Second Language Learning and Teaching

Breno B. Silva

Writing to Learn Academic Words

Assessment, Cognition, and Learning

Springer

Second Language Learning and Teaching

The series brings together volumes dealing with different aspects of learning and teaching second and foreign languages. The titles included are both monographs and edited collections focusing on a variety of topics ranging from the processes underlying second language acquisition, through various aspects of language learning in instructed and non-instructed settings, to different facets of the teaching process, including syllabus choice, materials design, classroom practices and evaluation. The publications reflect state-of-the-art developments in those areas, they adopt a wide range of theoretical perspectives and follow diverse research paradigms. The intended audience are all those who are interested in naturalistic and classroom second language acquisition, including researchers, methodologists, curriculum and materials designers, teachers and undergraduate and graduate students undertaking empirical investigations of how second languages are learnt and taught.

Breno B. Silva

Writing to Learn Academic Words

Assessment, Cognition, and Learning

Springer

Breno B. Silva
Institute of English Studies
University of Warsaw
Warsaw, Poland

ISSN 2193-7648 ISSN 2193-7656 (electronic)
Second Language Learning and Teaching
ISBN 978-3-031-06504-0 ISBN 978-3-031-06505-7 (eBook)
https://doi.org/10.1007/978-3-031-06505-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgements

I would like to thank all the students who participated in the studies reported in this book and the teachers who allowed me to come to their classrooms to collect the necessary data. I am also grateful to Prof. Agnieszka Otwinowska-Kasztelanic, the supervisor of my Ph.D. theses, on which this book is based. Her patience, commitment, and constant insightful input were invaluable and greatly appreciated. My appreciation is also extended to my colleague Katarzyna Kutyłowska, who helped me with some of the data collected in the studies and who provided valuable feedback on parts of the draft. Finally, I would like to thank the editors and anonymous reviewers from *English for Specific Purposes* and *Language Teaching Research*, the journals where two of the studies reported in this book have been published. Their input was invaluable to the quality of the studies.

Introduction

The importance of vocabulary to second language acquisition (SLA) has long been recognized and has attracted ever more attention from researchers, especially since the 1980s. Nevertheless, in order to function properly in a foreign language (FL), learners must acquire a large number of words. For example, current estimates suggest that FL learners may need to acquire about 34,000 words to perform satisfactorily in English. To achieve such knowledge, words must be acquired in a variety of ways. These include picking up words more indirectly (or incidentally) through meaning-focused activities such as reading or watching a movie; or learning words more directly (or intentionally) via word-focused tasks and language production (i.e. speaking and writing). Both indirect and direct learning are essential to help build one's lexical knowledge. However, research shows that knowledge of certain types of words, such as academic words, may benefit most from more direct learning. With this in mind, this book investigates the knowledge of academic vocabulary among university students and explores how writing tasks may benefit the acquisition of such words.

This book has two main goals. The first main goal is to assess the knowledge of academic words of Polish first- and second-year BA students at the Institute of English Studies, University of Warsaw. When doing this, I identify a tool that may be quickly and reliably used for placement purposes at the university. This tool is able to differentiate two groups of students, namely students who may need extra practice with academic words and learners who may not. The second main goal of this book is to explore the extent to which different writing tasks facilitate the acquisition of academic vocabulary. More specifically, I compare the lexical learning yielded by the writing of sentences and the writing of argumentative essays, with or without time pressure. I have chosen argumentative essays because they are tasks that pervade academic discourse. Therefore, the primary research findings discussed in this book may have clearer practical implications.

The first three chapters of this book discuss previous studies—both seminal and very recent findings—that are germane to the studies reported in later chapters. Chapter 1 of the literature review focuses on general characteristics surrounding lexical learning and assessment. Chapter 2 shifts attention to research on incidental

lexical learning, especially the learning generated through writing tasks. Chapter 3 underscores findings from L2 writing research and explores how characteristics that are intrinsic to the writing process may affect vocabulary learning.

Following the literature review, Chap. 4 provides the reader with a useful short overview of the entire project, that is, the three studies reported in this book. Chapter 4 then introduces some of the statistical analyses used in the project, divided into two main parts. The first part provides an outline on how basic statistical techniques were employed in the three studies. The second part introduces the reader to linear mixed-effects models, utilized in the second and third studies. Mixed-effects models are a more advanced and less common type of statistical analysis in applied linguistic studies, and this is why I considered it necessary to spend few pages describing this state-of-the-art statistical modelling technique.

As far as the studies are concerned, Chap. 5 reports on and discusses the findings of the first study, which measured the knowledge of academic words and developed a reliable tool for placement purposes, as outlined above. Chapters 6, 7, and 8 report on two other studies, which investigate compare the lexical learning potential of different writing tasks. Chapters 6 and 7 (i.e. the second and third studies) complement each other; therefore, their findings are discussed together in the general discussion provided in Chap. 8. Finally, Chap. 9 concludes the book. It lays out a number of practical implications, draws attention to some research limitations, and discusses promising venues for future research.

Importantly, the first and second studies reported in this book (Chaps. 5 and 6, and part of Chap. 8) are based on already published research. The first study was published in the journal *English for Specific Purposes* and is entitled "VST as a reliable academic placement tool despite cognate inflation effects" (Silva & Otwinowska, 2019). The second was published in the journal *Language Teaching Research* with the title "Learning academic words through writing sentences and compositions: Any signs of an increase in cognitive load?" (Silva et al., 2021). The reproduction of both papers in this book comply with the journals' copyright rules for authors' rights. The second and third authors of these studies—Agnieszka Otwinowska (both studies) and Katarzyna Kutyłowska (second study)—have agreed to the use of these papers in the current book.

In short, this book combines the results of three related studies towards one main goal: Improve the knowledge of academic vocabulary among Polish English majors. I have learnt enormously during the process of developing, conducting, and writing these studies. I believe that, despite limitations, the studies provide a valuable contribution to the existing body of research in the fields of SLA and L2 writing. More importantly, perhaps, they draw attention to possible new venues for future research.

Contents

Acronyms

AVT	Academic vocabulary test
AWL	Academic word list
BNC	British National Corpus
CEFR	Common European Framework of Reference for Languages
COCA	Corpus of Contemporary American English
CW	Composition writing
EAP	English for Academic Purposes
ILH	Involvement Load Hypothesis
L1	The first language (in order of acquisition) of a bilingual or multilingual speaker
L2	The second language (in order of acquisition) of a bilingual or multilingual speaker
SLA	Second-language acquisition
SW	Sentence writing
TESOL	Teaching English to Speakers of Other Languages
VLT	Vocabulary levels test
VST	Vocabulary size test
WM	Working memory

List of Figures

List of Tables

Chapter 1
The Importance of General and Academic Vocabulary Learning

1.1 Introduction

This chapter will explore the importance of lexical knowledge to second language learning and will highlight some of the challenges associated with vocabulary acquisition, especially in academic contexts. Importantly, the terms "learning" and "acquisition" will be used interchangeably throughout this book. Although some researchers differentiate between the terms (e.g., Krashen, 1982, 1989), the overwhelming majority do not. This is true in most fields connected to applied linguistics such as second language acquisition (SLA) research (see e.g., Chang, 2019; González-Fernández & Schmitt, 2020; Pellicer-Sánchez, 2019) and psycholinguistics (e.g., Elgort et al., 2018; Godfroid & Hui, 2020; Luke & Christianson, 2016).

First, this chapter will focus on general vocabulary, particularly on the number of words learners need to know in order to function properly in a language (i.e., vocabulary *breadth*). Then, it will shift attention to academic vocabulary, which is the focus of the studies reported in this book. Some benefits of and barriers to teaching and learning academic words at the university level will be discussed, followed by the need to accurately assess learners' knowledge of these words and the challenges involved in such undertaking.

1.2 The Essentialness of Lexical Learning and Its Challenges

The processes underlying second language (L2) lexical learning have been attracting ever-increasing attention since the 1980s, especially the 1990s (Laufer, 1989; Laufer & Nation, 2011). Nowadays, the essentialness of lexical knowledge to L2 language learning is universally accepted (Lewis, 1993; Nation, 2013; O'Dell, 1997; Schmitt, 2008, 2010; Sökmen, 1997). One of the reasons for this is the growing body of evidence demonstrating the importance of vocabulary learning to the development

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_1

of L2 skills. For example, L2 lexical knowledge has been found to correlate strongly with reading ($r = 0.69$–0.83; Hu & Nation, 2000; Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Milton et al., 2010; Paribakht & Webb, 2016; Stæhr, 2008) and listening comprehension ($r = 0.51$–0.70; Milton et al., 2010; Stæhr, 2008), as well with oral ($r = 0.63$; Milton et al., 2010) and written performance ($r = 0.73$–0.76; Milton et al., 2010; Paribakht & Webb, 2016; Stæhr, 2008, 2009). Put differently, L2 vocabulary knowledge may help explain from 26 to 69% of one's performance in English as a second language. This is a large proportion when considering the various other aspects involved in knowing a language. No wonder, then, that Vermeer (2001; p. 217) has asserted that "knowledge of words is now considered the most important factor in language proficiency and school success".

Unfortunately, acquiring a sufficient number of L2 English words may be a daunting task for most students. For instance, according to Schmitt (2007), Nation (2013) and Laufer and Nation (2011), the number of words that educated native speakers of English are able to understand (i.e., their *receptive lexical knowledge*) is around 20,000 word families. By word families I understand headwords with some of their inflections and derivations (Bauer & Nation, 1993). It is difficult to pinpoint exactly how many individual words this figure represents, not least because there is no agreement on what words should be included in a given family (Nation, 2013). Still, current estimates suggest that 8000 word families could amount to over 34,000 individual lexical items (Schmitt, 2010). This is a large number, but although far below the knowledge of educated native speakers, 8000 families may be what learners need to function independently in L2 English, as discussed below.

Initially, L2 English learners should focus on acquiring the most frequent 2000 word families (Meara, 1995; Nation, 2013), commonly referred to as *high-frequency* words. This is because knowledge of these words typically provides a coverage of, thus allowing learners to understand, at least 80% of the running words in both written and spoken texts (McCarthy, 1999; Nation & Waring, 1997; Read, 2004). Nevertheless, one word in five will still be unfamiliar, and this 20% has been shown to critically hinder understanding. In fact, researchers generally agree that understanding 98% of the running words in a text is the minimum necessary to attain satisfactory levels of unassisted comprehension (Hu & Nation, 2000; Schmitt et al., 2011).

Nation (2006) conducted a comprehensive corpus-based study using the *British National Corpus* (BNC), the *Wellington Corpus of Spoken English* and several other written and spoken corpora. He established that—if 98% is considered the necessary coverage for comprehension—8000–9000 or 6000–7000 families would be necessary in order to understand unsimplified reading material or spoken discourse, respectively, and this is assuming that proper nouns are easily understood. In subsequent studies, Laufer and Ravenhorst-Kalovski (2010) and Schmitt et al. (2011) investigated over one thousand participants with various native languages (L1s) and corroborated Nation's (2006) findings.

Indeed, mastering 9000 word families is far more manageable than the 20,000 families known receptively to educated native speakers, as mentioned above. Nevertheless, this is still a fairly large number, and it fails to consider at least two crucial

aspects that are central to vocabulary knowledge. First, this number takes into consideration only words as individual units, hence ignoring the existence of the *multiword items* that permeate discourse. Multiword items (e.g., phrasal verbs, fixed phrases, and idioms; see Moon, 1997; Nation & Meara, 2010 for more examples and categories) are also called, inter alia, *formulaic phrases* or *chunks* (Cameron, 2001), *phrasal lexical items* (Schmitt, 2008), and *lexical bundles* (Biber et al., 2004; Leńko-Szymańska, 2014). Some of them occur so frequently in English that they may be considered high-frequency items (Schmitt, 2008; Schmitt & McCarthy, 1997; Webb et al., 2013).

