسلى 8. يُرّكب/ يُجّمع/ يُشكل شيئاً ما 9. يُبرز/يُثير الانتباه لـ (تفاصيل) شيئاً ما	7. ● B C D E F 8. A B C ● E F 9. A B C D ● F

A. divide **B.** commit **C.** serve **D.** move on **E.** lay down **F.** go off	10. يوزع/ يُقسّم 11. يقدّم الطعام 12. ينتقل إلى موضوع أوعمل آخر	10. ● B C D E F 11. A B ● D E F 12. A B C ● E F

A. charge **B.** cover **C.** divide **D.** lay down **E.** go off **F.** make up	13. يُعوّض عن شيئاً ما 14. يُقسّم (رياضيات) 15. يُكلف شخصاً بعمل شيء ما	13. A B C D E ● 14. A B ● D E F 15. ● B C D E F

A. commit **B.** serve **C.** play **D.** bring out **E.** take up **F.** move on	16. يُصدر(كتاب)/ يُطلق (منتج) 17. يستخدم كـ (يمكن استخدامه كـ) 18. يُناقش (أمراً) مع	16. A B C ● E F 17. A ● C D E F 18. A B C D ● F

A. charge **B.** cover **C.** divide **D.** bring out **E.** take up **F.** move on	19. يُخرج 20. يسبب انقسام/اختلاف في الآراء 21. ينتقل (يغير) إلى مكان آخر	19. A B C ● E F 20. A B ● D E F 21. A B C D E ●

A. commit **B.** serve **C.** play **D.** lay down **E.** go off **F.** make up	22. يستلقي 23. انطلق لغرض معين 24. يلعب رياضة/ يشترك بالرياضة	22. A B C ● E F 23. A B C D ● F 24. A B ● D E F

A. cover **B.** divide **C.** commit **D.** go off **E.** make up **F.** bring out	25. يشمل 26. يتعهد بفعل شيئاً ما 27. يرّن (يبدأ بإطلاق صوت مزعج)	25. ● B C D E F 26. A B ● D E F 27. A B C ● E F

A. serve **B.** play **C.** charge **D.** take up **E.** move on **F.** lay down	28. يتّهم شخصاً بجريمة أو خطأ 29. فرض/يصدر قوانين أو قواعد جديدة 30. يواصل حياته بعد تجربة صعبة	28. A B ● D E F 29. A B C D E ● 30. A B C D ● F

A. cover **B.** divide **C.** commit **D.** take up **E.** move on **F.** lay down	31. أودع شخصاً ما السجن (مؤسسة) 32. يغطي التكاليف/النفقات 33. يُباشر/يتخذ هواية جديدة	31. A B ● D E F 32. ● B C D E F 33. A B C ● E F

A. cover **B.** serve **C.** play **D.** charge **E.** go off **F.** make up	34. يقرّر 35. انفجر 36. يعزف على الة موسيقية	34. A B C D E ● 35. A B C D ● F 36. A B ● D E F

A. run **B.** lift **C.** hit **D.** come around **E.** fill in **F.** turn up	37. يحل محل شخصاً ما 38. يعرّج على/ يزور 39. يصدم/ يضرب بسرعة وبشدة	37. A B C D **F** 38. A B C **E** F 39. A B **D** E F

A. relate **B.** order **C.** train **D.** take off **E.** break down **F.** pass on	40. يربط بين شيء وشيء 41. يطلب طعام/شراب إلخ 42. يخلع ملابسه	40. **A** B C D E F 41. A **C** D E F 42. A B C **E** F

A. run **B.** lift **C.** hit **D.** take off **E.** break down **F.** pass on	43. يرفع 44. تناقل الشيء من شخص إلى آخر 45. يتعطل/ يتوقف عن العمل	43. A **C** D E F 44. A B C D E **F** 45. A B C D **E** F

A. relate **B.** order **C.** train **D.** come around **E.** fill in **F.** turn up	46. يأمر شخصاً بفعل شيء ما 47. يُعلّم (يُدرّب) 48. يُطّلع شخصاً ما بالجديد (أو مافاته)	46. A **C** D E F 47. A B **D** E F 48. A B C D **E** F

A. run **B.** hit **C.** relate **D.** fill in **E.** turn up **F.** take off	49. يغادرمكان ما فجأة 50. يتفهم/يستوعب/يتعاطف مع 51. يركض	49. (A) (B) (C) (D) (E) ● 50. (A) (B) ● (D) (E) (F) 51. ● (B) (C) (D) (E) (F)

A. order **B.** train **C.** lift **D.** break down **E.** pass on **F.** come around	52. يُورث 53. يُنهي/ يُبطل قانون أو حكم قضائي 54. يتدرب/يتمرّن (رياضة)	52. (A) (B) (C) (D) ● (F) 53. (A) (B) ● (D) (E) (F) 54. (A) ● (C) (D) (E) (F)

A. lift **B.** hit **C.** relate **D.** break down **E.** pass on **F.** come around	55. يفكك/يجّزأ إلى أجزاء صغيرة 56. يسبب عناءً(ضرر)/ يؤثر عليه 57. يغير رأيه/ يقتنع برأي شخصاً آخر	55. (A) (B) (C) ● (E) (F) 56. (A) ● (C) (D) (E) (F) 57. (A) (B) (C) (D) (E) ●

A. order **B.** train **C.** run **D.** fill in **E.** turn up **F.** take off	58. ينظم/ يرتب 59. يظهر بعد فقدانه (شيء ضائع) 60. تولى إدارة شغلاً ما	58. ● (B) (C) (D) (E) (F) 59. (A) (B) (C) (D) ● (F) 60. (A) (B) ● (D) (E) (F)

A. hit **B.** relate **C.** order **D.** turn up **E.** come around **F.** break down	61. يحكي/يروي 62. يزيد/ يرفع (الصوت) 63. ينهارلدرجة البكاء	61. A ● C D E F 62. A B C ● E F 63. A B C D E ●

A. train **B.** run **C.** lift **D.** pass on **E.** come around **F.** fill in	64. يُصوّب/يُوجّه 65. يرفع معنويات او يغير مزاج شخص ما 66. يتكرر، يُعاد (الحدث أو الزمن)	64. ● B C D E F 65. A B ● D E F 66. A B C D ● F

A. hit **B.** relate **C.** order **D.** pass on **E.** take off **F.** fill in	67. يملأ شيئاً ما 68. يتوفى/ يموت 69. صل لمستوى أوحالة معينة (هبوط أو ارتفاع)	67. A B C D E ● 68. A B C ● E F 69. ● B C D E F

A. train **B.** run **C.** lift **D.** turn up **E.** take off **F.** break down	70. جاء/أتى 71. يجري (سوائل) 72. تقلع (الطائرة)	70. A B C ● E F 71. A ● C D E F 72. A B C D ● F

A. fall **B.** clear **C.** lead **D.** cut off **E.** go along **F.** throw out	73. يسقط/يقع 74. يقطع بشيء حاد (سكين) 75. يكون مسئول/ يقود/ يسيطر	73. ● B C D E F 74. A B C ● E F 75. A B ● D E F

A. introduce **B.** blow **C.** address **D.** back up **E.** get through **F.** hold up	76. يبدد المال/يُبذر 77. يجتاز 78. يكتب عنوان/يعنون	76. A ● C D E F 77. A B C D ● F 78. A B ● D E F

A. fall **B.** clear **C.** lead **D.** back up **E.** get through **F.** throw out	79. يرفض شيئاً ما 80. يقود إلى 81. يتقهقر(يرّتد)	79. A B C D E ● 80. A B ● D E F 81. A B C ● E F

A. introduce **B.** blow **C.** address **D.** cut off **E.** go along **F.** hold up	82. يزفر(يخرج الهواء) 83. يرفع إلى الأعلى 84. يتقدم، يحرز تقدماً	82. A ● C D E F 83. A B C D E ● 84. A B C D ● F

A. clear **B.** lead **C.** introduce **D.** go along **E.** get through **F.** back up	85. يساعد/ يساند 86. يُخلي 87. يقدم (يطرح) شئ للمرة الأولى	85. (A) (B) (C) (D) (E) ● 86. ● (B) (C) (D) (E) (F) 87. (A) (B) ● (D) (E) (F)

A. blow **B.** address **C.** fall **D.** get through **E.** throw out **F.** cut off	88. يتخلّص من شئ (يرمي) 89. ينخفض/ يصبح اقل 90. يوصل فكرته للشخص الآخر	88. (A) (B) (C) (D) ● (F) 89. (A) (B) ● (D) (E) (F) 90. (A) (B) (C) ● (E) (F)

A. clear **B.** lead **C.** introduce **D.** hold up **E.** throw out **F.** cut off	91. يجتاز بدون لمس 92. يُعرّف (يُقدّم) شخص 93. يقاطع شخصاً أثناء حديثه	91. ● (B) (C) (D) (E) (F) 92. (A) (B) ● (D) (E) (F) 93. (A) (B) (C) (D) (E) ●

A. blow **B.** address **C.** fall **D.** go along **E.** hold up **F.** back up	94. يعالج مشكلة 95. يتحمل/يصبر/يحافظ على قوته 96. يوافق/ يقبل/ يتماشى مع	94. (A) ● (C) (D) (E) (F) 95. (A) (B) (C) (D) ● (F) 96. (A) (B) (C) ● (E) (F)

A. lead **B.** introduce **C.** blow **D.** hold up **E.** back up **F.** get through	97. يفسد امراً ما/يُضيع فرصة 98. يصل (يتواصل مع شخصاً) عبر الهاتف 99. يؤدي إلى (طريق)	97. A B ● D E F 98. A B C D E ● 99. ● B C D E F

A. address **B.** fall **C.** clear **D.** throw out **E.** cut off **F.** go along	100. يُصبح (يتغير إلى حالة سلبية) 101. يُثبت براءة /يُبرَّأ 102. يوقّف إمداد شئ ما	100. A ● C D E F 101. A B ● D E F 102. A B C D ● F

A. lead **B.** introduce **C.** blow **D.** throw out **E.** cut off **F.** go along	103. يطرد شخصاً 104. يضيف شيئ على شيئ ما 105. يمر بمكان بدون تخطيط	103. A B C ● E F 104. A ● C D E F 105. A B C D E ●

A. address **B.** fall **C.** clear **D.** hold up **E.** back up **F.** get through	106. يُعطّل/ يؤخر (شخصاً أو شيئاً) 107. يخاطب برسمية 108. يدعم/يُثبت صحته	106. A B C ● E F 107. ● B C D E F 108. A B C D ● F

A. save **B.** treat **C.** collect **D.** give out **E.** get off **F.** put up	109. يُعَلِّق/يرفع شيئاً 110. يوزّع شيئاً، ينشر 111. يُعامل	109. (A) (B) (C) (D) (E) ●	 110. (A) (B) (C) ● (E) (F) 111. (A) ● (C) (D) (E) (F)
A. count **B.** settle **C.** enter **D.** set off **E.** hold back **F.** take in	112. يدخل مكان ما 113. يحُلّ/ يحسم (مشكلة أو مسألة ما) 114. يتحفظ/ يتكتّم على شئ ما	112. (A) (B) ● (D) (E) (F) 113. (A) ● (C) (D) (E) (F) 114. (A) (B) (C) (D) ● (F)
A. save **B.** treat **C.** collect **D.** set off **E.** hold back **F.** take in	115. يجمع 116. ينطلق في رحلة 117. يُسكن شخصاً في بيته	115. (A) (B) ● (D) (E) (F) 116. (A) (B) (C) ● (E) (F) 117. (A) (B) (C) (D) (E) ●
A. count **B.** settle **C.** enter **D.** give out **E.** get off **F.** put up	118. يعدّ/ يُحصي 119. يتحمل/يصبر على شئ مزعج 120. ينزل من القطار أو السيارة.. إلخ	118. ● (B) (C) (D) (E) (F) 119. (A) (B) (C) (D) (E) ● 120. (A) (B) (C) (D) ● (F)