Second, irrespective of being single or multiword units, each of these lexical items must be learnt in *depth.* That is, learners must become acquainted with the various aspects of a word's form, meaning, and use, both receptively and *productively* (i.e., the ability to use the words accurately in written and oral production). Nation's (2001, 2013) specification of the aspects involved in knowing a word seems to be the most comprehensive to date (González-Fernández & Schmitt, 2020; Schmitt, 2010, 2014) and has therefore been included here for reference (see Table 1.1).

To summarize, competence in L2 English comprehension and production can only be reached with enormous vocabulary knowledge, not only in terms of number of single words or multiword items, but also with regard to depth of learning. Put differently, acquiring words and deepening understanding of known vocabulary is as crucial to second language development as it is challenging. This is also true in academic contexts, to which this chapter turns now.

1.3 Vocabulary Knowledge in Higher Education and the Importance of Academic Vocabulary

Acquiring words is as important at university as it is in other contexts, perhaps even more so. In higher education, vocabulary knowledge has been shown to enhance L2 proficiency, especially regarding reading comprehension (Laufer & Ravenhorst-Kalovski, 2010; Li & Pemberton, 1994; Milton et al., 2010; Paribakht & Webb, 2016) and written production (Evans & Morrison, 2011; Hyland, 1997; Laufer & Ravenhorst-Kalovski, 2010; Milton et al., 2010). In fact, recent research suggests that university students need to know even more words than general English learners. Webb and Paribakht (2015) have demonstrated that knowledge of 14,000 word families may be necessary to attain sufficient comprehension of typical academic readings. This is far higher than the previous estimates of up to 9000 families mentioned above (Nation, 2006). Furthermore, similarly to Nation's estimates, this figure disregards the need to know words in depth and the ubiquity of multiword items in academic discourse (Biber et al., 2004; Byrd & Coxhead, 2010; Hyland, 2008, 2012; Simpson-Vlach & Ellis, 2010). Therefore, rather than attempting to acquire a large number of general English words, teachers and students may focus on academic vocabulary.

Table 1.1 What is involved in knowing a word (adapted from Nation, 2001, p. 27)

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 recognisable in this word?
		P	What word parts are needed to express the 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 of use (register, frequency…)	R	Where, when, and how often would we expect to meet this word?
		P	Where, when, and how often can we use this word?

Note R = receptive knowledge, P = productive knowledge

Academic words may be defined as words which are "reasonably frequent in a wide range of academic genres [while] relatively uncommon in other kinds of texts" (Hyland & Tse, 2007, p. 235). They contrast with *technical words* in that they are not specific to any field; they also differ from *low-frequency words* since these are items which are less frequent than the 2000 most common word families and are not specific to academic contexts (Nation & Waring, 1997; Wang et al., 2008). Coxhead's (2000) *Academic Word List* (AWL) is perhaps the most frequently adopted list of academic words to date (Paribakht & Webb, 2016) and will thus be introduced below in some detail.

The AWL was based on a corpus of written texts of 3.5 million words divided into the disciplines of arts, law, commerce, and science (see School of Linguistics and Applied Language Studies, 2020, for access to the AWL). As such, the AWL is considered suitable for a wide range of academic fields. The list contains 570 families, or about 3100 individual word forms (types). The AWL is divided into 10

sublists, and each sublist contains 60 families, except for sublist 10, which contains 30. The words become less frequent with each sublist, with sublist 10 containing the least frequent words. These are some examples of academic word families from the AWL (one word per sublist; from sublist 1 to 10): "analyze", "distinct", "ensure", "implement", "alter", "enhance", "differentiate", "complement", "anticipate", and "conceive".

The AWL is not without flaws, however. For one thing, it has been criticized because of its generalist character. That is, the 570 families of this list provide significantly different coverage across disciplines, including semantic and collocational differences (Cobb & Horst, 2004; Hyland & Tse, 2007; Martínez et al., 2009). Also, even though the AWL was supposed to only contain items outside the highest frequency range, its families do overlap with the 2000 most frequent words from more modern corpora, such as the BNC (Gardner & Davies, 2013; Nagy & Townsend, 2012). Consequently, field-specific lists have been compiled to counteract some of these problems (see Table 1.2 for some of these lists).

These lists are believed to be more suitable to target disciplines (e.g., collocationally and semantically) and to provide enhanced coverage, and few of them have indeed accomplished this (e.g., Lei & Liu, 2016; Liu & Han, 2015). Still, applied linguistics lists have been shown to increase the coverage of academic texts relative to the AWL by only 0.52% (Khani & Tazik, 2013); that is, learning these lists rather than the AWL would increase learners' comprehension by a rather narrow margin. Also, these lists overlap considerably with the AWL: 74.12% (Khani & Tazik, 2013) and 83.33% (Vongpumivitch et al., 2009). This means that any advantage of applied linguistics lists over the AWL would be minimal and unlikely to outweigh the yet-unknown flaws of these lists (Paribakht & Webb, 2016). Therefore, the AWL may still be considered the most suitable list to be used with many learners, especially with students from fields related to applied linguistics, such as the participants investigated in this book.

It stands to reason that university students should focus on academic words instead of low-frequency or technical vocabulary. The rationale behind this is that knowledge of academic words will allow learners to understand a much higher percentage of academic texts more quickly. For example, studies have shown that learning the 570 word families of Coxhead's (2000) AWL provides an additional coverage of at least 10% of academic texts (Paribakht & Webb, 2016). In applied linguistics, the AWL word families provide a textual coverage of between 11.17% and 13.11% (Chung & Nation, 2003; Cobb & Horst, 2004; Khani & Tazik, 2013; Vongpumivitch et al., 2009; see Table 1.2). Consequently, learning these words is more efficient for university students than, for example, acquiring the third most common 1000 families in the *British National Corpus* (BNC), which typically increases textual coverage by only 4.36% (Nation & Webb, 2011).

Another advantage of learning academic words may hinge upon their etymology. Most academic words are derived from Greek or Latin (Nagy & Townsend, 2012), with 91% of the AWL families being of Greco-Latin origin (Schmitt et al., 2001). This may make academic words relatively easy to acquire for learners who are proficient in languages that borrowed heavily from Greek or Latin (Otwinowska, 2015). This

Table 1.2 Studies utilizing corpora to investigate AWL coverage in specific fields and create field-specific lists (see Silva, 2018)

Study	Academic field	Created their own list?	AWL coverage in the field	Own list coverage in the field
Mudraya (2006)	Engineering	Yes	NA	NA
Chen and Ge (2007)	Medicine	No	10.07%	NA
Wang et al. (2008)	Medicine	Yes	NA	12.24%
Lei and Liu (2016)	Medicine	Yes	NA	20%
Hyland and Tse (2007)	Several	No	Range: 6.2–16% Average: 10.5%	NA
Martínez et al. (2009)	Agriculture	No	9.06%	NA
Li and Qian (2010)	Finance	No	10.46%	NA
Coxhead and Hirsh (2007)	Science	Yes	8.96%	3.79%
Liu and Han (2015)	Environmental sciences	Yes	12.82%	15.43%
Chung and Nation (2003)	Applied linguistics (AL)/Anatomy	No	AL: 13.1% Anatomy: NA	NA
Cobb and Horst (2004)	Linguistics + several others	No	Linguistics: 12.60% Average: 11.60%	NA
Vongpumivitch et al. (2009)	AL	No	11.17%	NA
Khani and Tazik (2013)	AL	Yes	11.96%	12.48%
Mean			10.97%	12.79%

Note NA = not available

is because many of these words will be *cognates*, i.e., "words that share form and meaning in two languages" (Lemhöfer et al., 2008, p. 12; see more details below). Even though Polish and English are typologically distant (a Slavonic vs. a Germanic language), both have borrowed extensively from Latin and Latin-based languages, such as French and Italian (Otwinowska, 2015). It is perhaps unsurprising, then, that 389 AWL types (12.54% of the list) are Polish-English cognates (see Appendix A for the full list) and hence likely easier to learn for Polish university students (see below). Given the significant number of Polish-English cognates present in the AWL, a more detailed discussion of the role of cognates in lexical learning is in order.

1.4 The Role of Cognateness in Vocabulary Learning

Defining cognateness is not as simple as implied above. Some researchers (e.g., Laufer & Levitzky-Aviad, 2018; Rogers et al., 2015) differentiate between "cognates" and "loanwords". The former refers to words shared between genetically related languages (e.g., Portuguese and Spanish); the latter is used to describe borrowings from unrelated languages. For instance, the English "precise" and the Polish "precyzyjny" are semantically and formally similar, but different from the Czech "přesný". Since the form "precyzyjny" is not strikingly similar to the word in Czech, a language genetically related to Polish, some researchers would say that "precise" and "precyzyjny" are loanwords, not cognates. By contrast, the English noun "mill" and the Polish "młyn" may be considered cognates since the word in Czech is "mlýn". Despite this difference, I will follow the more general definition by Lemhöfer et al. (2008), whereby cognates are words whose form and meaning are similar between languages, irrespective of their genealogy. One reason for this is that, as pointed out by Jarvis (2009), L2 learners are rarely able to differentiate between genealogical cognates and loanwords; another reason is that both types of words may well affect language learning similarly, rendering such distinction unnecessary in practice.

There is strong evidence in the literature to suggest that *cognates have a learning and processing advantage* over noncognates. In a highly relevant cross-sectional study with 120 Polish learners of English, Otwinowska and Szewczyk (2019) compared participants' knowledge of 35 Polish-English cognates and 35 Polish-English noncognates and demonstrated that cognates were 2.5 times more likely to be correctly translated than noncognates, even after controlling for cognate guessing. Helms-Park and Perhan (2016) found a similar advantage of cognates over noncognates, but this time exploring languages that do not share the same script (i.e., Ukrainian and English). In a psycholinguistic laboratory experiment, Lotto and De Groot (1998) investigated 56 Dutch university participants learning L2 Italian and used a paired associate task to compare learning and speed of retrieval of cognates and noncognates. In the paired associate task, participants were exposed either to the Dutch word and the corresponding Italian word (the word-learning condition) or to a picture and the corresponding Italian word (picture-learning condition). The results showed that cognates were learnt more and retrieved faster than noncognates in both conditions. Other laboratory experiments that used the paired-associate task and found similar results include De Groot and Keijzer (2000) with adults, as well as Tonzar et al. (2009) and Comesaña et al. (2012) with children (see also Puimège & Peters, 2019 for a very recent review on variables that affect lexical learning, including the advantage of cognates over noncognates). This cognate advantage may be because of the fact that when acquiring a cognate, learners do not need to map a novel L2 word form—both orthographic and phonological—onto the existing L1 concept; rather, learners need only validate the form-meaning connection between words in both languages (Ecke, 2015; Ringbom, 2007). For a similar reason, cognates appear to be translated faster and more accurately than noncognates (e.g., Jacobs et al., 2016; see De Groot, 2011 for an overview).

Psycholinguistic experiments have also shown that cognates are recognized faster than noncognates. This phenomenon is called the *cognate facilitation effect*, and persuasive evidence thereof abounds in the literature. Most of these studies have utilized lexical decision tasks (LDT) with adults. In these types of tasks, learners are presented with a word on a screen and must decide as quickly as possible if it is a word (e.g., cognate or noncognate) or a nonword (i.e., an artificially created word). Studies conducted with adult advanced L1-Dutch and L2-English bilinguals (e.g., Dijkstra et al., 2010; Lemhöfer & Dijkstra, 2004; Mulder et al., 2015) as well advanced L1-Spanish and L2-Catalan bilinguals (e.g., Comesaña et al., 2015) have consistently shown that cognates are recognized faster than cognates. Brenders et al. (2011) have found similar results with Dutch children who are intermediate learners of English, thus showing that less proficient and less experienced learners may also benefit from cognate facilitation effects (see below).

Other studies have reinforced these results with serial visual representation tasks. Here, participants saw sentences on the screen, one word at a time, with the critical words (i.e., keywords) shown in red. The participants were instructed to say the critical word out loud in a microphone as quickly and as accurately as possible. Schwartz and Kroll (2006) found shorter response times for cognates than for noncognates when investigating 23 English-speaking participants of intermediate or advanced proficiency in Spanish. Similar results were reported by Duyck et al. (2007) with advanced Dutch learners of L2 English. These experiments were conducted with isolated words, and the cognate facilitation effect, as defined in the psycholinguistic literature, refers only to the faster recognition of cognates mostly when presented in isolation. However, this might mean that bilinguals could recognize cognates faster than noncognates in reading tasks, and this faster recognition may lead to enhanced learning (see Sect. 2.2).

The proficiency level of participants may also affect the advantage afforded by cognates. Usually, higher-proficiency students, better acquainted with the formal similarities between languages, may benefit the most from cognateness (Nagy et al., 1993; Otwinowska & Szewczyk, 2019). One reason is that the formal overlap between cognates varies considerably (Jarvis, 2009; Ringbom, 2007), with more divergent pronunciation and/or orthography blurring the similarity between L1 and L2 words to a larger extent (Comesaña et al., 2015; De Groot, 2011; Dijkstra et al., 2010; Duyck et al., 2007; Mulder et al., 2015; Otwinowska & Szewczyk, 2019). As a result, learners at higher proficiency levels may be better able than lower-proficiency students to notice and learn cognates with less similarity (Dressler et al., 2011; Nagy et al., 1993) because, for instance, they are better acquainted with word morphology and grapheme-to-phoneme mappings (Otwinowska & Szewczyk, 2019). Moreover, the degree of semantic similarity between cognates may also affect learning. Cognate pairs may be fully equivalent semantically, or the meanings may only partially overlap; also, in the case of polysemous words, not all available meanings may exist (and/or be similar) in both languages (see Ringbom, 2007 or Otwinowska, 2015 for a more comprehensive discussion). Again, more proficient learners are better equipped to understand the nuances in meaning, thereby being more liable to benefit from cognateness than less advanced learners.

In short, in the case of Polish learners of English as an L2, over 12% of the AWL items are Polish-English cognates (see Appendix A). Based on the existing evidence presented above, such a number of cognates may facilitate the acquisition of academic vocabulary for Polish learners. On the other hand, research has consistently shown that acquiring L2 academic words may be problematic, more so than acquiring low-frequency and technical vocabulary. The next section will focus on these findings.

1.5 The Difficulty in Learning Academic Vocabulary

Research shows that acquiring L2 academic vocabulary may be challenging for university students, particularly when teachers fail to draw learners' attention more explicitly to these types of words. Academic words are mostly abstract, low in imagery, and morphologically complex (Corson, 1997; Lubliner & Hiebert, 2011; Vidal, 2011), and often occur rather infrequently outside academic contexts to be learnt incidentally via input (Hyland & Tse, 2007; Lei & Liu, 2016; Vidal, 2011). A series of studies conducted at an English-medium university in Australia with non-native English speakers of various nationalities showed no significant gains in academic vocabulary knowledge after six months (Storch, 2009), one year (Knoch et al., 2014), and even three years (Knoch et al., 2015) of instruction. In all studies, the participants did not take part in any course focused on developing their English proficiency, such as English for Academic Purposes (EAP) programs. Had they had access to these types of courses, they would likely have had more opportunities to practice academic words (e.g., through academic writing) and would have received more explicit instruction on academic vocabulary. In fact, when interviewed, participants reported producing precious little writing during those three years; also, they complained that when given the opportunity to write in English, they received little to no corrective feedback from their teachers. In other words, considering the little opportunity to practice and limited explicit academic vocabulary instruction, it appears that these students were expected to learn incidentally solely through exposure to input during their years at university. Nevertheless, even living in Australia and attending and English-medium university, this exposure was clearly insufficient to improve their knowledge of academic words.

In another longitudinal study, this time in China, Zhang and Lu (2014) used Schmitt et al.'s (2001) *Vocabulary Levels Test* (VLT) to measure the knowledge of 30 academic lexical items. At the beginning of the study, the Chinese participants were able to recognize the meaning of 69.95% of the 30 words; after 2 years at an English-medium university, their scores improved by only 16.95%, to 86.9%. This is very little improvement considering that the VLT is a *meaning-recognition* test—that is, test takers are only required to match words to the meanings provided. Therefore, the VLT measures solely the early stages of lexical acquisition, when the form-meaning link between words starts to be established (see Laufer et al., 2004; Laufer & Goldstein, 2004; Schmitt, 2010). Put differently, recognizing the meaning of a word when definitions are provided, such as in the VLT, does not mean learners

are able understand the words in texts (i.e., *meaning recall*), much less produce them (González-Fernández & Schmitt, 2020; Laufer & Paribakht, 1998; Mondria & Wiersma, 2004; Webb, 2005).

Given the above, there is convincing evidence to suggest that long-term exposure to the L2 in academic contexts is insufficient to promote significant learning of academic words, even when participants live in English-speaking countries. When learning occurred, it was very limited (Zhang & Lu, 2014; see also Vidal, 2011) and failed to reach significance when productive knowledge was measured (Knoch et al., 2014, 2015; Storch, 2009 see also Lin & Morrison, 2010). By comparison, Storch and Tapper (2009) followed postgraduate learners taking part in an EAP course with an explicit focus on academic vocabulary and found significant learning of AWL items after only 10 weeks. Additionally, Helms-Park and Perhan (2016) subjected Ukrainian participants to the explicit instruction of AWL items and demonstrated that these participants outperformed the participants in the reading treatment—without explicit AWL instruction—in both receptive and productive test measurements. These results show why researchers often recommend a more explicit focus on the learning of academic words (e.g., Corson, 1997; Gardner & Davies, 2013; Kuehn, 1996; Nagy & Townsend, 2012; Vongpumivitch et al., 2009).

One reason for the dearth of learning of academic vocabulary despite extended exposure may lie in the role these words play in discourse. For instance, Vidal (2011) exposed Spanish students of English to written and oral academic English input and demonstrated that academic vocabulary was considerably more difficult to acquire than technical and low-frequency words. She went on to explain that academic vocabulary may be less salient in discourse than low-frequency and technical vocabulary. Due to this, the researcher asserted that learners attended to academic words infrequently even when reading, when they "had the opportunity to reread parts of the text and attend to language form" (Vidal, 2011, p. 248). Additionally, as explained by others, technical words are discipline-specific and often central to a topic, thus frequently being highlighted and explained in the classroom (Hancioğlu et al., 2008; Li & Pemberton, 1994; Strevens, 1973). Academic words, conversely, adopt a more supportive role in discourse, hence lacking salience (Vidal, 2011), and are thus usually ignored by students and subject specialists, who assume these words are known to learners (Farrell, 1990).

This book has so far argued that learning L2 academic vocabulary is crucial to academic success, but that this learning may be affected by some variables. On the one hand, cognateness among academic words may facilitate the learning of the AWL items; on the other, there is strong evidence that the acquisition of academic vocabulary is highly problematic, even after years of exposure in English-speaking contexts, and that therefore learners often lack sufficient knowledge of these words. This has led researchers to suggest that academic vocabulary may be a major barrier for university students, both natives and non-natives (Baumann & Graves, 2010; Evans & Green, 2007), particularly in reading and writing (Chen & Ge, 2007; Kuehn, 1996; Shaw, 1991).

At least two conclusions may be drawn from the above. First, it is necessary to ensure that university students, Polish students for the purpose of this book, possess

sufficient knowledge of academic vocabulary. Second, if learners fail to provide evidence of such knowledge, it is crucial to investigate ways to facilitate the learning of academic words. These are exactly the two overarching aims of this book: (1) to find a reliable way to measure academic vocabulary knowledge of Polish university students and (2) to compare the effectiveness of different types of writing to the learning of academic words. I will now draw attention to research that is germane to the former.

1.6 Cognate Inflation Effects in Vocabulary Tests

As discussed above, cognates may be easier to learn, retain, translate, and recognize than noncognates (see Sect. 1.3). While this could provide an advantage for learners, it may also give rise to problems when assessing lexical knowledge. This is because cognateness may benefit speakers of some languages more than others. Since English contains a large number of Latin and Greek words (Otwinowska, 2015; Petrescu et al., 2017), and most of the academic words are of Greco-Latin origin, cognateness is more advantageous for learners whose language(s) borrowed from Latin or Greek than for students who are proficient in languages that contain few or no such words. One drawback of this cognate advantage, especially for researchers and educational institutions, is that scores in standardized international vocabulary tests—i.e., tests designed for all students, irrespective of linguistic background and proficiency level—may be artificially inflated, making students appear more proficient than they really are.

Research conducted by Petrescu et al. (2017) with Vietnamese and Romanian students of similar English proficiency (i.e., upper-intermediate level) has demonstrated such cognate inflation. The researchers adopted the VLT (Schmitt et al., 2001) and utilized the academic level, the 10,000-word level (already considered an advanced level), and a newly designed level with words with frequencies between 12,000 and 20,000. Thus, only the knowledge of academic words and low-to-very-low-frequency words (i.e., very advanced vocabulary) was assessed. As expected, Romanian participants, whose language is rich in Latinate and Greek vocabulary, benefited more than Vietnamese learners. Romanians had significantly higher scores in all levels of the test, including the level with academic words. Moreover, the cognate advantage was more prominent at higher levels, suggesting a more pronounced inflation among very low-frequency words. If this test had been used for university placement purposes, for example, the Romanian learners would have been erroneously deemed more proficient than their Vietnamese colleagues. Perhaps even worse, they might have been inadvertently considered highly proficient in English. This is because, owing to the presence of cognates, they scored relatively high (i.e., 53.10%) even in the 12–20,000 level (with words such as “verdure”, “macerate”, “abjure”, and “asperity”—all English-Romanian cognates; see Petrescu et al., 2017, p. 24).

Evidence of the cognate inflation effect has also been found in standardized tests measuring general vocabulary proficiency. Allen (2018) also adopted the VLT, this time with Japanese learners of English, and lent support to Petrescu et al.'s (2017) results. Allen (2018) measured participants' vocabulary knowledge in frequency bands 2000, 3000, 5000, and 10,000, as well the academic vocabulary level. Signs of cognate inflation were found overall, but the researcher did not report on the results for each level separately; therefore, it remains unclear whether cognate inflation was detected among academic words. Still, one finding stands out in this study: cognates were recognized more accurately—and hence the cognate inflation was more pronounced—when they were highly frequent in participants' L1 (i.e., Japanese), regardless of their frequency in English. Put differently, even rather infrequent and hence advanced English words were more readily recognized provided that they were common in Japanese. Again, if the VLT had been used to assess participants' proficiency, the accurate recognition of advanced words may have made participants look more proficient than they are. In another study also with Japanese learners, Jordan (2012) corroborated Allen's (2018) findings, but this time using a meaning recall (L2-L1 translation) test. Here, and in line with Petrescu et al.'s (2017) results, the researcher found evidence of higher cognate inflation among lower-frequency words.