A. treat **B.** collect **C.** count **D.** get off **E.** put up **F.** set off	121. يُفجّر(القنبلة) 122. يُقلّ شخصاً/شيئاً من مكان ما 123. يُضمّن (يُدخل في الحسبان)	121. A B C D E ● 122. A ● C D E F 123. A B ● D E F
A. settle **B.** enter **C.** save **D.** hold back **E.** take in **F.** give out	124. يستقرّ/يستوطن 125. يُنقذ من الخطر 126. يستوعب، يفهم	124. ● B C D E F 125. A B ● D E F 126. A B C D ● F
A. treat **B.** collect **C.** count **D.** hold back **E.** take in **F.** give out	127. يُعالج (طبياً) 128. يُصرح/ يُعلن شيئاً ما (أخبار..إلخ) 129. يقيّد/يمنع/يعيق شخصاً (من التقدم)	127. ● B C D E F 128. A B C D E ● 129. A B C ● E F
A. settle **B.** enter **C.** save **D.** get off **E.** put up **F.** set off	130. يبني / يُشّيد 131. يبدأ (بداية جيدة او سيئة) 132. يلتحق/ ينضم لـ مجموعة	130. A B C D ● F 131. A B C ● E F 132. A ● C D E F

A. collect **B.** count **C.** settle **D.** put up **E.** set off **F.** hold back	133. يكبت نفسه/يكبح مشاعره 134. يجمع أفكاره (يستجمع نفسه) 135. سبب في بداية او حدوث شيئاً ما	133. A B C D E ● 134. ● B C D E F 135. A B C D ● F

A. enter **B.** save **C.** treat **D.** take in **E.** give out **F.** get off	136. يخدع شخصاً 137. يوفّر/يحافظ على 138. يتجنب عقوبة/ يفلت من العقوبة	136. A B C ● E F 137. A ● C D E F 138. A B C D E ●

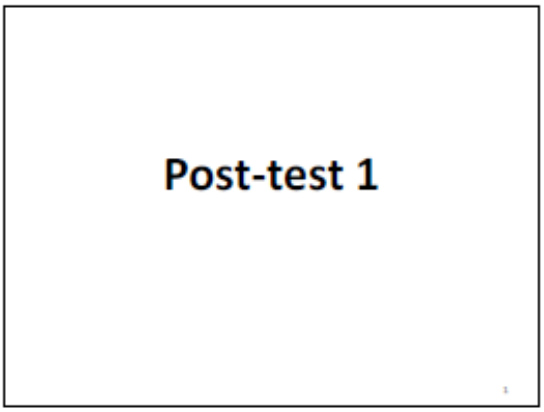
A. collect **B.** count **C.** settle **D.** take in **E.** give out **F.** put up	139. يهُـمّ/ يعتبر أهم 140. يجلس بـ وضعية مريحة 141. يتعطّل: عضو جسدي	139. A ● C D E F 140. A B ● D E F 141. A B C D ● F

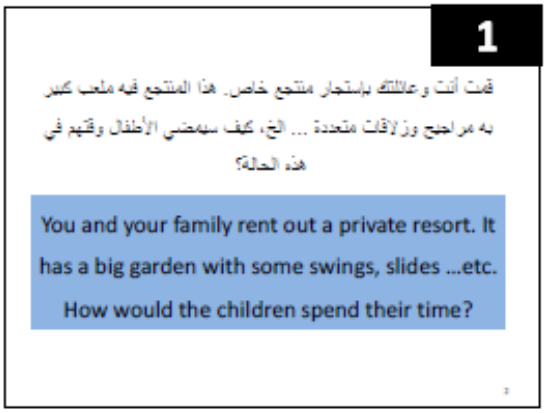
A. enter **B.** save **C.** treat **D.** get off **E.** set off **F.** hold back	142. يُغطى/يُعالج بـ مادة لحمايته 143. يدّخر للمستقبل 144. يكتب/ يدُخل للكمبيوتر	142. A B ● D E F 143. A ● C D E F 144. ● B C D E F

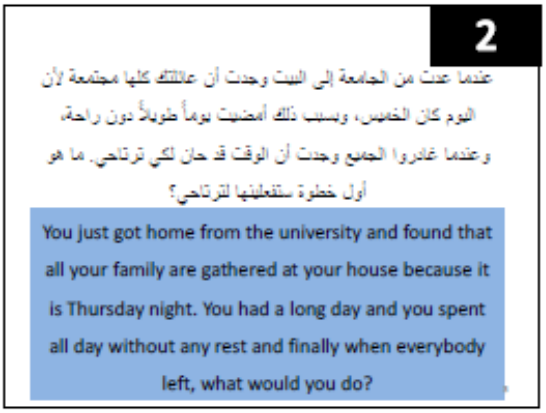
Appendix 5

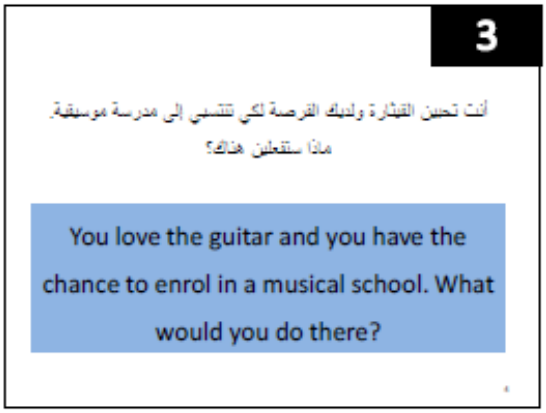
Study 1

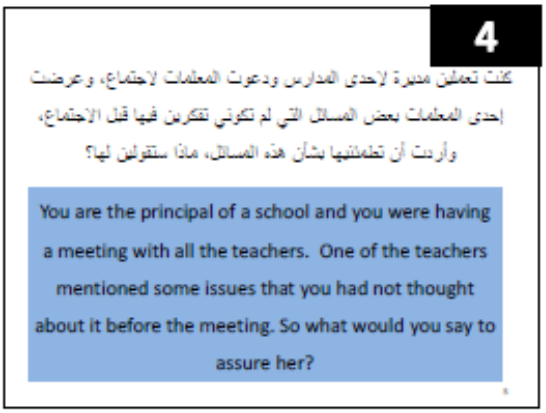
The productive test (form-meaning recall)

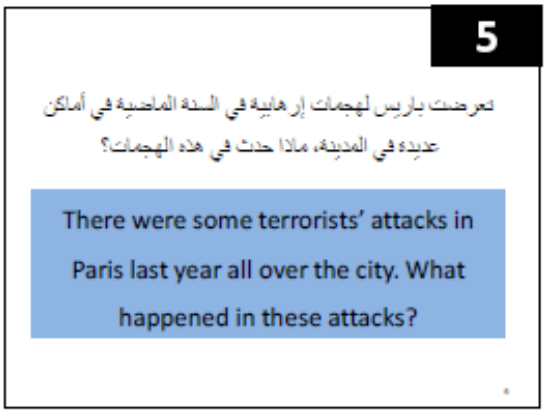












6

من المحزن أن الامة العراقية بعد وفاة صدام حسين دخلت في أزمات متعددة، هل ممكن ان تذكري مثال لـ ما حصل للشعب العراقي؟

Sadly, after Saddam Hussein death, the Iraqi nation went through a lot of crises. Can you give an example of what happened to the people and religious groups?

7

ما هي وظيفة الشرطي؟ وما هو الشيء الذي يفعله الشرطي للمجتمع؟

What does a police officer do for the community?

8

كنت تعيشين في جدة وكنت تحبين هذه المدينة، ولكن عندما بدأت العمل تم نقلك إلى ينبع ولم تعجبك ينبع إطلاقاً. والآن تقاعدت ورأيت أنه يجب أن لا تعيشي في ينبع بعد الآن، فماذا ستفعلين؟

You used to live in Jeddah and you loved it there. But when you started working in Yanbu, you started living there. You didn't like Yanbu at all. Now, you are retired and you still don't like Yanbu, what might you do?

9

فعلت شيئاً أغضب والدتك واعتذرت لها شفوياً ولكنها لم تقبل ذلك، وبدأت تفكرين في الطريقة التي تتالين بها رضاها فقررت في شراء بعض الزهور مصحوبة ببطاقة مكتوب عليها "إنني آسفة حقاً"، لماذا تفعلين ذلك؟

You did something that made your mom angry and you apologised many times but she is still angry. You are thinking how you can make it right and you decided to buy her some flowers with a card saying you are really sorry. Why are you doing that?

10

يملك والدك مزرعة بها بعض الأحصنة ويذهب إلى المزرعة كل يوم لكي يعتني بها (مثل إطعامها وغير ذلك) ولديه عامل واحد وهو "علي" الذي يساعده في ذلك. كان والدك يريد أن يسافر لبضعة أيام، فماذا تعتقدين أنه سيفعل ليضمن أن الأحصنة ستتم العناية بها؟

Your father owns a farm with some horses. He goes to the farm every day and takes care of these horses. He has one worker 'Ali' who helps him with that. If your father wants to travel somewhere, what do think he will do to make sure that the horses are taken care of?

11

عندما يمارس شخص فعلاً غير قانوني مثل السرقة وهو يعاني من مشكلة عقلية، فماذا ستفعل المحكمة؟

When somebody do something illegal such as stealing but he has a mental problem. What do you think the court should do?

12

أمضيت الليلة مع خالتك، ونام أخوك الصغير في الغرفة التي كانت باردة جداً، وطلبت من خالتك أن تعطيك بطانية، فماذا ستفعلين له لكي تجعليه يحس بالدفء؟

You were spending the night at your aunt's house. Your little brother slept there and the room was very cold. There is a blanket in the room. What would you do with it?

13

كنت تقرأين كتاباً لبعض الوقت وبدأت تشعرين بالألم في عينيك وتشعرين بالتعب على الرغم من أنك كنت تستمتعين بالقراءة، فماذا ستفعلين؟

You were reading a book for a while. Your eyes started hurting and you started to feel tired even though that you are enjoying the reading. What should you do?

14

عندما تقوم والدتك بتعليمك كيف تطبخين اللحم كانت دائماً تقول لك استخدمي بعض البهارات لكي تعطي اللحم شيئاً من النكهة، لماذا تريد منك أن تفعلي ذلك؟ وما هو تأثير البهارات؟

When your mother teaches you how to cook meat, she always tells you that you need to use some spices. Why do you think she wants you to do that?

15

جاء بعض صديقات والدتك للعشاء، وعندما جلست معهم لاحظت أن إحداهن ليس أمامها طبق، وهناك بعض الأطباق في دولاب المطبوخ، فماذا ستفعلين؟

Some of your mother's friends came over for dinner. When they sat down at the table, you noticed that one of them didn't have a plate. There are some extra plates in the kitchen, so what would you do?

16

في يوم من الأيام أرادت إحدى صديقاتك أن تقوم بعمل فيه شيء من التهور وأن تدخن في الحمام وأردت تحذيرها من وجود جهاز الإنذار، ماذا ستقولين لها؟

One day in the school, one of your friends wanted to do something reckless and thought about smoking a cigarette in the bathroom. You want to advise her not to do that and told her that she would be caught out because …

17

يعمل أخوك في جامعة الملك عبد العزيز ويريد أن يكمل دراسته خارج المملكة، في المملكة المتحدة أو في الولايات المتحدة بعد أن حصل على منحة دراسية من الجامعة ولكنك تريدين أن تذهبي لتمضي معه شهرين من سيتحمل تكاليف رحلتك؟

Your brother works in King Abdul-Aziz university. He is planning to complete his studies outside Saudi Arabia, UK or US, since he has a scholarship from the university. But you are thinking of going with him for the first two months. Who would be responsible for your expenses?

18

تم منحك وظيفة معلمة براتب قدره ١٢٠ الف ريال في السنة، وتريدين أن تعرفي كم هو راتبك في الشهر ، فماذا ستفعلين؟

You are offered a teacher job for 120,000 Riyals per year. You want to know how much you would make every month. What would you do?

19

في يوم من الأيام قبل أن تخرج والدتك طلبت منك أن تنظفي البيت قبل أن تعود، ولكنك كنت تذاكرين في ذلك الوقت ولم تستطيعي أن تفعلي ذلك قبل أن تعود، تريدين ان تخبريها بذلك كيف ستقولين لها ذلك؟

Before your mom go out one day she asked you to clean the house before she gets back. You were studying all the time and you couldn't do that before she's back. What would you say to her?

20

تقدم شاب للزواج من أختك وكان يعمل طبيباً في مستشفى ولكنها كانت تريد أن تتزوج من قريبها الذي لا يعمل حتى الآن. أنت تعتقدين أنه ليس جاد في ذلك وتريدين منها أن تقرر ما إذا كانت تريد الزواج منه أم لا، كيف ستقولين لها ذلك؟

A young man proposed to your sister. He works as a doctor in the city hospital but she always thought that she would marry your cousin who doesn't have a job yet and you feel that he is not serious with her. You want to tell her that she needs to decide. How would you say that?

21

كانت صديقتك تعمل في نفس الوظيفة لمدة ٢٠ سنة وقد بدأت تشعر بالتعب وتريدين نصحها بـ أن تغير وظيفتها فكيف ستقولين ذلك؟

Your friend has been working in the same job for 20 years. She is tired and would like to change it. What would you advise her to do?

22

أنت تدرسين فصولاً من كتاب "New Headway" في حصص اللغة الإنجليزية في سنة التحضيري، كم عدد الوحدات الموجودة في الكتاب؟

You are studying New Headway in your English classes in the preparatory year. How many units are in the book?

23

محمد نور رجل ثري من أهل مكة ما الذي يفعله لكي يعيش منه؟

Mohammad Nour is a very rich man in Makkah. What does he do for living?

24

سمعت في الأخبار عن سرقة احد البنوك في جده ولكن من حسن الحظ انهم قبضوا على السارقين، ماذا ستفعل المحكمة لهم؟

You heard about in the news about some thieves who stole a bank in Jeddah, luckily the police catch them what do you think the police will do?

25

كنت وزوجك تعملان لفترة طويلة وشعرتما بالتعب وقد حان الوقت لبعض الراحة وقررتما السفر إلى دبي لإمضاء العطلة هناك، كيف ستقولين له ذلك؟

You and your husband have been working for a long time. Your anniversary is coming around and you both are really tired. You want to suggest that you both travel to Dubai on holiday, how can you say that?

26

سمعت بشخص أمضى في السجن 16 عاماً وتم الآن إطلاق سراحه إذ اتضح أنه بريء بعد كل هذه السنوات، وأنت تشعرين الآن بأسف لذلك، لماذا؟

You heard about a man who spent 16 years in prison and now he has been released because he was found innocent after all these years. You felt sorry for that man, why?

27

تخيلي أنك تعملين كنادلة في أحد المطاعم، فما هو الدور الذي تلعبينه بالضبط؟

Imagine that you are working as a waitress at a restaurant. Exactly what do you do?

28

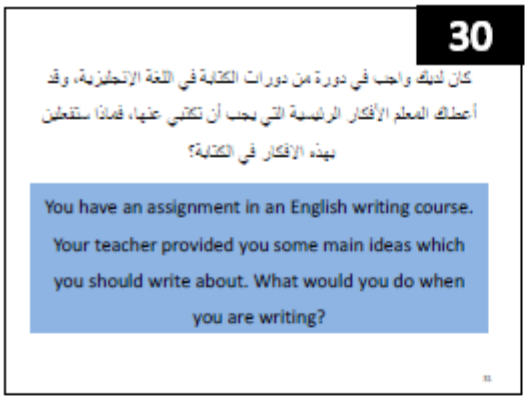
كنت تعملين في مدرسة ابتدائية، وتم تحويلك الآن إلى الصف الثاني، واليوم هو أول يوم لك معهم في الفصل وقد وجدت أن التلميذات مز عجات للغاية، فماذا ستفعلين لكي تضبطيهم من أول يوم؟

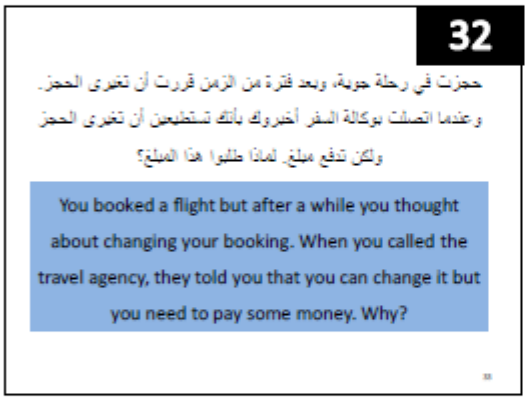
You are working in an elementary school. You have been assigned to second year children and today is your first class with them. You found out that the class is a very noisy one. You need to control it from the beginning, how can you do that?

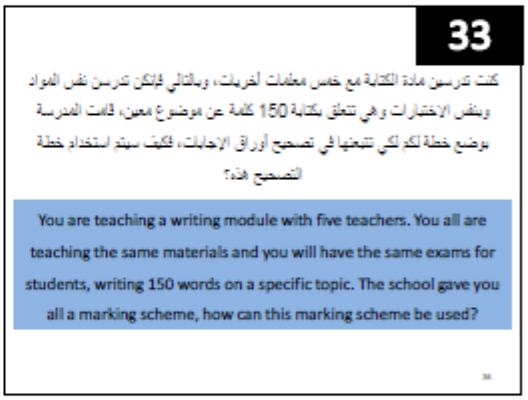
29

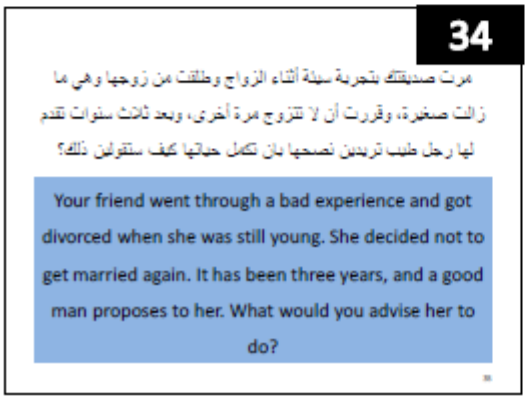
تمضي أختك وقتاً طويلاً في العمل في المكتب وفي المنزل، وقد أصابها الملل وتريد أن تفعل شيئاً مسلياً، تريدين نصحها بان تتخذ هوايه جديدة، كيف ستقولين ذلك؟

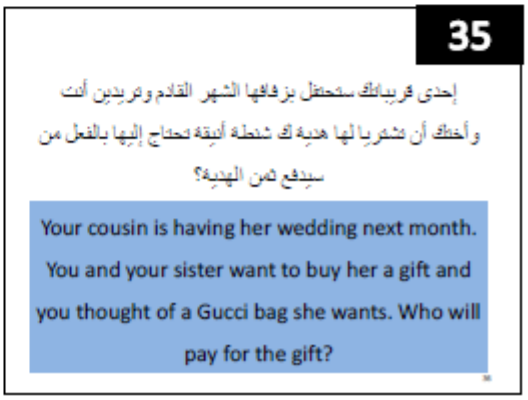
Your sister spends a lot of her time working either at work or at home. She is really bored and wants to do something interesting. What can you advise her?











Note. Only samples are included here. The full materials are available in the CD.

Appendix 6

Study 1

Ethics information sheet

ورقة معلومات المشارك (Information Sheet)

عنوان المشروع: تعلم الأفعال المفردة والأفعال المركبه متعددة المعاني عن طريق التحدث المحفز المهام

اسم الباحثة: هناء المطيري

اسم المشرفين: الدكتور مايكل رودجرز والدكتورة آنا بيليكر سانتشيز

عزيزتي الطالبة

أود أن أفيدك أنني أنا هناء المطيري أقوم حالياً بتحضير رسالة دكتوراه في مجال اللغويات التطبيقية بجامعة نوتنجهام بالمملكة المتحدة. والدراسة المطلوب منك المشاركة فيها هي عبارة عن بحث الغرض منه معرفة تأثير استخدام مخرجات التحدث المحفز على تدريس مفرادات اللغة الإنجليزية. وسوف تشتمل المشاركة على تدريس بعض الدروس، والجلوس لبعض اختبارات الاختيار المتعدد، وتسجيل اثنين من اختبارات التحدث والمحادثات أثناء إكمال النشاطات المتعلقة بالمشاركة. والمشروع طوعي وبامكانك الانسحاب في أي وقت دون إبداء أي أسباب. ولكنني أرجو أن تشاركي في هذا المشروع الذي أعتقد أنك يمكن أن تستفيدي منه.

وسوف يتم حفظ كل البيانات التي يتم جمعها من خلال هذه الدراسة في الحاسب الآلي وتتم حمايتها بكمات المرور. ولن يتم الإطلاع على بياناتك الشخصية إلا من قبل شخصي أنا والمشرف. وسوف يتم التعامل مع المعلومات بصورة سرية. كما سيتم الاحتفاظ بكل البيانات طوال فترة البحث، وبعد إكمال البحث سوف يتم حفظ البيانات في إرشيف لدى الباحثة لفترة من الزمن (سبع سنوات بعد تاريخ نشر النتائج) ومن ثم يتم محوها. وتؤكد الباحثة أن البيانات سوف يتم حفظها بسرية تامة ولن يتاح التعرف على هويه الشخص الذي أدلى بالبيانات عند عرض النتائج. ولن يتم استخدام البيانات التي تدلي بها سوى لأغراض هذا البحث. وإذا كان لديك أي استفسارات بشأن الدراسة فالرجاء عدم التردد في الاتصال على البريد الألكتروني الخاص بي وهو: aexhma@nottingham.ac.uk.

التوقيع: .....

الاسم: ..... **التاريخ** .....