Adding to the existing body of evidence, Elgort (2013) obtained similar results with Russian learners of English. She utilized the *Vocabulary Size Test* (VST; Nation & Beglar, 2007), a meaning-recognition multiple-choice test. This test comprises 14 frequency bands (i.e., levels). Each band contains 10 lexical items, totaling 140 items, and the score is multiplied by 100 (max. = 14,000), representing learners' receptive vocabulary size. Corroborating the findings reported above, Elgort (2013) detected higher cognate inflation among the lower-frequency bands and estimated that the scores of her Russian participants had been inflated by at least 1000 words. More recently, Laufer and Mclean (2016) found similar results, although with speakers of Hebrew and Japanese. In this study, the researchers utilized only the first 8 bands of the VST, adapting them to measure three types of the form-meaning knowledge: active-recall (L1-L2 translation), passive-recall (L2-L1 translation), and active-recognition (choose the correct spelling of the word) tests (see Laufer et al., 2004). Once again, the presence of cognates inflated the results.

Thus far, except for Petrescu et al.'s (2017) study with Vietnamese and Romanian learners, all research reviewed here utilized participants with L1s whose scripts differed from the English alphabet (i.e., Japanese, Hebrew, and Russian). Still, cognate inflation was reported even though participants could only rely on phonological and semantic, but not orthographic similarity, which may have hampered cognate recognition (Lemhöfer et al., 2008; Nagy et al., 1993; see Berthele, 2011 for how phonological and orthographic similarities facilitate cognate recognition differently). That is, cognate inflation effects may be significantly higher when learners' L1 shares the English script (such as Polish).

Consequently, a growing body of existing evidence suggests that the presence of cognates distorts the results of standardized English tests. Since these tests are designed for all learners, cognateness may confer an unfair advantage or disadvantage

to some students, and researchers cannot agree on a viable solution to this problem. Gyllstad et al. (2015) have suggested keeping cognates in tests since these cross-linguistic formal and semantic similarities should be borne in mind when assessing learners. However, this does not solve the problem of cognate inflation. Petrescu et al. (2017) have argued in favor of scoring cognates and noncognates separately to give more flexibility to teachers, researchers, and institutions, so they can better decide when to include cognates. Still, the researchers are unclear as to when incorporating cognates would be acceptable; also, excluding cognates may underestimate lexical size estimates. Additionally, cognateness is inextricably intertwined to learners' L1, making their identification an insurmountable task when learners from various L1 are involved. Finally, Allen (2018), Elgort (2013), and Laufer and McLean (2016) have recommended that the proportion of cognates in a test be similar to the proportion in learners' L1. While this appears reasonable, it is impractical when students originate from diverse linguistic backgrounds, not to mention that verifying such proportion may be difficult, if not impossible, in many contexts.

To illustrate, researchers are still unaware of the number of cognates that exist between Polish and English. Both languages, although typologically distant, have borrowed heavily from other languages, particularly French, Italian, and Latin (Otwinowska, 2015). Polish has also borrowed from English for the last 150 years, and some words have changed considerably, making their identification onerous. The most comprehensive Polish dictionaries contain approximately 140,000 words. Overall, 13% may be considered borrowings, 9% stemming from Latin or English (Otwinowska, 2015). Still, the proportion of highly similar cognates is likely much lower, not to mention that some words are only partial cognates or false cognates (formally similar but semantically different).

It appears, then, that it is a formidable task to arrive at the proportion of cognates for each language for which a vocabulary test is needed, as suggested by some researchers (Allen, 2018; Elgort, 2013; Laufer & McLean, 2016). What is necessary instead is a novel, efficient, and accurate way to estimate learners' receptive L2 vocabulary size, thus obtaining a reliable measurement of their English proficiency level (Hu & Nation, 2000; Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Milton et al., 2010; Paribakht & Webb, 2016; Stæhr, 2008). Such a measure, or test, may then be used by teachers, and researchers, or for placement at university.

1.7 Conclusion

This chapter has explained that vocabulary knowledge is as important when learning a language as it is challenging. It has also demonstrated that, for university students, learning academic words is paramount, but may necessitate more direct instructional interventions. Additionally, This chapter has highlighted the role of cognates among academic words and has showed that while cognateness may facilitate learning, they may also blur test results, making it difficult to reliably utilize vocabulary tests to assess language proficiency or for placement purposes. Developing a reliable test for

university placement is the first goal of this book, which is investigated in Study 1, reported in Chap. 5. Once learners have been properly placed, it will be necessary to ensure that those learners in need to improve their knowledge are able to acquire academic words effectively. To this aim, Chap. 2 explores research underpinning lexical learning.

References

Allen, D. (2018). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism*, 1–16. https://doi.org/10.1177/1367006918781063

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279.

Baumann, J. F., & Graves, M. F. (2010). What is academic vocabulary? *Journal of Adolescent & Adult Literacy, 54*(1), 4–12.

Berthele, R. (2011). The influence of code-mixing and speaker information on perception and assessment of foreign language proficiency: An experimental study. *International Journal of Bilingualism, 16*(4), 453–466.

Biber, D., Conrad, S., & Cortes, S. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Brenders, P., Van Hell, J. G., & Dijkstra, T. (2011). Word recognition in child L2 learners: Evidence from cognates and false friends. *Journal of Experimental Child Psychology, 109*, 383–396.

Byrd, P., & Coxhead, A. (2010). *On the other hand: Lexical bundles in academic writing and in the teaching of EAP* (Vol. 5, pp. 31–64). University of Sydney Papers in TESOL.

Cameron, L. (2001). *Teaching languages to young learners*. Cambridge University Press.

Chang, A. C.-S. (2019). Effects of narrow reading and listening on L2 vocabulary learning. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263119000032

Chen, Q., & Ge, G.-C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes, 26*, 502–514.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116.

Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in second language: Selection, acquisition, and testing* (pp. 15–38). John Benjamins.

Comesaña, M., Soares, A. P., Ferré, P., Romero, J., Guasch, M., & García-Chico, T. (2015). Facilitative effect of cognate words vanishes when reducing the orthographic overlap: The role of stimuli list composition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 614–635.

Comesaña, M., Soares, A. P., Sánchez-Casas, R., & Lima, C. (2012). Lexical and semantic representations in the acquisition of L2 cognate and non-cognate words: Evidence from two learning methods in children. *British Journal of Psychology, 103*, 378–392.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliquée, 12*(2), 65–78.

De Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction*. Psychology Press.

De Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: The role of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1), 1–56.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-linguistic similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*, 284–301.

Dressler, C., Carlo, M. S., Snow, C. E., August, D., & White, C. E. (2011). Spanish-speaking students' use of cognate knowledge to infer the meaning of English words. *Bilingualism: Language and Cognition, 14*(2), 243–255. https://doi.org/10.1017/S1366728910000519

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 663–679.

Ecke, P. (2015). Parasitic vocabulary acquisition, cross–linguistic influence, and lexical retrieval in multilinguals. *Bilingualism: Language and Cognition, 18*(2), 145–162. https://doi.org/10.1017/S1366728913000722

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing, 30*(2), 253–272.

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

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes, 6*, 3–17.

Evans, S., & Morrison, B. (2011). Meeting the challenges of English-medium higher education: The first-year experience in Hong Kong. *English for Specific Purposes, 30*, 198–208.

Farrell, P. (1990). *A lexical analysis of the English of electronics and a study of semi-technical vocabulary* (CLCS Occasional Paper No. 25). Trinity College (ERIC Document Reproduction Service No. ED332551). Retrieved from http://files.eric.ed.gov/fulltext/ED332551.pdf. Accessed November 07, 2017.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327.

Godfroid, A., & Hui, B. (2020). Five common pitfalls in eye-tracking research. *Second Language Research*, 1–29. https://doi.org/10.1177/0267658320921218

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481–505.

Gyllstad, H., Vilkaite, 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. https://doi.org/10.1075/itl.166.2.04gyl

Hancioğlu, S., Neufeld, S., & Eldridge, J. (2008). Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes, 27*, 459–479.

Helms-Park, R., & Perhan, Z. (2016). The role of explicit instruction in cross-script cognate recognition: The case of Ukrainian-speaking EAP learners. *Journal of English for Academic Purposes, 21*, 17–33.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 23*(1), 403–430.

Hyland, K. (1997). Is EAP necessary? A survey of Hong Kong undergraduates. *Asian Journal of English Language Teaching, 7*, 77–99.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150–169.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly, 41*(2), 235–253.

Jacobs, A., Fricke, M., & Kroll, J. F. (2016). Cross-language activation begins during speech planning and extends into second language speech. *Language Learning, 66*, 324–353. https://doi.org/10.1111/lang.12148

Jarvis, S. (2009). Lexical transfer. In A. Pavlenko (Ed.), *The bilingual mental lexicon: Interdisciplinary approaches* (pp. 99–124). Multilingual Matters.

Jordan, E. (2012). Cognates in vocabulary size testing—A distorting influence? *Language Testing in Asia, 2*(3), 5–17.
Khani, R., & Tazik, K. (2013). Towards the development of an academic word list for applied linguistics research articles. *RELC Journal, 44*(2), 209–232.
Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing, 21*, 1–17.
Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university. *Journal of Second Language Writing, 28*, 39–52.
Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Pergamon.
Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence by the input hypothesis. *The Modern Language Journal, 73*(4), 440–464.
Kuehn, P. (1996). *Assessment of academic literacy skills: Preparing minority and limited English proficient (LEP) students for post-secondary education.* California State University (ERIC Document Reproduction Service No. ED415498). Retrieved from file:///C:/Users/breno/OneDrive/University%20of%20Nottingham/Courses/Issues%20in%20Teaching%20EAP/Extra/Studies%20with%20Academic%20vocabulary/Academic%20vocabulary%20difficulty/Kuehn%201996.pdf. Accessed November 11, 2017.
Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Multilingual Matters.
Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399–436.
Laufer, B., & Levitzky-Aviad, T. (2018). Loanword proportion in vocabulary size tests: Does it make a difference? *International Journal of Applied Linguistics, 169*(1), 95–114.
Laufer, B., & Mclean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly, 13*(3), 202–217.
Laufer, B., & Nation, I. S. P. (2011). Vocabulary. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 163–176). Routledge.
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., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. Retrieved in March 2018 from http://files.eric.ed.gov/fulltext/EJ887873.pdf
Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*, 202–226. https://doi.org/10.1191/0265532204lt277oa
Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42–53.
Lemhöfer, K., & Dijkstra, T. (2004). Recognizing cognates and interlingual monographs: Effects of code similarity in language-specific and generalized lexical decision. *Memory & Cognition, 32*(4), 533–550.
Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 12–31.
Leńko-Szymańska, A. (2014). The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics, 19*(2), 225–251. https://doi.org/10.1075/ijcl.19.2.04len
Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Cengage Learning, Inc.
Li, E. S.-L., & Pemberton, R. (1994). An investigation of students' knowledge of academic and subtechnical vocabulary. *Proceedings of the Joint Seminar on Corpus Linguistics and Lexicology*

(pp. 183–196). Hong Kong University of Science and Technology. Retrieved in November 2017 from: http://repository.ust.hk/ir/bitstream/1783.1-1089/2/entertext05.pdf
Li, Y., & Qian, D. D. (2010). Profiling the academic word list (AWL) in a financial corpus. *System, 38*, 402–411. https://doi.org/10.1016/j.system.2010.06.015
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*, 255–266.
Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes, 39*, 1–11.
Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48*(1), 31–69.
Lubliner, S., & Hiebert, E. H. (2011). An analysis of English-Spanish cognates as a source of general academic language. *Bilingual Research Journal, 34*(1), 76–93.
Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60.
Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based research. *English for Specific Purposes, 28*, 183–198.
McCarthy, M. (1999). What constitutes a basic vocabulary for spoken communication? *Studies in English Language and Literature, 1*, 233–249.
Meara, P. (1995). The importance of an early emphasis on L2 vocabulary. *The Language Teacher, 19*(2), 8–11.
Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. D. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–97). Multilingual Matters.
Mondria, J., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 79–100). John Benjamins Publishing.
Moon, R. (1997). Vocabulary connections: Multi-word items in English. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 40–63). Cambridge University Press.
Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*, 235–256.
Mulder, K., Dijkstra, T., & Baayen, R. H. (2015). Cross-language activation of morphological relatives in cognates: The role of orthographic overlap and task-related processing. *Frontiers in Human Neuroscience, 9*(16), 1–18.
Nagy, W. E., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91–108.
Nagy, W. E., García, G. E., Durgunoğlu, A. Y., & Hancin-Bhatt, B. (1993). Spanish-English bilingual students' use of cognates in English reading. *Journal of Reading Behaviour, 25*(3), 241–259.
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–81.
Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.
Nation, I. S. P., & Meara, P. (2010). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed., pp. 34–52). Hodder Education.
Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6–19). Cambridge University Press.
Nation, I. S. P., & Webb, S. (2011). *Researching and analysing vocabulary*. Heinle Cengage Learning.