<u>بيانات الاتصال</u>

الباحثة: aexhma@nottingham.ac.uk

المشرفون: Michael.Rodgers@nottingham.ac.uk, ana.pellicer-sanchez@nottingham.ac.uk

منسق الجوانب الأخلاقية للأبحاث: gareth.carrol@nottingham.ac.uk

Ethics consent form

نموذج موافقة المشارك (Consent Form)

عنوان المشروع: تعلم الأفعال المفردة والأفعال المركبه متعددة المعاني عن طريق التحدث المحفز

اسم الباحثة: هناء المطيري

اسم المشرفين: الدكتور مايكل رودجرز والدكتورة آنا بيليكر سانتشيز

الرجاء وضع علامة في المكان المناسب:

☐	☐ أوكد أن غرض الدراسة قد تم توضيحه وأنني قد فهمته
☐	☐ لقد أتيحت لي الفرصة لكي أوجه الأسئلة وقد تلقيت الإجابات الكافية عليها
☐	☐ إنني أعلم أن المشاركة في هذه الدراسة طوعية وأنني أستطيع أن أنسحب منها بدون إبداء أي أسباب ودون الخوف من أي عاقبة
☐	☐ لقد علمت أن البيانات لن تنسب إلي وأنه لن تكون هناك صلة بين البيانات الشخصية والبيانات التي أوفرها أنا
☐	☐ أقر أن البيانات سوف يتم حفظها في الحاسب الآلي من قبل الباحثة وأنه سوف تتم حمايته بواسطة كلمات المرور
☐	☐ أقر أن حديثي سوف يتم تسجيله وأن بياناتي سوف يتم نسخها والإشارة إليها دون أن تنسب لأحد في النماذج المكتوبة بعد النشر
☐	☐ أعلم أنه لن تكون هناك أي مخاطر من جراء المشاركة في هذه الدراسة
☐	☐ أعلم أنني أستطيع أن أتصل بالباحثة لأحصل على معلومات إضافية عن البحث وأستطيع أن أتصل بمنسق الجوانب الأخلاقية للأبحاث بكلية التربية بجامعة نوتنجهام إذا أردت أن أتقدم بشكوي بشأن مشاركتي في البحث
☐	☐ أوكد أنني قد قرأت وفهمت المعلومات الموضحة أعلاه ومن ثم أوافق على المشاركة في هذه الدراسة
☐	☐ أوكد أنن عمري فوق 16 عاماً

التوقيع: .....

الاسم: ..... **التاريخ** .....

بيانات الاتصال

الباحثة: aexhma@nottingham.ac.uk

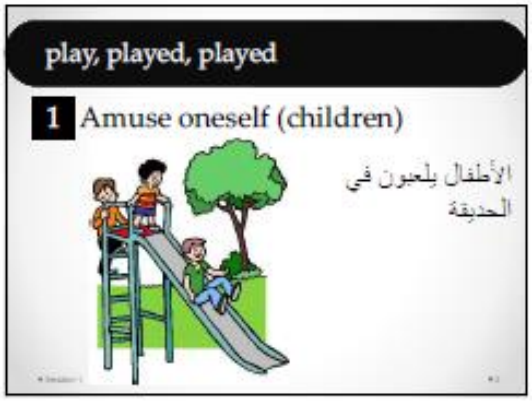
المشرفون: Michael.Rodgers@nottingham.ac.uk, ana.pellicer-sanchez@nottingham.ac.uk

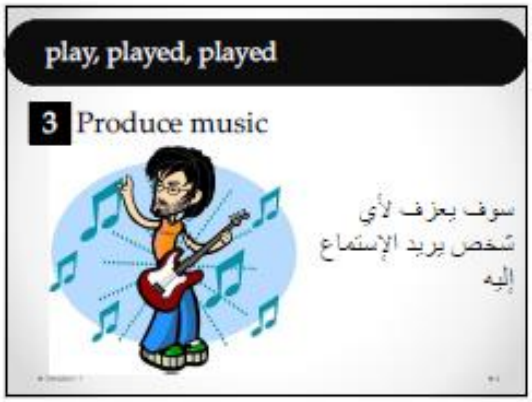
منسق الجوانب الأخلاقية للأبحاث: gareth.carrol@nottingham.ac.uk

Appendix 7

Study 1

PowerPoint presentation of the target items





charge, charged, charged

3 Accuse formally

سوف يرفع على مدير البنك دعوى ويتهمه بالرشوة

serve, served, served

1 Work

اشتغل والدي بإخلاص في الجيش لمدة أربعين عاماً

serve, served, served

2 Provide food or drink

دعني أقدم لك بعضاً من لحم البقر مع المكرونة

serve, served, served

3 Function as

ممكن أن تستخدم هذه الكنبة أيضاً كسرير إضافي

lay down, laid down, laid down

Verb+ (.....) + Prep.

1 Put something away/down

ضعوا أقلامكم جانباً لقد انتهى وقت الاختبار

lay down, laid down, laid down

Verb+ (.....) + Prep.

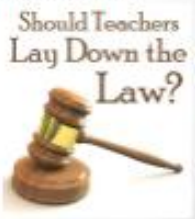
2 Lie flat on a surface

هل أستطيع فحسب أن أستلقي هنا وآخذ قسطا من الراحة ؟

lay down, laid down, laid down
Verb+ (.....) + Prep.

3 Officially make new rules

لذا علي الآن
أن أضع بعض القوانين



go off, went off, gone off
Intransitive

1 Go somewhere for a particular purpose

لنترك المدينة
ونذهب للشواطئ
الرائعة في
"طوكيو"

go off, went off, gone off
Intransitive

2 Start making loud noise

لقد جهزت هذا
المنبه كي ينطلق كل
ثلاث ساعات

go off, went off, gone off
Intransitive

3 Explode or fires

لم يكن من
المفترض
أن تنفجر هذه القنبلة

move on, moved on, moved on
Intransitive

1 Change the subject or the job

الآن، لننقل حديثنا
إلى موضوع آخر

move on, moved on, moved on
Intransitive

2 Change physical location, spot …etc.

لقد انتقلنا إلى منزل
جديد في الشهر
الماضي

Note. Only samples are included here. The full materials are available in the CD.

Appendix 8

Study 2

The sentence reconstruction activity **Session 1**

Student A:

A. Your partner is going to read some sentences to you. Listen carefully.

B. Now, Listen to your partner again. This time try to remember some words of the text. You can use Box A to take some notes.

Box (A)
1. .....
2. .....
3. .....
4. .....
5. .....

C. Try to retell the sentences.

D. Now, it's your time to read the sentences in Box (B) to your partner.

Box (B)	
6.	I'm taking extra lessons to compensate for the time I missed because of my accident.
7.	He lost his chance at getting the job when he arrived late to the interview. Maybe being fired from his last job has prevented him from progressing in his career.
8.	My mom couldn't stop her anger when she heard my brother talking about how he spent all his money on a new car he doesn't need.
9.	He started learning gardening after he was sent to prison for fraud.
10.	Even though the White Nile is bigger than the Blue Nile, it only forms 15% of the flow of the Nile River, which flows through North Africa.

E. Read the sentences again and this time give her some time to take notes.

F. Listen to your partner while she is retelling the sentences. Try to remind her to replace all the verbs.

Student B:

A. Read the following sentences to your partner and allow her to take some notes while you are reading.

Box (A)	
1.	I can't believe that he was mistakenly spent 10 years in prison for a crime that he didn't do.
2.	She obligated herself to take care of her family after her mother's death, which required all her time, as she has five brothers.
3.	I wanted my friend to decide for herself whether to marry my brother or not, without any constraints. So, I did not say my opinion.
4.	My brother began managing the family business after he graduated from university last Monday. On his first day, he called a meeting to discuss all employees' complaints and issues.
5.	There was a man breathing rapidly on the bus, sending air out of his mouth. He must have been moving quickly before he got on the bus.

B. Read the sentences again and this time give her some time to take notes.

C. Listen to your partner while she is retelling the sentences. Try to remind her to replace all the verbs.

D. Now, it's your time to listen to your partner while she is reading some sentences. Listen carefully and try to remember some words of the text. You can use Box B to take some notes.

Box (B)	
6.	.....
7.	.....
8.	.....
9.	.....
10.	.....

5. Now, try to retell the sentences.

The sentence reconstruction activity **Session 2**

Student A:

A. Your partner is going to read some sentences to you. Listen carefully.

B. Now, Listen to your partner again. This time try to remember some words of the text. You can use Box A to take some notes.

Box (A)
1. .....
2. .....
3. .....
4. .....
5. .....

C. Try to retell the sentences.

D. Now, it's your time to read the sentences in Box (B) to your partner.

Box (B)
6. The summer holiday usually takes place every year in Shaban and lasts for three months so that students come back refreshed.
7. The company provided its employees with a course to teach them how to use the new software. However, the trainer couldn't make it that day, so they had an external trainer substitute for her.
8. The news of her friend's death strongly affected her, so I tried reminding her that her friend had moved on to a better place.
9. The temperature reached 53°C yesterday, so the football players could not exercise as much as usual.
10. My grandmother used to tell a story about a man who had a fortune and no sons to inherit it, so all his money was donated to the poor.

5. Read the sentences again and this time give her some time to take notes.
6. Listen to your partner while she is retelling the sentences. Try to remind her to replace all the verbs.

Student B:

A. Read the following sentences to your partner and allow her to take some notes while you are reading.

Box (A)
1. My mother bought a new wardrobe, but she was complaining about a little mark on the door, so I suggested that she patch it with some wood fillers, and nobody would know the mark was there.
2. My sister's friend Rana asked us to visit her house. When we went there, we met a friend of her mother's who is in a wheelchair. We could understand her because our grandmother was in a wheelchair in her last years.
3. With five guns suddenly pointed at him, he scrambled away so quickly that he ran into the street barriers.
4. The teacher gave Sara a paper on which to write some personal information and then instructed her to give it to the student beside her.
5. If we compare the money we're paying to the services being provided, they may change their opinion in the end and reduce the price.

B. Read the sentences again and this time give her some time to take notes.

C. Listen to your partner while she is retelling the sentences. Try to remind her to replace all the verbs.

D. Now, it's your time to listen to your partner while she is reading some sentences. Listen carefully and try to remember some words of the text. You can use Box B to take some notes.

Box (B)
6. .....
7. .....
8. .....
9. .....
10. .....

E. Now, try to retell the sentences. Remember to replace all the verbs with words from your own.

Note. Only samples are included here. The full materials are available in the CD.

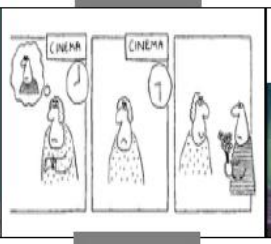
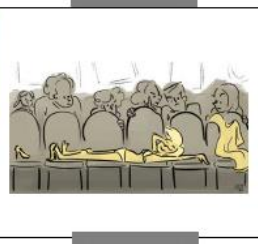
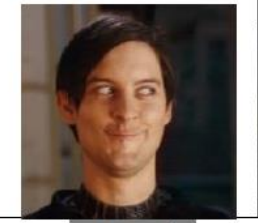
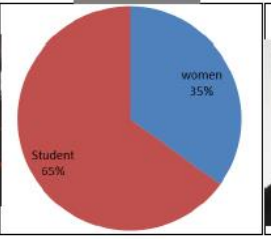
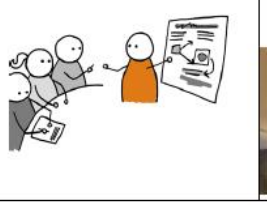
Appendix 9

Study 2

The picture description activity **Session 1**

Student A: Describe the picture with asterisk (*) to your partner. You also need to listen to your partner talking about her pictures. If the picture is the same as your partner's, then you both write (S) beside it. If the picture is different from your partner's picture, then you both write (D) beside the picture. The following words can help you describe the pictures:

blow **make up** **run** **take up** **commit** **hold back**

1 __ *	2 __	3 __ *	4 __	5 __	6 __ *
					
7 __ *	**8 __**	**9 __ ***	**10 __**	**11 __ ***	**12 __**
					
13 __ *	**14 __**	**15 __**	**16 __ ***	**17 __**	**18 __ ***
					

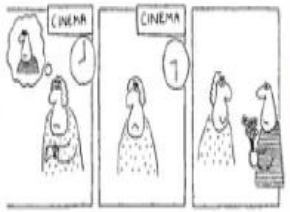
Student B: Describe the picture with asterisk (*) to your partner. You also need to listen to your partner talking about her pictures. If the picture is the same as your partner's, then you both write (S) beside it. If the picture is different from your partner's picture, then you both write (D) beside the picture. The following words can help you describe the pictures:

blow **make up** **run** **take up** **commit** **hold back**

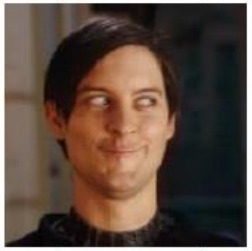
1 __ 2 __ * 3 __ 4 __ * 5 __ * 6 __



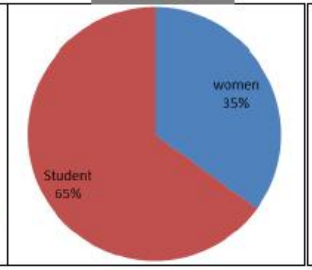




7 __ 8 __ * 9 __ 10 __ * 11 __ 12 __ *

13 __ 14 __ * 15 __ * 16 __ 17 __ * 18 __

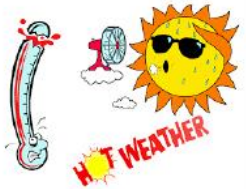
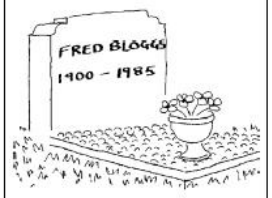
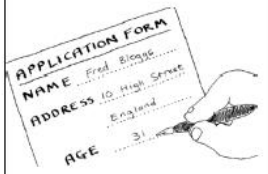
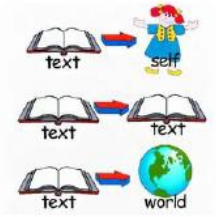
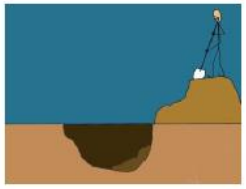






The picture description activity **Session 2**

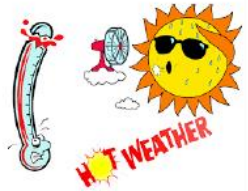
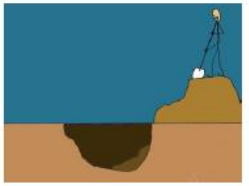
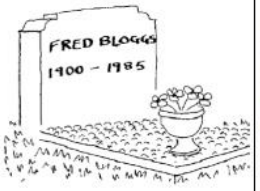
Student A: Describe the picture with asterisk (*) to your partner. You also need to listen to your partner talking about her pictures. If the picture is the same as your partner's, then you both write (S) beside it. If the picture is different from your partner's picture, then you both write (D) beside the picture. The following words can help you describe the pictures:

hit **fill in** **train** **come around** **relate** **pass on**

1 __ *	2 __	3 __ *	4 __	5 __	6 __ *
					
7 __ *	**8 __**	**9 __ ***	**10 __**	**11 __ ***	**12 __**
					
13 __ *	**14 __**	**15 __**	**16 __ ***	**17 __**	**18 __ ***
					

Student B: Describe the picture with asterisk (*) to your partner. You also need to listen to your partner talking about her pictures. If the picture is the same as your partner's, then you both write (S) beside it. If the picture is different from your partner's picture, then you both write (D) beside the picture. The following words can help you describe the pictures:

hit **fill in** **train** **come around** **relate** **pass on**

1 __	2 __ *	3 __	4 __ *	5 __ *	6 __
					
7 __	**8 __ ***	**9 __**	**10 __ ***	**11 __**	**12 __ ***
					
13 __	**14 __ ***	**15 __ ***	**16 __**	**17 __ ***	**18 __**
					

Note. Only samples are included here. The full materials are available in the CD.

Appendix 10

Study 2

The listen-and-retell meaning activity **Session 1**

Student A

Part A

A. Your partner is going to read some sentences to you. Listen carefully. Try to answer the questions she is going to ask about the word meaning and translation.

Part B

B. Read the following sentences to your partner. Ask her about the meaning and translation of the underlined words. Tick (√) Yes for correct answers. Tick (√) No for wrong answers or no answers.

10.	Women **make up** nearly 50% of law school students.			
	(A) Meaning?	form	☐ Yes	☐ No
	(B) Translation?	يُشكل شيئاً ما	☐ Yes	☐ No
11.	He was **committed** to a state prison.			
	(A) Meaning?	put for keeping	☐ Yes	☐ No
	(B) Translation?	أودع شخصاً ما السجن	☐ Yes	☐ No
12.	I was thinking about **taking up** photography.			
	(A) Meaning?	doing something for pleasure	☐ Yes	☐ No
	(B) Translation?	يُباشر/يتّخذ هواية جديدة	☐ Yes	☐ No
13.	He has no idea how to **run** the hotel.			
	(A) Meaning?	manage	☐ Yes	☐ No
	(B) Translation?	تولى إدارة شغلاً ما	☐ Yes	☐ No
14.	If your coffee is too hot, **blow** on it before taking a sip.			
	(A) Meaning?	send out air	☐ Yes	☐ No
	(B) Translation?	يزفر (يخرج الهواء)	☐ Yes	☐ No
15.	She just managed to **hold back** her anger.			
	(A) Meaning?	stop showing this emotion	☐ Yes	☐ No
	(B) Translation?	يكبت نفسه/يكبح مشاعره	☐ Yes	☐ No
16.	I **run** about three miles every morning.			
	(A) Meaning?	move quickly	☐ Yes	☐ No
	(B) Translation?	يركض	☐ Yes	☐ No
17.	Those big desks **take up** too much space.			
	(A) Meaning?	use space or time	☐ Yes	☐ No
	(B) Translation?	يستحوذ على شيئاً ما (مساحة/وقت)	☐ Yes	☐ No
18.	Her bad health **held** her **back** from being more successful.			
	(A) Meaning?	prevent from progress	☐ Yes	☐ No
	(B) Translation?	يقيّد/يمنع/يعيق شخصاً (من التقدم)	☐ Yes	☐ No

Student B

Part A

A. Read the following sentences to your partner. Ask her about the meaning and translation of the underlined words. Tick (√) Yes for correct answers. Tick (√) No for wrong answers or no answers.

1.	I haven't **made up** my mind whether to go yet.			
	(A) Meaning?	decide	☐ Yes	☐ No
	(B) Translation?	يقرر	☐ Yes	☐ No
2.	You **blew** your second chance by arriving late for the interview.			
	(A) Meaning?	lose chance	☐ Yes	☐ No
	(B) Translation?	يفسد امراً ما/يُضيع فرصة	☐ Yes	☐ No
3.	He **committed** himself to charity works and helping others.			
	(A) Meaning?	obligate, promise	☐ Yes	☐ No
	(B) Translation?	يتعهد بفعل شيئاً ما	☐ Yes	☐ No
4.	In Egypt, the Nile River **runs** into the Mediterranean Sea.			
	(A) Meaning?	flow (liquid)	☐ Yes	☐ No
	(B) Translation?	يجري (سوائل)	☐ Yes	☐ No
5.	Recently, terrorists **committed** some murders in some mosques in Saudi.			
	(A) Meaning?	do something wrong	☐ Yes	☐ No
	(B) Translation?	يرتكب خطأ او جريمة	☐ Yes	☐ No
6.	Let's **take up** this problem with the manager.			
	(A) Meaning?	discuss	☐ Yes	☐ No
	(B) Translation?	يُناقش (أمراً) مع	☐ Yes	☐ No
7.	The witness **held back** important information about the case and kept silent.			
	(A) Meaning?	don't say something	☐ Yes	☐ No
	(B) Translation?	يتحفظ/ يتكتّم على شئ ما	☐ Yes	☐ No
8.	Sara **blew** all her Eid money on a diamond ring.			
	(A) Meaning?	spend money carelessly	☐ Yes	☐ No
	(B) Translation?	يبدد المال/يُبذر	☐ Yes	☐ No
9.	I forget Sara's birthday yesterday so I'll try to **make** it **up** to her by taking her to a nice place.			
	(A) Meaning?	compensate	☐ Yes	☐ No
	(B) Translation?	يُعوّض عن شيئاً ما	☐ Yes	☐ No

Part B

B. Your partner is going to read some sentences to you. Listen carefully. Try to answer the questions she is going to ask about the word meaning and translation.

The listen-and-retell meaning activity **Session 2**

Student A

Part A

A. Your partner is going to read some sentences to you. Listen carefully. Try to answer the questions she is going to ask about the word meaning and translation.

Part B

B. Read the following sentences to your partner. Ask her about the meaning and translation of the underlined words. Tick (√) Yes for correct answers. Tick (√) No for wrong answers or no answers.

10.	You should **come around** to our house more often.			
	(A) Meaning?	come near	☐ Yes	☐ No
	(B) Translation?	يمُرّ على/ يزور	☐ Yes	☐ No
11.	All staff should be **trained** in first aid.			
	(A) Meaning?	teach	☐ Yes	☐ No
	(B) Translation?	يُعلّم (يُدرّب)	☐ Yes	☐ No
12.	The company will **fill in** the holes with concrete.			
	(A) Meaning?	put materials into something	☐ Yes	☐ No
	(B) Translation?	يملأ شيئاً ما	☐ Yes	☐ No
13.	A sudden storm **hit** the ship and destroyed it.			
	(A) Meaning?	touch quickly with force	☐ Yes	☐ No
	(B) Translation?	يصدم/ يضرب بسرعة وبشدة	☐ Yes	☐ No
14.	You must be feeling awful. I went through something similar myself last year, so I can **relate** to your feelings.			
	(A) Meaning?	feel sympathy with someone	☐ Yes	☐ No
	(B) Translation?	يتفهم/يستوعب/يتعاطف مع	☐ Yes	☐ No
15.	He **trained** the lights into the hole to see what is moving inside.			
	(A) Meaning?	point	☐ Yes	☐ No
	(B) Translation?	يُصوّب/يُوجّه	☐ Yes	☐ No
16.	Her father **passed on** the house and all its belongings to her when he died.			
	(A) Meaning?	inherit	☐ Yes	☐ No
	(B) Translation?	يُورث	☐ Yes	☐ No
17.	At first she didn't like the idea, but she **came around** to our way of thinking in the end.			
	(A) Meaning?	change one's opinion	☐ Yes	☐ No
	(B) Translation?	يغير رأيه/ يقتنع برأي شخصاً آخر	☐ Yes	☐ No
18.	We listened carefully as she **related** the whole exciting story.			
	(A) Meaning?	tell a story	☐ Yes	☐ No
	(B) Translation?	يحكي/يروي	☐ Yes	☐ No

Student B

Part A

A. Read the following sentences to your partner. Ask her about the meaning and translation of the underlined words. Tick (√) Yes for correct answers. Tick (√) No for wrong answers or no answers.