O'Dell, F. (1997). Incorporating vocabulary into the syllabus. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 258–278). Cambridge University Press.

Otwinowska, A. (2015). *Cognate vocabulary in language acquisition and use: Attitudes, awareness, activation.* Multilingual Matters.

Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism, 22*(8), 974–991.

Paribakht, T. S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes, 21*, 121–132.

Pellicer-Sánchez, A. (2019). Examining second language vocabulary growth: Replications of Schmitt (1998) and Webb & Chang (2012). *Language Teaching, 52*(4), 512–523.

Petrescu, M. C., Helms-Park, R., & Dronjic, V. (2017). The impact of frequency and register on cognate facilitation: Comparing Romanian and Vietnamese speakers on the vocabulary levels test. *English for Specific Purposes, 47*, 15–25.

Puimège, E., & Peters, E. (2019). Learner's English vocabulary knowledge prior to formal instruction: The role of word-related and learner-related variables. *Language Learning, 69*(4), 943–977.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146–161.

Ringbom, H. (2007). *The importance of cross-linguistic similarity in foreign language learning: Comprehension, learning and production.* Multilingual Matters.

Rogers, J., Webb, S., & Nakata, T. (2015). Do the cognacy characteristics of loanwords make them more easily learned than noncognates? *Language Teaching Research, 19*, 9–27.

Schmitt, N. (2007). Current trends in vocabulary learning and teaching. In J. Cummins & C. Davison (Eds.), *The international handbook of English language teaching* (Vol. 2). Springer.

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows? *Language Learning, 64*(4), 913–951.

Schmitt, N., & McCarthy, M. J. (Eds.) (1997). *Vocabulary: Description, acquisition and pedagogy.* Cambridge University Press.

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.

Schwartz, A. I., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory and Language, 55*, 197–212.

Shaw, P. (1991). Science research students' composing processes. *English for Specific Purposes, 10*, 189–206.

Silva, B. (2018). *Lack of English academic vocabulary at a polish university: Outlining negative consequences and suggesting potential solutions* [Unpublished manuscript]. School of Education, University of Nottingham.

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

Sökmen, A. J. (1997). Current trends in teaching second language vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 237–257). Cambridge University Press.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal, 36*(2), 139–152.

Stæhr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition, 31*, 577–607.

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing, 18*, 103–118.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes, 8*, 207–223.

Strevens, P. (1973). Technical, technological, and scientific English (TTSE). *ELT Journal, 27*(3), 223–234.

Tonzar, C., Lotto, A., & Job, R. (2009). L2 vocabulary acquisition in children: Effects of learning method and cognate status. *Language Learning, 59*(3), 623–646.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics, 22*(2), 217–234.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258.

Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (2009). Frequency analysis of the words in the academic word list (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes, 28*, 33–41.

Wang, J., Liang, S.-I., & Ge, G.-C. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*, 442–458.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*, 33–52.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test. *English for Specific Purposes, 38*, 34–43.

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

Zhang, X., & Lu, X. (2014). A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics, 35*(3), 283–304.

Chapter 2
Incidental Lexical Learning and the Involvement Load Hypothesis

2.1 Introduction

The focus of this chapter is on *incidental lexical learning*, particularly on the incidental acquisition of words through writing activities. The first section will discuss the construct of incidental learning. It will compare contrasting views in the literature and will offer a rationale for the definition adopted in this book. (The reader is advised to also consult Sect. 10.5. for a more thorough discussion on relevant issues connected with the proceduralizing of incidental lexical learning.) The second and third sections will shift focus to the potential of input and output to lexical learning. Finally, this chapter will discuss Laufer and Hulstijn's (2001) *involvement load hypothesis* (ILH). After describing the components and premises of the hypothesis, it will outline the growing body of studies that have investigated the ILH, including some counterevidence and criticism leveled at the hypothesis. At the end, this chapter will explore lexical learning via different types of writing through the lens of the ILH.

2.2 Defining Incidental Lexical Learning

The definition of incidental learning is complex and controversial. Krashen (1982, 1989), singling out input—particularly reading—as the sole responsible for subconscious acquisition, defines incidental learning as subconscious learning: "language is subconsciously acquired—while you are acquiring, you don't know you are acquiring (…) Thus, the acquisition process is identical to what has been termed 'incidental learning'" (Krashen, 1989, p. 440). However, for most researchers, unconscious learning, or learning without awareness—what Krashen (1989) has called "subconscious acquisition"—better describes *implicit learning*, not incidental learning (see DeKeyser, 1994, 2013; Ellis, 1994; Hulstijn & Schmidt, 1994; Laufer & Hulstijn, 2001; Schmidt, 1990, 1994, 2010). Importantly, Laufer and Hulstijn (2001) note that incidental learning may be both implicit or *explicit* (i.e., conscious). They add that

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_2

incidental contrasts with *intentional learning*: that is, learning with the "deliberate decision to commit words to memory", typically when learners are forewarned of an upcoming retention test (Laufer & Hulstijn, 2001, pp. 10 and 11).

Nevertheless, there is still much disagreement on what constitutes incidental learning. Similarly to Krashen (1982, 1989), some researchers believe that incidental learning can only occur via input. One case in point is Webb (2019), who considers incidental only learning that stems from L2 meaning-focused input, such as "reading, listening and viewing for the purpose of interest, information and enjoyment" (p. 226; see also Uchihara et al., 2019, p. 3 for a similar opinion). However, Webb (2019, p. 226) acknowledges that even "with meaning-focused L2 input, it may not be possible to rule out that there is some intention to learn language". In other words, it is problematic to contrast incidental with intentional learning in that it may be impossible to ascertain complete lack of intentionality. As a result, Webb (2019, p. 226) prefers to define incidental learning as a "by-product of meaning focused activities or tasks", irrespective of intentionality. Elgort et al. (2017) agree that when exposed to input, learners may deliberately (i.e., intentionally) attempt to attend to form and infer meaning (see also Paribakht & Wesche, 1999 below). Thus, the researchers prefer the term "contextual word learning" to "incidental learning" to eschew any assumption of lack of intentionality when learning through input. Furthermore, De Vos et al. (2018)—also agreeing with the problematic nature of any assumption of lack of intentionality—advise researchers to interview participants post-experiment in order to eliminate any risk of intentional learning (see also Rice & Tokowicz, 2020, p. 25). As a result, the common definition of incidental learning as being unintentional, or not deliberate (e.g., see Barcroft, 2004; Hulstijn & Laufer, 2001; Hu & Nassaji, 2016; Laufer, 2003; Laufer & Hulstijn, 2001; Ortega, 2009 for such definitions), may be misleading to the point of not being useful.

Other also widely mentioned criteria that learning needs to meet to be deemed incidental do not assume lack of intentionality and do not restrict incidental learning to learning that occurs through input only. In agreement with Webb (2019) and Uchihara et al. (2019), many researchers maintain that incidental learning should come about as a "*by-product*" *of another activity*: Learners perform a primary task (e.g., reading or writing) while processing information (e.g., lexical items), which could therefore be learnt incidentally (e.g., De Vos et al., 2018; Hu & Nassaji, 2016; Hulstijn & Laufer, 2001; Laufer, 2003; Ortega, 2009). This argument aligns well with Schmidt's (1990, p. 149) perspective, whereby incidental learning occurs "when the demands of a task focus attention on what is to be learned". This means that, for instance, if learners are asked to write a text with pre-specified novel words, the writing task demands will oblige students to incorporate the novel lexical items, likely yielding incidental lexical learning. Such an understanding of "incidental learning" underlies the two quasi-experimental studies that will be reported in Chaps. 6 and 7 of this book.

Since incidental learning must be secondary to the performance of the main task, researchers studying L2 incidental learning often adopt some strategies to make any learning as unintentional as possible. Some of these strategies have become part of the definition of incidental learning. Most, if not all researchers, *do not inform*

learners that they will be tested post-experiment (e.g., Hulstijn, 2003; Hulstijn & Laufer, 2001; Pichette et al., 2012). Hiding the existence of posttests is effectively a long-standing practice, as illustrated by Eysenck's (1982, p. 198, as cited in Laufer & Hulstijn, 2001, p. 10) comparison of incidental and intentional learning: "In operational terms, incidental and intentional learning can be distinguished simply in terms of prelearning instructions that either do, or do not, forewarn subjects about the existence of a subsequent retention test". The reason for this is that if learners are informed of a posttest, the real goal of the experiment may be inadvertently revealed, and participants will likely make a deliberate effort to retain information. Obviously, perhaps, what follows is the need to *hide the true purpose of the experiment* from participants, which is also a criterion that can be found in most definitions of incidental learning (e.g., Craik & Lockhart, 1972; De Vos et al., 2013; Horst, 2005; Silva & Otwinowska, 2018; see De Vos et al., 2018 for a discussion).

Considering the above, it appears that there are three main criteria a study measuring lexical learning must meet to be considered incidental:

1. Learning must be a by-product of the main task.
2. The participants cannot be informed of the purpose of the experiment.
3. Thus, learners cannot be aware of an upcoming vocabulary posttest.

In line with these criteria, in this book incidental learning is operationalized as learning that takes place when participants perform a primary task (i.e., writing) involving the processing of some information (i.e., novel academic words) without being aware of the true purpose of the experiment and without being told in advance that they will be tested afterwards on their recall of that information.

Nonetheless, for participants to be able to properly incorporate novel lexical items into their writing, they must be provided with the *keywords* (i.e., the vocabulary items whose knowledge will be assessed in the posttest), their definitions, and illustrations of use (see Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017 for examples of studies with a similar design). Exposing research participants to the keywords simulates instances when leaners consult a monolingual dictionary (or a teacher) for words they want to use but do not know fully. This is inevitable when trying to use novel words in writing, but will likely increase exposure to, processing of, and hence acquisition of the keywords (Hulstijn et al., 1996; Rott, 2005; Schmitt, 2008; Watanabe, 1997). Still, in the studies reported in this book, when using the keywords not only will all the three criteria for incidental learning be met; also, further measures will be taken to ensure learners are unaware of the upcoming posttest and the purpose of the experiment (see Chap. 6 for a more detailed discussion of these measures). Granted, this enhanced exposure means that incidental learning through writing may be considered more explicit than incidental learning through reading, but both are incidental, and will be treated as such in this book. Incidental learning through writing, as proceduralised here, is definitely not intentional inasmuch as learners are in no way directly or indirectly encouraged to memorize these words.

This section has attempted to shed light on some of the issues around defining incidental learning and has made it clear why learning academic words through writing may be deemed incidental. The next sections will briefly discuss the importance of

incidental learning to language acquisition, will underscore the need to enhance the learning of academic words, and will introduce the reader to the involvement load hypothesis (Laufer & Hulstijn, 2001).

2.3 Research on Incidental Vocabulary Learning Through Input

Research has demonstrated that learners acquire a large number of words via exposure to input, particularly through reading. One case in point is the study conducted by Krashen (1989), which reviewed 144 studies with L1 readers and found strong evidence of incidental lexical learning. (For other studies investigating incidental vocabulary learning with L1 readers see, for example, Batterink & Neville, 2011; Borovsky et al., 2010; Frishkoff et al., 2010). Still, as rightly noted by Horst et al. (1998) and others (e.g., Huckin & Coady, 1999), native speakers may be better equipped than L2 students to learn from input. This is because they are more likely to fully understand the surrounding words and the semantic nuances of the context where the keyword is embedded, thereby increasing the likelihood of accurate inferencing and lexical acquisition. Put differently, a better-understood context may provide more useful information about the meaning of a novel word, making the context more informative. *Context informativeness*, as it is often called in the literature, has indeed been shown to facilitate incidental learning (e.g., Chen et al., 2017; Elgort & Warren, 2014; Frishkoff et al., 2010; Hu, 2013; Joseph & Nation, 2018; Mulder et al., 2019).

More recent studies with L2 readers have found persuasive evidence of incidental learning of word meanings (e.g., Brown et al., 2008; Elgort et al., 2017; Godfroid et al., 2018; Horst et al., 1998; Laufer & Rozovski-Roitblat, 2015; Pellicer-Sánchez & Schmitt, 2010; Sonbul & Schmitt, 2010; Vidal, 2011; Waring & Takaki, 2003) and other aspects of a word's knowledge, such as spelling, association, and syntax (e.g., Chen & Truscott, 2010; Laufer et al., 2004; Laufer & Rozovski-Roitblat, 2015; Webb, 2005, 2007; Webb et al., 2013; see Webb, 2019 for a recent literature review on incidental learning through input). In effect, evidence exists suggesting that even one or two encounters with an unknown word may result in some learning if the context is sufficiently informative (e.g., Bisson et al., 2014; Rott, 1999).

Still, for learning to be effective and long lasting, input needs to be constant and extensive. For example, native-speaking children typically read about one million words a year (Nagy et al., 1985) and "will be exposed to over 40,000 h of their home language by the end of 6 years of schooling" (Elley & Mangubhai, 1983, p. 55). While it is true that adults' experience learning languages and enhanced cognitive capacity may enable them to acquire vocabulary incidentally faster than children do (DeKeyser, 2013; Muñoz, 2006; Murphy, 2014), the process remains slow and gradual (e.g., Berens et al., 2018; Chen & Truscott, 2010; Horst et al., 1998; Nation & Wang, 1999; Schmitt, 2007). In the words of Waring and Takaki

(2003, p. 151), "learners will have to read several hundred or several thousand words in order to learn one new word from their reading".

This is because not all novel words in input are attended to equally. Many of these words will be ignored (Paribakht & Wesche, 1999), and thus will not be processed. If input is not consciously noticed and explicitly attended to, it cannot become intake and hence be retained, a tenet underpinning Schmidt's *noticing hypothesis* (Schmidt, 1990, 1994, 2010; Schmidt & Frota, 1986). Even when words are noticed, they may not be inferred at all, or may be inferred incorrectly (Laufer, 2003; Rott et al., 2002), especially when the surrounding context lacks sufficient informativeness (Hu, 2013; Joseph & Nation, 2018; Randy & Morris, 2017; Webb, 2008). Finally, even in the instances when words are inferred accurately, very few of them will be retained (Laufer, 2003; Mondria & Wit-de Boer, 1991), particularly if there is no further exposure and hence recycling of lexical knowledge (Nation & Wang, 1999).

The prospect of acquiring academic words solely through exposure is even poorer than with general vocabulary, as discussed in Sect. 1.4. Academic vocabulary lacks salience in discourse and is typically abstract and morphologically complex, all of which may hamper learning significantly (Corson, 1997; Lubliner & Hiebert, 2011; Nagy & Townsend, 2012; Vidal, 2011). If this is true in English-speaking contexts, where exposure is high, even after years of instruction in English-medium higher education institutions (e.g., Knoch et al., 2015), then exposure alone is insufficient in foreign language contexts such as Poland. As a result, a more explicit approach to the learning of academic words is needed. One that takes advantage of quotidian university tasks, enhances processing, and consequently accelerates learning.

2.4 Research on Incidental Vocabulary Learning Through Output

A more explicit focus on lexical learning typically results in more effective and efficient learning and retention and may enhance the likelihood of attaining higher levels of productive proficiency (Elgort, 2011; Lyster, 2007; Nation, 2013; Nation & Meara, 2010; Schmitt, 2008). Thus, university students' acquisition of academic vocabulary will be facilitated if they are required to incorporate academic words in production. One simple way to do this is by having students embed pre-specified academic keywords in customary tasks such as academic essay writing and, if necessary, consult dictionaries. If so, learners will perform tasks designed to evaluate their knowledge of content (i.e., the essays) while concurrently exercising their linguistic skills (i.e., attempting to accurately utilize academic words). This is nothing more than the principal underpinning tenet of the now-in-vogue *Content and Language Integrated Learning approach* (CLIL; see Coyle et al., 2010). Namely, in CLIL, which has a dual focus on the teaching of content and the teaching of language, content and language must be integrated, with one not taking precedence over the other (Lasagabaster, 2008), both by learning content through an L2 and learning an

L2 through content (Ruiz de Zarobe, 2008). This book will explore the latter, that is, learning academic words in L2 English through writing essays, the skill learners need to practice in their writing classes. Below, a few factors will be discussed that may help explain why the incorporation of academic keywords in writing may facilitate lexical learning. These are *depth of processing*, *elaboration*, the *output hypothesis*, and *generation.*

Craik and Lockhart's (1972) depth of processing hypothesis argues that information is more likely to be stored in long-term memory when it is processed more deeply, where "greater 'depth' implies a greater degree of semantic or cognitive analysis" (p. 675). Put differently, semantic (deeper) processing may be more facilitative of lexical learning and retention than phonological or orthographic processing. This hypothesis has found support in research conducted by Hyde and Jenkins (1969, 1973). The researchers found that tasks focused on semantic processing yielded higher recall rates than tasks requiring morphological, phonological (1969), or syntactic processing (1973).

Later, Craik and Tulving (1975) postulated that the elaboration of stimuli may be better able to explain retention than depth of processing. For the researchers, depth of processing is a reliable predictor of retention in that "a minimal semantic analysis is more beneficial for memory than an elaborate structural analysis" (Craik & Tulving, 1975, p. 291). Nevertheless, the elaboration of a basic stimulus with further encodings—structural, phonemic, or semantic—is a better predictor of retention than depth of processing. That is, richly encoded material, or material that is processed also quantitatively, is more conducive to learning than material that is processed only semantically, or qualitatively. The writing of complex texts may be a good example of both types of processing, that is, semantic and quantitative. First, it is mostly a meaning-oriented task (Byrnes & Manchón, 2014; Manchón & Williams, 2016; Ruiz-Funes, 2015), and this ensures a large amount of semantic processing. Second, the composition process is recursive—i.e., it involves constant planning, writing, and reviewing; see Chap. 3—which improves the likelihood of different types of encoding occurring. That is, writing argumentative essays may deepen and enrich processing, therefore enhancing lexical learning and retention. Psycholinguistic experiments have also demonstrated the effectiveness of elaboration to lexical learning (see Rice & Tokowicz, 2020 for a recent discussion).

Effectively, language production has long been recognized as a valuable source of learning. Swain's (1985) output hypothesis posits that during production, communication breakdowns or the realization of a linguistic problem *pushes* learners to modify their output. This *pushed output* allows them to notice a gap between their production and the target form (see Schmidt & Frota, 1986) and to test hypotheses in the target language (Swain, 1995). All these deepen language processing, enhance learning, and likely lead to more accurate production (Qi & Lapkin, 2001; Swain, 2000; Swain & Lapkin, 1995). In a similar vein, Joe (1998), drawing on the construct of generation (Slamecka & Graf, 1978; Wittrock, 1974), has demonstrated that a more creative use of vocabulary items in original contexts results in better lexical acquisition. She suggested that incorporating words in novel contexts, such as composing sentences

and essays, as in this book, pushes writers to draw connections between already-acquired knowledge and the new information being processed during production, which facilitates learning. In fact, there is a growing body of persuasive evidence demonstrating learning (see below).

For example, in two studies, Elgort et al. (2018) explored lexical learning with 47 Chinese (study 1) and 50 Dutch speakers (study 2). In each study, learners were required to either actively infer the meaning of the keywords through context or to write down (i.e., copy) the keyword after having read it in sentences. In both studies, the word-writing procedure resulted in more learning of word form and meaning than the active-inference procedure. Importantly, the active-inference procedure forced learners to infer the meaning from context, likely yielding more learning than ordinary reading, where (accurate) inferencing is far from guaranteed (see above). Also, participants in the word-writing group were simply copying the keywords, which is liable to generate less learning than producing words in original contexts.

It seems then that, when compared to widely used tasks such as reading a text or writing an essay, learning through reading in the study was inflated (as inferencing was obligatory) while learning through writing was compromised (since participants merely copied the words); and yet the word-writing group registered more learning. In another relevant study, Pichette et al. (2012) used 203 French-speaking intermediate and advanced learners of English to investigate the effectiveness of reading and writing sentences for incidental lexical learning. Again, writing sentences with keywords resulted in significantly higher gains than reading sentences, both in the immediate and one-week delayed posttests. Laufer (2003) has also found more incidental learning following sentence writing than following reading, even when learners consulted a dictionary. Other studies demonstrating high learning rates following writing tasks are Swain and Lapkin (1995, 2002) and Webb (2005) see also Hulstijn & Laufer (2001), Keating (2008), Kim (2008), Zou (2017) in Sects. 2.4 and 2.5.

Unfortunately, it appears that all existing studies measuring lexical acquisition through writing tasks have utilized sentences or short texts (see Kim, 2008 below for a possible exception) while longer tasks such as essay writing may warrant the employment of demanding cognitive processes that may hinder or facilitate language learning (see Chap. 3 for a comprehensive discussion). As a result, it is still unclear the extent to which complex tasks such as argumentative essay writing benefits lexical learning (Byrnes & Manchón, 2014; Pichette et al., 2012). To help address this issue, the next section will draw on Laufer and Hulstijn's (2001) involvement load hypothesis (ILH).

2.5 The Involvement Load Hypothesis

The ILH is the most widely adopted method designed to evaluate the incidental lexical-learning potential of tasks (Nation, 2013; Nation & Webb, 2011) and is therefore the chosen method for this book.[1] The ILH combines the motivational-cognitive constructs of *need*, *search*, and *evaluation* to identify task-specific criteria that can be "observed, manipulated, and measured" (Laufer & Hulstijn, 2001, p. 13). Each of these components may be absent, moderately present, or strongly present, resulting in an *involvement load* (IL) of 0, 1, and 2, respectively. The ILH assumes that the sum of these three components will provide a measure of the learning potential of tasks, with higher ILs being indicative of more lexical learning and retention.

The motivational component of *need* refers to learners' drive to fulfil the task requirements, which may be imposed on the learner by an external source (e.g., the teacher or the task) or may be self-imposed. *Need* is moderate when imposed by the task, such as when learners are required to fill gaps in a text or write sentences or texts with pre-specified keywords. *Need* is strong when the learners impose on themselves the requirement to incorporate keywords in the task. An example would be when learners, during text composition, decide to use a novel word, perhaps a synonym just found in a thesaurus.