1.	Your father has been **training** hard for the race for several weeks now.			
	(A) Meaning?	exercise	☐ Yes	☐ No
	(B) Translation?	يتدرب/يتمرّن (رياضة)	☐ Yes	☐ No
2.	Eid al-Fitr will **come around** again after Ramadan.			
	(A) Meaning?	happen again (regular)	☐ Yes	☐ No
	(B) Translation?	يتكرر ، يُعاد (الحدث أو الزمن)	☐ Yes	☐ No
3.	The economy has been **hit** badly by high unemployment of young people.			
	(A) Meaning?	Affect badly	☐ Yes	☐ No
	(B) Translation?	يسبب عناءً(ضرر)/ يؤثر عليه	☐ Yes	☐ No
4.	Khaled will **fill me in** on the latest changes in the company.			
	(A) Meaning?	give extra information	☐ Yes	☐ No
	(B) Translation?	يُطلع شخصاً ما بالجديد (ماقاته)	☐ Yes	☐ No
5.	Just **pass on** the magazine to someone else when you finish.			
	(A) Meaning?	give a thing to someone	☐ Yes	☐ No
	(B) Translation?	تناقل الشيء من شخص إلى آخر	☐ Yes	☐ No
6.	Temperatures **hit** 50° degrees yesterday.			
	(A) Meaning?	reach	☐ Yes	☐ No
	(B) Translation?	يصل لمستوى (هبوط/ ارتفاع)	☐ Yes	☐ No
7.	Ahmad had an emergency so I will **fill in** for him.			
	(A) Meaning?	Substitute for someone	☐ Yes	☐ No
	(B) Translation?	يحل محل شخصاً ما	☐ Yes	☐ No
8.	His stories never seem to **relate** to anything in the real world.			
	(A) Meaning?	Show connection	☐ Yes	☐ No
	(B) Translation?	يربط بين شيء وشيء	☐ Yes	☐ No
9.	Jim's uncle **passed on** last year. He was just 35 and he left two children behind.			
	(A) Meaning?	die	☐ Yes	☐ No
	(B) Translation?	يتوفى/ يموت	☐ Yes	☐ No

Part B

B. Your partner is going to read some sentences to you. Listen carefully. Try to answer the questions she is going to ask about the word meaning and translation.

Note. Only samples are included here. The full materials are available in the CD.

Appendix 11

Study 2

The ILH rating sheet for the activities

The participant first received a sample of each activity presented in Appendices 8-10.

Please rate the activity according to the following:

Need:	0	1	2
Whether knowledge of words is required to complete a task?	if the activity does not impose any requirement for knowing the word.	if need is imposed by an external factor: the teacher, the requirement of the activity, etc.	if need is imposed by learners themselves.
Search:	0	1	
The attempt learners make to determine the meaning of unknown words in an activity.	when learners must seek the meaning of unknown words to complete a task.	when learners do not need to seek the meaning of unknown words to complete a task.	
Evaluation:	0	1	2
It entails comparing a new word with other words and making a decision as to its suitability in a given context.	No any sort of evaluation is involved.	when learners must recognize differences between words provided in a given context.	when the activity requires making decisions about new words and combining them with known words in original contexts.

Appendix 12

Study 2

The TFA rating sheet for the activities

The participant first received a sample of each activity presented in Appendices 8-10

Motivation		
1. Is there a clear vocabulary learning goal?	0	1
2. Does the activity motivate learning?	0	1
3. Do the learners select the words?	0	1
Noticing		
4. Does the activity focus attention on the target words? *Attention draw to words through bolding, underlining, glossing words in context, or through instructions that require learners to pay attention to words*	0	1
5. Does the activity raise awareness of new vocabulary learning? *If the activity raise awareness to learning words (e.g., meeting words in original sentences, selecting words from choices, using words in context)*	0	1
6. Does the activity involve negotiation?	0	1
Retrieval		
7. Does the activity involve retrieval of the word? *If both form and meaning are provided in the activity, then the activity receives 0* *If the activity require learners to retrieve either form or meaning, then the activity receive 1*	0	1

8. Is it productive retrieval? *If the activity requires learners to retrieve the meaning (Receptive retrieval), then the activity receives 0* *If the activity requires learners to retrieve the form (Productive retrieval), then the activity receives 1*	0	1
9. Is it recall? *If the activity require recognition only, i.e., when the correct words or meaning is retrieved through seeing or hearing a number of choices, then the activity receives 0* *If there are no choices in the activity and learners need to retrieve the word or meaning from memory, then the activity receives 1*	0	1
10. Are there multiple retrievals of each word?	0	1
11. Is there spacing between retrievals?	0	1
Generation		
12. Does the activity involve generative use? *Receptive generative use involves meeting a word in new context whereas productive generative use refers to using a word in a new way that learners has not met before.* *If the activity does not involve any kind of generative use, then the activity receives 0.* *If the activity involves any kind of generative use, then the activity receives 1.*	0	1
13. Is it productive? *If the activity involves receptive generative use, then the activity receives 0.* *If the activity involves productive generative use, then the activity receives 1.*	0	1
14. Is there a marked change that involves the use of other words?	0	1

Retention		
15. Does the activity ensure successful linking of form and meaning?	0	1
16. Does the activity involve instantiation? *Instantiation involves seeing an instance of a word such as when the word is used in a meaningful situation where the object, action, or quality referred to is visually present.*	0	1
17. Does the activity involve imaging? *Imaging involves deliberately seeing or imaging a visual image related to the meaning of the word.*	0	1
18. Does the activity avoid interference? *If two or more unknown items are related in form or meaning in the activity, then the activity receives 0.* *If the activity avoid any interference of forms and meaning of the unknown items, then activity receives 1.*	0	1

Appendix 13

Study 2

The receptive test (form-meaning recognition)

Name _____ الاسم

Group _____ الشعبة

Instructions

- This is a vocabulary test. You must choose the right word to go with each meaning.
- هذا اختبار للمفردات لذا يجب عليكِ اختيار الكلمة الصحيحة التي تتماشى مع كل معنى
- Match the word in the left column with the meaning in the centre column. Shade the letter of the correct answer.
- صلي الكلمات في العمود الأيسر بمعناها في العمود الوسط ثم ظللي حرف الإجابة الصحيحة
- You answer it in the following way
- تستطيعِ الإجابة كالنحو التالي

A. feed **B.** claim **C.** leave **D.** come down **E.** make out **F.** break off	1. ينفصل عن شيء 2. يُشجّع شخصاً ما 3. يُغادر	1.	Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ ●
		2.	● Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ
		3.	Ⓐ Ⓑ ● Ⓓ Ⓔ Ⓕ

- Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are **claim**, **come down**, **make out**
- هناك بعض الكلمات موجودة في الإختبار لجعله أكثر صعوبة لذا لا يجب عليكِ البحث عن معاني هذه الكلمات

الكلمات في المثال أعلاه **claim**, **come down**, **make out** تعتبر مثال لهذه الكلمات

- Try to do every part of the test.
- حاولي الإجابة على كل أجزاء الإختبار

A. commit **B.** run **C.** blow **D.** make up **E.** take up **F.** hold back	1. يركض 2. يستحوذ على شيئاً ما (مساحة/وقت) 3. يبدد المال/يُبذر	1. A ● C D E F 2. A B C D ● F 3. A B ● D E F

A. train **B.** hit **C.** relate **D.** pass on **E.** come around **F.** fill in	4. يحل محل شخصاً ما 5. يصدم/ يضرب بسرعة وبشدة 6. يربط بين شيء وشيء	4. A B C D E ● 5. A ● C D E F 6. A B ● D E F

A. commit **B.** run **C.** blow **D.** pass on **E.** come around **F.** fill in	7. يعّرج على/ يزور 8. تناقل الشيء من شخص إلى آخر 9. يرتكب خطأ او جريمة	7. A B C D ● F 8. A B C ● E F 9. ● B C D E F

A. train **B.** hit **C.** relate **D.** make up **E.** take up **F.** hold back	10. يتفهم/يستوعب/يتعاطف مع 11. يُعّلم (يُدّرب) 12. يُرّكب/ يُجّمع/ يُشكل شيئاً ما	10. A B ● D E F 11. ● B C D E F 12. A B C ● E F

A. run **B.** blow **C.** train **D.** take up **E.** hold back **F.** pass on	13. يتدرب/يتمرّن (رياضة) 14. يُورث 15. تولى إدارة شغلاً ما	13. Ⓐ Ⓑ ● Ⓓ Ⓔ Ⓕ 14. Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ ● 15. ● Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ	

A. hit **B.** relate **C.** commit **D.** come around **E.** fill in **F.** make up	16. يحكي/يروي 17. يُطّلع شخصاً ما بالجديد (أو مافاته) 18. يغير رأيه/ يقتنع برأي شخصاً آخر	16. Ⓐ ● Ⓒ Ⓓ Ⓔ Ⓕ 17. Ⓐ Ⓑ Ⓒ Ⓓ ● Ⓕ 18. Ⓐ Ⓑ Ⓒ ● Ⓔ Ⓕ	

A. run **B.** blow **C.** train **D.** come around **E.** fill in **F.** make up	19. يُصوّب/يُوجّه 20. يُعوّض عن شيئاً ما 21. يتكرر، يُعاد (الحدث أو الزمن)	19. Ⓐ Ⓑ ● Ⓓ Ⓔ Ⓕ 20. Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ ● 21. Ⓐ Ⓑ Ⓒ ● Ⓔ Ⓕ	

A. hit **B.** relate **C.** commit **D.** take up **E.** hold back **F.** pass on	22. يتعهد بفعل شيئاً ما 23. يتحفظ/ يتكتّم على شئ ما 24. يُناقش (أمراً) مع	22. Ⓐ Ⓑ ● Ⓓ Ⓔ Ⓕ 23. Ⓐ Ⓑ Ⓒ Ⓓ ● Ⓕ 24. Ⓐ Ⓑ Ⓒ ● Ⓔ Ⓕ	

A. blow **B.** train **C.** hit **D.** hold back **E.** pass on **F.** come around	25. يسبب عناءً(ضرر)/ يؤثر عليه 26. قيّد/يمنع/يعيق شخصاً (من التقدم) 27. يزفر(يخرج الهواء)	25. (A) (B) ● (D) (E) (F) 26. (A) (B) (C) ● (E) (F) 27. ● (B) (C) (D) (E) (F)

A. relate **B.** commit **C.** run **D.** fill in **E.** make up **F.** take up	28. يقرّر 29. يُباشر/يتخذ هواية جديدة 30. يجري (سوائل)	28. (A) (B) (C) (D) ● (F) 29. (A) (B) (C) (D) (E) ● 30. (A) (B) ● (D) (E) (F)