The cognitive components are *search* and *evaluation*, and they are predicated on noticing and the deliberate processing of these novel lexical items (Laufer & Hulstijn, 2001; see also Schmidt, 1990, 1994, 2010). *Search* may be absent or moderate, but it is debatable whether it can also be strong. Originally, Laufer and Hulstijn (2001) asserted that *search* is moderate when there is an attempt to find the meaning of a word by consulting an external source—such as the teacher, a colleague, or a dictionary—otherwise no *search* is induced by the task. *Search* is also absent when the definitions of keywords are provided, such as in a glossary accompanying a task. For instance, a fill-in reading task with the deleted words "listed at the bottom of the text with their translations or explanations [i.e., the glossary] (…) induces moderate need, [but] no search (the words are explained) …" (Laufer & Hulstijn, 2001, p. 17; see also Table 2.1). By contrast, Nation and Webb (2011, p. 4) have pointed out that *search* could be moderate and strong. It would be moderate when there is a need to retrieve or find the meaning of a word, such as when learners come across a novel lexical item in a text. *Search* would be strong when learners need to seek or retrieve the word form (e.g., when writing a text, learners attempt to incorporate a novel word or use a forgotten word by consulting a dictionary or thesaurus). This distinction makes sense, not least as it aligns well with the concepts of receptive (moderate *search*) and productive knowledge (strong *search*), with the latter being more difficult (see Mondria & Wiersma, 2004; Webb, 2005), but also likely more conducive to learning (see Sect. 2.3). Still, I preferred to define *search* as originally conceptualized for two reasons. First, only few researchers have adopted Nation and Webb's (2011)

[1] To my knowledge, there is one other method that serves a similar purpose but that has not received as much attention from researchers as the ILH. It is called *technique feature analysis* (Nation & Webb, 2011).

Table 2.1 Examples of the different tasks' involvement loads

Task	Status of keywords	Need	Search	Evaluation	Task IL
Reading and comprehension questions	Glossed in text but irrelevant to task	–	–	–	0
Reading and comprehension questions	Glossed in text and relevant to task	+[a]	–	–	1
Reading and comprehension questions	Not glossed but relevant to task	+	–/+[b]	–/+[c]	1–3
Reading and comprehension questions and filling gaps	Relevant to reading comprehension. Listed with glosses at the end of text	+	–	+	2
Writing original sentences	Listed with glosses	+	–	++	3
Writing a composition	Listed with glosses	+	–	++	3
Writing a composition	Concepts selected by the teacher (and provided in L1). The L2 learner-writer must look up the L2 form	+	+	++	4
Writing a composition	Concepts selected (and looked up) by L2 learner-writer	++	+	++	5

Note – = absent; + = moderate; ++ = strong
Adapted from Laufer and Hulstijn (2001, p. 18)
[a]Need to understand the word to answer comprehension questions
[b]moderate if learners look for the meaning of the word
[c]moderate if the word is polysemous

suggestion (e.g., Hu & Nassaji, 2016; Kohler, 2014). Second, *search* is absent in all tasks explored in this book, and thus such distinction is of less importance to the results presented here.

Evaluation is the "comparison of a word with other words, a specific meaning of a word with its other meanings, or combining a word with other words in order to assess whether a word … does or does not fit its context" (Laufer & Hulstijn, 2001, p. 15). It is moderate, for example, when learners are given words to fill the gaps in a text or, when consulting a dictionary, they must decide on the most appropriate meaning of a polysemous word. *Evaluation* will be strong when vocabulary items are incorporated in novel contexts, such as sentence or essay writing. Table 2.1 contains several illustrations of tasks and their respective involvement loads.

Generally, research has corroborated the predictive value of the ILH (Nation & Webb, 2011). For example, Huang et al. (2012) meta-analytic study analyzed published and unpublished research and found strong support for the hypothesis. Below, I will briefly outline findings only from research conducted with adults. This is because most research to date has been done with this age group, including the studies reported in this book. The reader may refer to Alcaraz Mármol and

Sánchez-Lafuente (2013) and Silva and Otwinowska (2018) for studies conducted with children.

Hulstijn and Laufer (2001) were the first to set out to find evidence for their hypothesis. They investigated Israeli and Dutch advanced English learners and compared three tasks: reading comprehension (IL = 1), reading + gap filling (IL = 2), and text composition (IL = 3). In line with the predictions of the ILH, the researchers found more learning following tasks with higher involvement loads, especially text composition. Later studies utilized comparable tasks and substantiated Hulstijn and Laufer's (2001) findings, albeit with participants of varied L2s and different proficiencies. Exploring the incidental learning of English words, studies have corroborated the tenets of the ILH with 77 Iranian (Keyvanfar & Badraghi, 2011) and 60 Chinese (Qin & Teng, 2017) low-intermediate learners as well as with 162 (Sarani et al., 2013) and 140 (Soleimani & Rahmanian, 2015) Iranian intermediate participants. Studies have found similar results with 120 advanced learners of English from Iran (Ghabanchi et al., 2012) and 104 upper-intermediate and advanced adults of different nationalities (Kim, 2008). Keating (2008) investigated 79 American elementary learners of Spanish and also found supporting evidence for the ILH (see also Martínez-Fernández, 2008 for another study exploring the learning of Spanish words).

Other studies have measured the incidental vocabulary learning yielded by different tasks and appear to confirm the predictive value of the ILH. Some examples include studies conducted by Teng (2015), with elementary learners of English, and Tajeddin and Daraee (2013), using participants with intermediate proficiency; there have also been studies measuring the lexical learning of advanced learners with listening tasks (e.g., Jing & Jianbin, 2009) and reading tasks (e.g., Hu & Nassaji, 2016; Nassaji & Hu, 2012), and those comparing receptive to productive tasks inducing the same IL (Sarani et al., 2013; Yang & Cao, 2020). Overall, then, there appears to be strong evidence in support of the ILH.

This is not to say that research has not found evidence contra the predictions of the hypothesis. For example, Nassaji and Hu (2012) adopted three tasks, namely reading + multiple-choice glosses (IL = 2; learners need to match the keywords to the definitions provided); reading with comprehension questions and dictionary use (IL = 3); reading with derivationally different keywords (IL = 5). The researchers showed that the first task yielded more learning than the second. Nevertheless, nothing is mentioned about whether in-task performance was monitored, making it impossible to ensure that participants in task 2 (reading comprehension and dictionary use) (a) tried to infer the meaning of keywords in the text, or (b) inferred the meaning, although inaccurately, or (c) consulted a dictionary when necessary. If inferencing was not performed and/or dictionary-use failed to occur, no IL was induced, hence explaining the findings. Effectively, there is some evidence suggesting that learners often fail to use a dictionary when required (e.g., Hulstijn et al., 1996), and when they do, many of them, even of advanced proficiency, misunderstand the meaning or use of the words or focus on the wrong definition of polysemous lexical items (e.g., Nesi & Hail, 2014).

Similarly, Li's (2014) sentence writing task (IL = 3) resulted in less lexical learning than reading + gap filling (IL = 2), with glosses provided for both tasks. Once again, task performance was not monitored, and nothing is written about the types of glosses provided. Therefore, it is impossible to ascertain that participants had sufficient information to fully understand the meaning and usage of the words. If information was insufficient (e.g., only the L1 translation was provided, not examples of use), learners may have been able to perform the gap filling task satisfactorily, but their attempts to incorporate the keywords in semantically and grammatically accurate sentences may well have been foiled. Any learning yielded by the sentence writing task would then have been impeded. In fact, the amount of information provided in glosses—or rather, the insufficiency thereof—appears to be an existing and yet crucial issue that haunts research on the ILH and that will hence be discussed in some detail in Chap. 8.

Few other studies have produced ambivalent results (e.g., Hu & Nassaji, 2016; Sarani et al., 2013); also, there is some disagreement on how the three components may be applied to certain types of tasks. However, these conflicting results and disagreements will not be discussed here as they do not affect our tasks of choice (i.e., sentence writing and composition writing). Interested readers are therefore encouraged to consult Silva and Otwinowska (2018, pp. 211–212; see also Folse, 2006) for more details.

Importantly, however, at least part of the counterevidence against the ILH may be explained by its lack of granularity. In other words, many factors that are known to impact lexical learning are not considered by the hypothesis (see Peters, 2019; Puimège & Peters, 2019; Webb, 2019 for interesting discussions of these factors). Regarding learner-related factors, higher L2 proficiency may increase the likelihood of incidental lexical learning (e.g., Elgort et al., 2015; Feng & Webb, 2019; Kormos & Trebits, 2012; Qian & Lin, 2019; Webb & Chang, 2015). Learners' working memory has also been found to correlate positively with learning (Elgort et al., 2018; see also Biedroń & Pawlak, 2016, pp. 406–408 for a discussion on how working memory may influence language learning). Some context-related factors also mediate vocabulary acquisition. One example is the number of times a word appears in a text, with more repetitions typically yielding more learning (e.g., Brown et al., 2008; Malone, 2018; Rott, 1999; Saragi et al., 1978; Teng, 2019, 2020; Uchihara et al., 2019; Webb, 2019). Also, more informative contexts tend to be more conducive to learning than less informative contexts, as discussed in Sect. 2.2.

Finally, there are several word-related factors known to affect learning. Different parts of speech, for instance, are learnt to different degrees, with nouns typically being acquired faster than verbs or adjectives (Godfroid et al., 2018; Kweon & Kim, 2008; Luke & Christianson, 2016; Pigada & Schmitt, 2006). Also, words that occur more frequently in the L2 are known to be generally learnt before less frequent words (Brysbaert & New, 2009). Additionally, concrete lexical items (i.e., words that are easy to visualize, thus highly imageable, such as "tower") are acquired more easily than abstract words, such as "vision" (Brysbaert et al., 2014; Ellis & Beaton, 1993; Reilly & Kean, 2007; Sadoski, 2005; Schmitt, 2010). Furthermore, words that are central to the main topic of the text (or the "keyness" of a word) may be acquired faster

than other words (Elgort & Warren, 2014). For instance, in a text about immigration, words such as "refugee" may be easier to acquire than "unprecedented". Last, longer words, noncognates, and polysemous lexical items are more challenging to learn than shorter words, cognates, and monosemous items (Vidal, 2011; see also Otwinowska et al., 2020; Otwinowska & Szewczyk, 2019).

Indeed, Laufer and Hulstijn (2001), Hulstijn (2005) acknowledge that the construct of involvement was intended only as a first step, and does not, and cannot, cover all grounds. It represents instead an initial attempt towards proceduralizing widely utilized and often obscure constructs in SLA theory such as *input*, *output*, *noticing*, *depth*, *elaboration*, and *information processing*. This was done to achieve simplicity, so the ILH could be accessible to language teachers—thus enabling them to make better informed decisions—and to researchers. It behooves the latter to control for as many word-, content-, and learner-related factors as the study necessitates. Unfortunately, a dearth of rigor in existing research design appears to have distorted findings, leading to inconsistent results that remain unexplained. The next section will focus on one of these conflicting findings. More specifically, it will analyze ILH studies comparing the lexical learning yielded by sentence writing (SW) and composition writing (CW) tasks, which are the ones investigated in this book.

2.6 The ILH as Applied to Writing

According to the ILH, writing sentences and compositions with pre-specified keywords provided in a glossary induce an IL of 3. There is a moderate *need* (+) to utilize the keywords in writing, moderate since the requirement to use the keywords is imposed by the task, not self-imposed; there is also strong *evaluation* (++) in that the keywords must be used in original contexts. *Search* is absent because a glossary is provided. (See Laufer & Hulstijn, 2001; Hulstijn & Laufer, 2001 for the ILs of SW and CW, respectively.) Truly, writing compositions such as argumentative essays may entail production and cognitive processes that are far different from those involved in SW. Even so, the ILH does not differentiate between the tasks and thus assumes that both yield similar lexical learning. This is because, as explained by Laufer (in personal communication with Kim, 2008), CW may be more complex than SW and may induce a higher "overall task involvement", not least due to the need to maintain cohesion and coherence. Still, Laufer contends that the keywords themselves will be processed similarly, and hence be subjected to the same level of involvement. Only four studies have hitherto compared the lexical learning potential of SW and CW, although with inconsistent results (see Table 2.2).