A. blow **B.** train **C.** hit **D.** fill in **E.** make up **F.** take up	31. يصل لمستوى أوحالة معينة (هبوط أو ارتفاع) 32. يفسد امراً ما/يُضيع فرصة 33. يملأ شيئاً ما	31. (A) (B) ● (D) (E) (F) 32. ● (B) (C) (D) (E) (F) 33. (A) (B) (C) ● (E) (F)

A. relate **B.** commit **C.** run **D.** hold back **E.** pass on **F.** come around	34. أودع شخصاً ما السجن (مؤسسة) 35. يكبت نفسه/يكبح مشاعره 36. يتوفى/ يموت يتوفى/ يموت	34. (A) ● (C) (D) (E) (F) 35. (A) (B) (C) ● (E) (F) 36. (A) (B) (C) (D) ● (F)

A. fall		37. (A) (B) ● (D) (E) (F)

B. clear **C.** introduce **D.** hold up **E.** back up **F.** get through	37.	يقدم (يطرح) شئ للمرة الأولى		
	38.	يرفع إلى الأعلى	38.	A B C ● E F
	39.	يسقط/يقع	39.	● B C D E F

A. collect **B.** count **C.** settle **D.** put up **E.** set off **F.** take in	40.	ينطلق في رحلة	40.	A B C D ● F
	41.	يحُلّ/ يحسم (مشكلة أو مسألة ما)	41.	A B ● D E F
	42.	يعدّ/ يُحصي	42.	A ● C D E F

A. fall **B.** clear **C.** introduce **D.** put up **E.** set off **F.** take in	43.	يُعرّف (يُقدّم) شخص	43.	A B ● D E F
	44.	يُسكن شخصاً في بيته	44.	A B C D E ●
	45.	يُعَلِّق/يرفع شيئاً	45.	A B C ● E F

A. collect **B.** count **C.** settle **D.** hold up **E.** back up **F.** get through	46.	يجتاز	46.	A B C D E ●
	47.	يُضمّن (يُدخل في الحسبان)	47.	A ● C D E F
	48.	يتقهقر(يرّتد)	48.	A B C D ● F

A. clear **B.** introduce **C.** collect **D.** back up **E.** get through **F.** put up	49.	يبني / يُشّيد	49.	A B C D E ●
	50.	يجمع	50.	A B ● D E F
	51.	يُخلي	51.	● B C D E F

A. count **B.** settle **C.** fall **D.** set off **E.** take in **F.** hold up	52.	يُفجّر(القنبلة)	52.	A B C ● E F
	53.	يستوعب، يفهم	53.	A B C D ● F
	54.	ينخفض/ يصبح اقل	54.	A B ● D E F

A. clear **B.** introduce **C.** collect **D.** set off **E.** take in **F.** hold up	55.	يُقلّ شخصاً/شيئاً من مكان ما	55.	A B ● D E F
	56.	يتحمل/يصبر/يحافظ على قوته	56.	A B C D E ●
	57.	يجتاز بدون لمس	57.	● B C D E F

A. count **B.** settle **C.** fall **D.** back up **E.** get through **F.** put up	58.	يستقرّ/يستوطن	58.	A ● C D E F
	59.	يوصل فكرته للشخص الآخر	59.	A B C D ● F
	60.	يساعد/ يساند	60.	A B C ● E F

A. introduce **B.** collect **C.** count **D.** get through **E.** put up **F.** set off	61. يهُـمّ/ يعتبر أهم 62. تسبب في بداية او حدوث شيئاً ما 63. يضيف شيئ على شيئ ما	61.  62.  63. 

A. settle **B.** fall **C.** clear **D.** take in **E.** hold up **F.** back up	64. يثُبت براءة /يُبرَّأ 65. يصبح (يتغير إلى حالة سلبية) 66. يُعطّل/ يؤخر (شخصاً أو شيئاً)	64.  65.  66. 

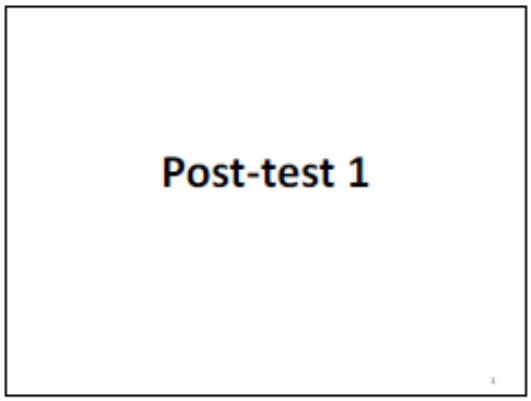
A. introduce **B.** collect **C.** count **D.** take in **E.** hold up **F.** back up	67. يدعم/يُثبت صحته 68. يجمع أفكاره (يستجمع نفسه) 69. يخدع شخصاً	67.  68.  69. 

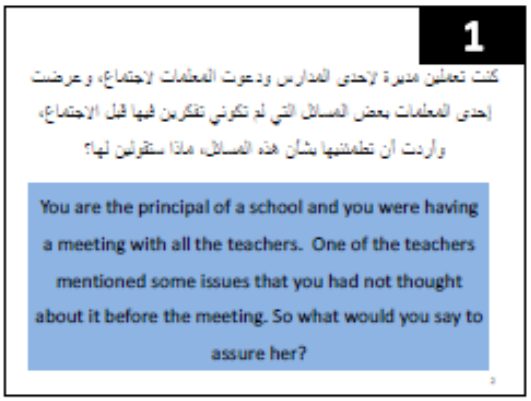
A. settle **B.** fall **C.** clear **D.** get through **E.** put up **F.** set off	70. يتحمل/يصبر على شئ مزعج 71. يجلس بـ وضعية مريحة 72. يصل (يتواصل مع شخصاً) عبر الهاتف	70.  71.  72. 

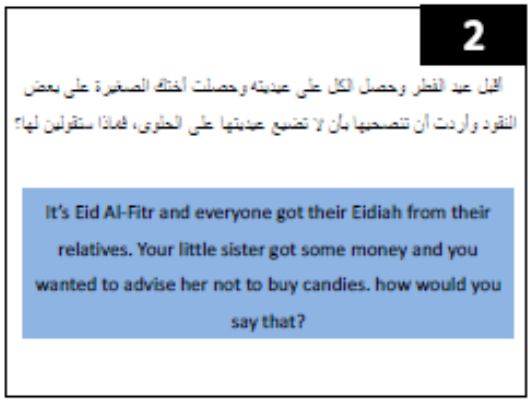
Appendix 14

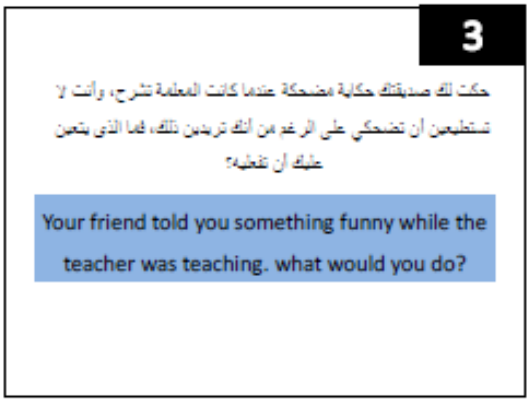
Study 2

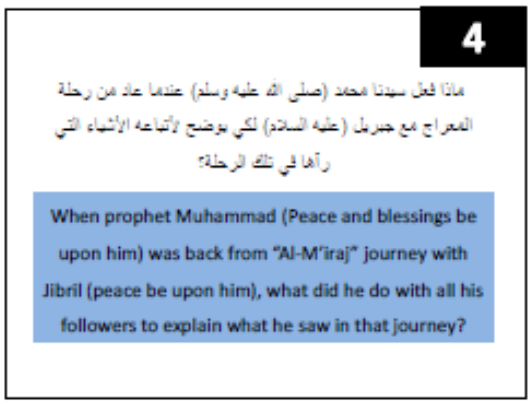
The productive test (form-meaning recall)

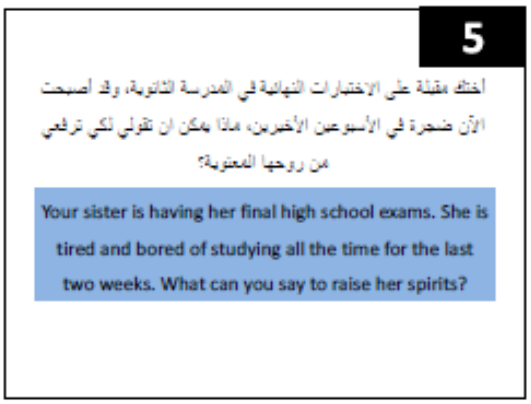












6

تمت إقامة مباريات بين فرق كرة القدم المدرسية في مدارس جدة. ما الذي ستلتقطه هذه المدارس مع فرقها؟

There is a football competition between all the football teams of Jeddah schools. What should these schools do with their teams?

7

انتقلتوا الى منزل جديد وكان مناسب جداً لكن بعض الجدران يوجد فيها حفر مسامير مما شوه منظر الجدار، ولا تتحمل ميزانيتكم الدفع لطلاء المنزل، ماذا يمكنك فعله لتغطية هذه الحفر؟

You moved to a new apartment which is the best option you found in the city. But there were nails holes in some of the walls and you can't afford to pay for new painting, what would you do with these nails holes?

8

عندما يمارس شخص فعلاً غير قانوني مثل السرقة وهو يعاني من مشكلة عقلية، فماذا ستفعل المحكمة؟

When somebody do something illegal such as stealing but he has a mental problem. What do you think the court should do?

9

كنت تواسين إحدى صديقاتك التي توفيت والدتها المصابة بالسرطان وقد عانت منه لوقت طويل، فماذا ستقولين لها وتخبريها بأن والدتها الآن تعيش حياة أسعد من حياتها السابقة؟

You are trying to comfort a friend that is going through a bad time because of her mother's death, who had cancer and was suffering for a long time. What might you say to her?

10

الصيف في السعودية حار للغاية إذ أن درجة الحرارة تكون أحياناً عالية جداً، فما الدرجة التي قد تصل إليها؟

Summer in Saudi Arabia is very hot. The temperature is sometimes so high. How high it can be?

11

كان لديك إختبار يبدأ الساعة الثامنة صباحاً، وصلت قبل الإختبار بخمس دقائق وانت تعلمين أن الطالبات المتأخرات لا يسمح لهن بدخول الاختبار، فماذا ستفعلين لكي تصلي في الوقت المطلوب؟

You have an early exam that starts at 8 am at the university. You arrived five minutes before the exam and you know that the examiner would not allow any late students to enter. What would you do to enter the room on time?

12

فعلت شيئاً أغضب والدتك واعتذرت لها شفوياً ولكنها لم تقبل ذلك، وبدأت تفكرين في الطريقة التي تتالين بها رضاها فقررت في شراء بعض الزهور مصحوبة ببطاقة مكتوب عليها "إنني آسفة حقاً"، لماذا تفعلين ذلك؟

You did something that made your mom angry and you apologised many times but she is still angry. You are thinking how you can make it right and you decided to buy her some flowers with a card saying you are really sorry. Why are you doing that?

13

توفي جارك في الشهر الماضي وترك وراءه ابنين، وكان رجلاً ثرياً وترك وراءه ثروة كبيرة، فمن الذي سيرث هذه الثروة؟

You neighbour passed away last month leaving two sons behind. He was a rich man and he left a fortune behind, to whom do you think all his fortune will go?

14

في يوم من الأيام قبل أن تخرج والدتك طلبت منك أن تنظفي البيت قبل أن تعود، ولكنك كنت تذاكرين في ذلك الوقت ولم تستطيعي أن تفعلي ذلك قبل أن تعود، تريدين ان تخبريها بذلك كيف ستقولين لها ذلك؟

Before your mom go out one day she asked you to clean the house before she gets back. You were studying all the time and you couldn't do that before she's back. What would you say to her?

15

توفيت والدتك عندما كان عمرك 7 سنوات، ولديك ثلاثة من الإخوة وأخت عمرها 18 سنة، وقد قامت بكل شيء كانت تفعله والدتك لكي تعتني بكم ولذلك لم تتزوج حتى تكبروا، لماذا فعلت ذلك في الأساس؟

Your mother passed away when you were 7 years old. You have three younger brothers and one 18-years sister. She took care of you all as your mom and didn't get married until you all got older. Why she did that?

16

لم تحضري أحد الدروس وهو الدرس الذي اخبرت فيه المعلمة معلومات مهمه عن الاختبارات ماذا ستفعلين؟

You missed one of your classes and it was the one your teacher gave important information for the final exam. You may ask you friend to ……

17

كنت تناقشين مسائل معينة مع صديقاتك واتفقتن على رأي ما عدا أحدى صديقاتك المعروفة بالعناد ولكنك تعلمين أنها سوف تقتنع في النهاية لأنك كنت دائماً تقومين بإقناعها كيف تشرحي ذلك لصديقاتك؟

You were discussing something with your friends. You all come to agree on something except one of you who didn't agree. You know that she will agree at the end because she always does that. How can you explain that to your friends?

18

تمر صديقتك بأزمة نفسية بسبب موت أخيها، وكنت أنت قد مررت بأزمة مماثلة في السنة الماضية، ما هو شعورك الآن؟

Your friend is going through some bad time because of her brother's death. You went through a similar thing last year. How would you feel about it?

19

يمتلك والدك شركة خاصة كان يعمل فيها منذ 30 سنة، ويريد أن يتقاعد ويترك مكانه في الشركة، ويدرس أخوك إدارة أعمال فما الذي سيفعله أخوك في الشركة في رأيك؟

Your father owns a private company and he has been working there for 30 years now. He wants to retire now. Your brother is studying Business Management. What do you think your brother will do there?

20

عندما يتحرى ضابط الشرطة بشأن جريمة حدثت أمام الشهود، ما هو أول شيء يقوله للشاهد عندما يطلب منه أن يسرد ما رآه؟

When a police officer investigates a murder that happened in front of a witness, what is the first thing they tell that witness when he asks them?

21

كانت العادة في الزفاف أنه عندما تبدأ في السير في الممر يتم إطفاء كل الأنوار، فكيف يستطيع الناس أن يروا العروس، ماذا يستخدمون لذلك؟

In a wedding, when the bride starts walking down the aisle, all the venue lights will be turned off. How can everyone see the bride? What they usually do?

22

عادة عندما تذهب أختك إلى سريرها لتنام تشعل شمعات خلف سريرها، وفي الليلة الماضية تركت الشمعة مشتعلة طوال الليل وسببت الرائحة إزعاجاً لك أثناء النوم، ماالذي ستفعلينه حتى تكملي نومك براحة أكثر؟

Your sister usually lights some candles in the room before she goes to bed. She left one of them burning and the smell annoyed you. You couldn't sleep, what would you do with the candle?

23

تقدم شاب للزواج من أختك وكان يعمل طبيباً في مستشفى ولكنها كانت تريد أن تتزوج من قريبها الذي لا يعمل حتى الآن. أنت تعتقدين أنه ليس جاد في ذلك وتريدين منها أن تقرر ما إذا كانت تريد الزواج منه أم لا، كيف ستقولين لها ذلك؟

A young man proposed to your sister. He works as a doctor in the city hospital but she always thought that she would marry your cousin who doesn't have a job yet and you feel that he is not serious with her. You want to tell her that she needs to decide. How would you say that?

24

كان أخوك يقود السيارة في جدة وأنت معه بسرعة شديدة ثم جاءت سيارة من الأمام وكادت أن تصطدم بسيارتكم مما جعل أخوك ينحرف إلى جانب الطريق لكي يتجنبها، ثم رأيت الحواجز الجانبية أمامك ماذا سيحدث؟

Your brother was driving you to Jeddah. He was driving so fast and suddenly a car come in front of you, and your brother had to slide to the side to avoid crashing that car. You saw the street barriers in front of you, What likely will happen?

25

سمعت بشخص أمضى في السجن 16 عاماً وتم الآن إطلاق سراحه إذ اتضح أنه بريء بعد كل هذه السنوات، وأنت تشعرين الآن بأسف لذلك، لماذا؟

You heard about a man who spent 16 years in prison and now he has been released because he was found innocent after all these years. You felt sorry for that man, why?

26

رسبت أختك في الفترة الدراسية الأولى في الجامعة في إحدى المواد، ولم تكن قد رسبت من قبل في أي مرحلة من مراحل الدراسة، والآن فقد أصبحت مكتئبة بسبب هذا الرسوب الوحيد ولم تعد تذاكر بجد، كيف يؤثر عليها هذا الاكتئاب؟

Your sister failed in the first semester in the university in one course. She has never failed in all her study life and this was her first time. Now, she is depressed thinking about that single failure & doesn't study hard anymore. What do you think her depression might do to her?

27

تقوم صديقتك ووالدتها بزيارتكم لأول مرة، ووالدتك تعرف دائماً كيف تؤانس والدتها ولذلك فقد أمضوا اليوم كله في منزلكم، وعندما أردتما أن تغادرا المنزل أردت أنت أن تقترحي تكرار الزيارة، كيف ستقولين ذلك؟

Your friend and her mother visit you for the first time. Your mom get along with her mom and they spent the whole day at your house. They are leaving now and you want to suggest that they should do it again, what would you say?

28

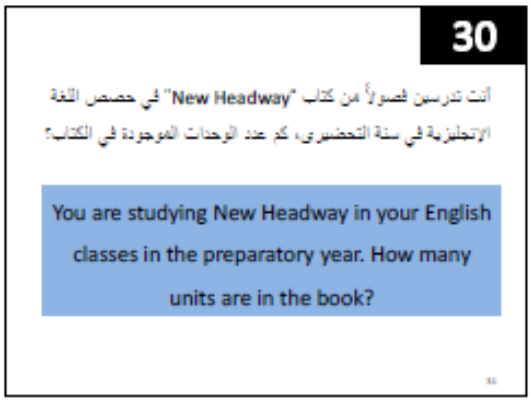
كنت تعملين في إحدى المدارس الابتدائية عندما أنجبت إحدى المعلمات طفلاً، وقامت المدرسة بمنحها عطلة للعناية بطفلها ولكن المدرسة لم تحدد ماذا ستفعل بشأن الحصص التي تدرسها هذه المعلمة، فماذا ستفعل المدرسة في رأيك؟

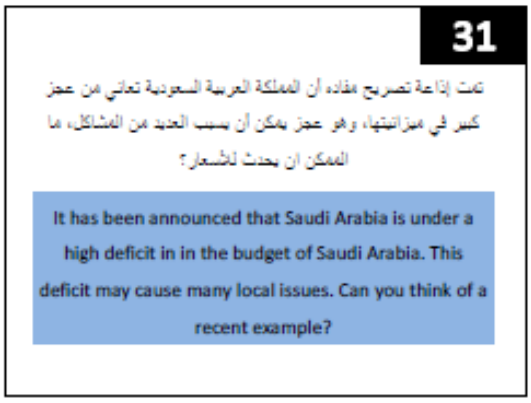
You are working in an elementary school and one of the teachers at school just had a baby. The school gave her maternity leave. What do you think the school will do about the classes she is teaching?

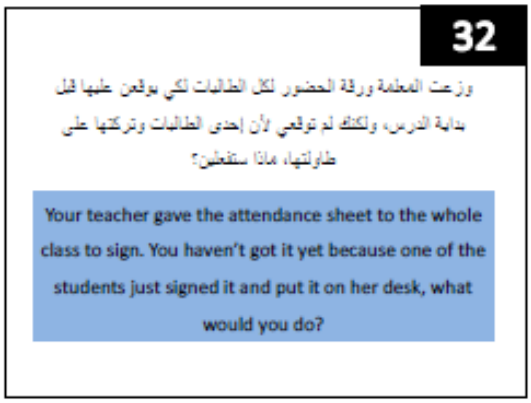
29

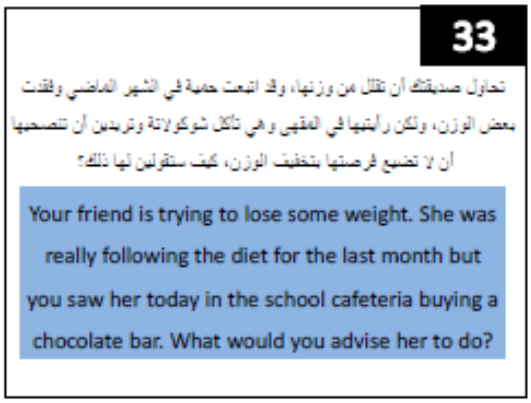
تمضي أختك وقتاً طويلاً في العمل في المكتب وفي المنزل، وقد أصابها الملل وتريد أن تفعل شيئاً مسلياً، تريدين نصحها بأن تتخذ هوايه جديدة، كيف ستقولين ذلك؟

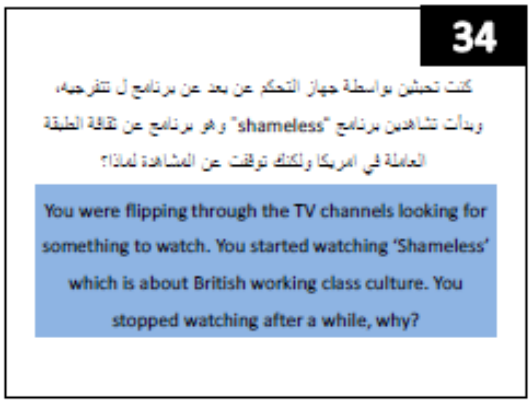
Your sister spends a lot of her time working either at work or at home. She is really bored and wants to do something interesting. What can you advise her?

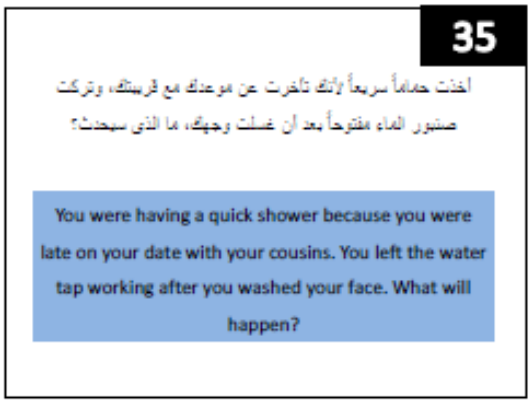












Note. Only samples are included here. The full materials are available in the CD.