Almost all studies, barring Zou (2017), found support for the ILH since SW resulted in similar learning to CW. However, at least two of these studies have flaws in research design that may have rendered the results unreliable: namely, Gohar et al. (2018) and Tahmasbi and Farvardin (2017; see Table 2.2. for relevant details on the design of these studies). For one thing, in neither study was there a requirement in the

Table 2.2 Studies comparing sentence writing (SW) and composition writing (CW)

Study	Participants	Task: relevant information	Tests and results
Kim (2008) Experiment 2	Learners from an intensive English program (IEP) in the US and undergraduate students ($M_{age} = 25$) English proficiency: upper-intermediate and advanced	*SW*: sentences minimum length of seven words *CW*: write a descriptive or an argumentative essay with one to three paragraphs; write a "well-developed answer" (p. 324) *Keywords*: 10 low-frequency words *Glosses*: part of speech and quick definition (in English) provided	– VKS[a] – Immediate and delayed posttest – SW = CW
Zou (2017)	Chinese non-English major freshmen at a university in China (Age range = 18–21) English proficiency: IELTS 5.5 (intermediate)	*SW*: sentences minimum length of 10 words *CW*: "write a composition that coherently connects the 10 target words, and correct use of all words was required for task completion" (p. 57). No topic given, no min./max. length specified, and no time limit stipulated *Keywords*: 10 low-frequency words from a text on procrastination *Glosses*: part of speech and quick definition (in English) provided	– VKS – Immediate and delayed posttests – CW > SW
Tahmasbi and Farvardin (2017)	Junior high school students in Iran ($M_{age} = 14.7$) English proficiency: elementary	*SW and CW*: "write semantically acceptable and grammatically correct sentences or paragraphs in 15 min" (p. 4) No min./max. length stipulated *Keywords*: 30 novel words (10 keywords × 3) *Glosses*: not provided. Participants had access to bilingual and monolingual dictionaries	– VKS – Immediate and delayed posttests – SW = CW

(continued)

Table 2.2 (continued)

Study	Participants	Task: relevant information	Tests and results
Gohar et al. (2018)	ESL students in Iran (Age range = 15–25) English proficiency: TOEFL paper-based: 460–490 (intermediate)	*SW*: 10 min to write all sentences; no minimum length specified *CW*: write a letter. No topic given, no min./max. length stipulated, and no time limit specified *Keywords*: 10 novel words *Glosses*: L1 definitions and examples	– L2-L1 translation – Delayed posttest – SW = CW

Note An equal sign (=) means no statistically significant difference in learning between the groups
Adapted from Silva (2019)
[a]The Vocabulary Knowledge Scale (VKS) was developed by Wesche and Paribakht (1996). It is based on a 1 (no knowledge) to 5 (productive use) scale, and data is collected by means of self-reported and demonstrated knowledge (writing a sentence with the keyword). It measures both receptive (lower score) and productive knowledge (higher score)

SW task to produce sentences with a minimum length. This being the case, participants may have written over simplistic sentences that do not even warrant understanding of the keywords. For instance, learners may have produced sentences such as "The proportion is good", which does little to nothing to demonstrate understanding of the keyword "proportion" (keyword taken from Gohar et al., 2018).

With regard to the CW tasks, neither Gohar et al. (2018) nor Tahmasbi and Farvardin (2017) controlled for text length and quality, and consequently their participants may have failed to produce complex texts (see also Zou, 2017 below). However, for Laufer, one of the authors of the ILH, text composition tasks imply the need to maintain cohesion and coherence (see above). Thus, the texts produced in these studies may not have induced the involvement load predicted by the hypothesis. Furthermore, Tahmasbi and Farvardin (2017) did not specify a text type—and different types may have been processed differently, thus affecting learning (see Chap. 3). Also, Gohar et al. (2018) did not impose a time limit, hence allowing learners to allocate attention to the keywords indefinitely and possibly enhancing lexical learning and retention. By contrast, Tahmasbi and Farvardin (2017) allocated only 15 min to SW and CW. This is truly little for their elementary learners, especially because they were required to use dictionaries. Here, CW participants, even assuming they were willing to consult the dictionaries under such time pressure, may have failed to look up the keywords assiduously and to adequately evaluate their meanings in context (see Sect. 2.4), both of which may have affected vocabulary acquisition.

Kim (2008) and Zou (2017) appear to have adopted more rigorous study designs and will therefore make up the thrust of the discussion here and later in this book. These two studies were rather similar in design and yet reported contradictory results: Kim (2008) found similar learning following SW and CW while Zou (2017) showed higher lexical gains for CW participants, which runs counter to the predictions of the ILH. Zou (2017) claimed that CW is more conducive to lexical learning than SW because of the need to chunk information and to organize them hierarchically. This means that learners need to process information, including the keywords, in order to organize them into semantically related chunks (i.e., coherently and cohesively connected groups of sentences). Participants also needed to ensure that each newly written chunk was meaningfully connected to the previously generated ones. Moreover, these chunks needed to be organized into a coherent whole, which the author called *hierarchical organization*, just as paragraphs need to be carefully organized in a text. All this chunking and effort to maintain coherence and cohesion likely enhanced processing of the keywords, therefore increasing learning. On the other hand, sentence writers needed only compose separate contexts for each TW.

Zou's (2017) argument appears reasonable; however, her results may also be explained by how her participants performed the CW task. Zou (2017) instructed learners to write a composition while accurately and coherently connecting the 10 keywords but did not specify a minimum or maximum length. This may have given learners free rein to compose too short or too long texts, likely distorting results. A closer look at a sample essay produced by one of Zou's (2017) participants may help elucidate this point (see Fig. 2.1).

People who has assiduous trait will get lassitude easily. **BECAUSE** they usually worry about some events are indispensable, pernicious or apprehensive. Sometimes, it is unnecessary to care too much about them, because these events are ostensible.

FOR EXAMPLE, if Jim divulges a secret to his friends and let them keep it confidential. However, later, Jim may worry about whether his friends will renege or not, or taunt him about it in some other places.

Fig. 2.1 Sample composition from Zou's (2017) participant (p. 67). *Note* The keywords are underlined. The words in bold are linking devices used by participants to maintain coherence

Based on the sample essay, it appears unlikely that learners were able, or even attempted, to write a coherently well-developed composition, as expected by the ILH and claimed by Zou (2017). For one thing, the text has only 74 running words, 10 of which (or 13.51%) are the keywords. The result is that the keywords are chunked, often as a list, which makes it rather difficult, if not impossible, to ensure they were used accurately, and that their meaning was understood. For instance, in the sentence "Because they usually worry about [*sic*] some events are indispensable, pernicious, or apprehensive", all three keywords could mean many things.

Additionally, it appears that learners focused on the keywords to the detriment of the text. First, the message is rather unclear, particularly the three sentences in the first paragraph. Second, the text contains many errors, making understanding even harder. This lack of clarity and the many existing inaccuracies may indicate learners' unwillingness or incapacity to focus on textual construction and keyword use concurrently, rather allocating most of their attentional resources to the latter. This explanation appears all the more reasonable when considering that both CW and SW participants took, on average, 34 min to complete the tasks. Writing a text, especially a coherent one, as intended by Zou (2017), should take considerably longer than writing individual sentences, unless the writer's main concern is to incorporate the keywords. This a problem since textual composition, not keyword use, is supposed to be the primary task in an incidental learning study such as this one (see Sect. 2.1). And yet, this extra attention devoted to the keywords may be exactly what made CW more conducive to learning than SW.

As a result, one possible explanation for the contradictory findings in Zou's (2017) and Kim's (2008) studies may be the following. Zou's (2017) participants, being non-English majors with lower-intermediate proficiency in English, lacked sufficient writing and linguistic skills to be able to allocate attentional resources to both the composition process and keyword use, and hence opted to focus on the latter, which enhanced lexical learning. By contrast, learners in Kim's (2008) experiment were not only more proficient, but they were also students in pre-university Intensive English Programs and undergraduate students at an English-medium university in the US.

Consequently, it is highly likely that they were far more accustomed to writing in English than Zou's (2017) learners. Being more experienced in writing and more proficient in English, these learners did not have to choose where to focus their cognitive resources on, but rather allocated some attention to both text composition and keyword use. This led to lexical learning, but only to the same level as SW. What this all means is that the amount of vocabulary learning yielded by CW may depend on learners' capacity to allocate attentional resources to the keywords, which is predicated on how cognitively overwhelming the task is. This hypothesis is addressed in the second and third studies reported in this book.

2.7 Conclusion

This chapter has focused on research on incidental lexical acquisition and has argued for the need for a more explicit approach to the teaching of academic words, such as by incorporating these words in academic writing. To this aim, the chapter has focused on the ILH and on studies drawing on this hypothesis, including studies on learning via written production. Still, as mentioned above, there is little research measuring vocabulary learning through the writing of texts, and most studies have implemented short texts with low complexity. This means that research on the lexical learning potential of more complex writing, such as argumentative essay writing, remains largely unexplored (Byrnes & Manchón, 2014; Pichette et al., 2012). This has led Manchón and Williams (2016, p. 577) to wonder "whether L2 knowledge can actually be created as a result of production processes", describing it as an essential question for SLA and L2 writing research.

I have argued here that the learning yielded by CW may depend on the cognitive complexity of the task and on learners' capacity to process information. The participants investigated in this book are all Polish speakers of similar English proficiency and background, and all of them first-year university students at the Institute of English Studies, University of Warsaw. As a result, they may possess similar cognitive capacity and experience when writing in English. The variable of interest is therefore the complexity of the writing task and how it may affect learning. Chapter 3 will explore this in detail.

References

Alcaraz Mármol, G., & Sánchez-Lafuente, A. A. (2013). The involvement load hypothesis: Its effect on vocabulary learning in primary education. *Revista Española de Linguística Aplicada/Spanish Journal of Applied Linguistics*, *26*, 11–24.

Barcroft, J. (2004). Effects of sentence writing in second language lexical acquisition. *Second Language Research, 20*(4), 303–334. https://doi.org/10.1191/0267658304sr233oa

Batterink, L., & Neville, H. (2011). Implicit and explicit mechanisms of word learning in a narrative context: An event-related potential study. *Journal of Cognitive Neuroscience, 23*(11), 3181–3196.

Berens, S. C., Horst, J. S., & Bird, C. M. (2018). Cross-situational learning is supported by propose-but-verify hypothesis testing. *Current Biology, 28*, 1132–1136.

Biedroń, A., & Pawlak, M. (2016). The interface between research on individual difference variables and teaching practice: The case of cognitive factors and personality. *Studies in Second Language Learning and Teaching, 6*(3), 395–422. https://doi.org/10.14746/ssllt.2016.6.3.3

Bisson, M.-J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). The role of repeated exposure to multilingual input in incidental acquisition of foreign language vocabulary. *Language Learning, 64*(4), 855–877.

Borovsky, A., Kutas, M., & Elman, J. (2010). Learning to use words: Event-related potentials index single-shot contextual word learning. *Cognition, 116*, 289–296.

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition through reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language, 20*(2), 136–163.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.

Byrnes, H., & Manchón, R. M. (2014). Task, task performance and writing development: Advancing the constructs and the research agenda. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 267–299). John Benjamins.

Chen, B., Ma, T., Liang, L., & Liu, H. (2017). Rapid L2 word learning through high constraint sentence context: An event-related potential study. *Frontiers in Psychology, 8*, 1–14.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics, 31*(5), 693–713.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718.

Coyle, D., Hood, P., & Marsh, D. (2010). *Content and language integrated learning.* Cambridge University Press.

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

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology, 104*(3), 263–294.

De Vos, J. F, Schriefers, H., & Lemhöfer, K. (2013). Noticing vocabulary holes aids incidental language word learning: An experimental study. *Bilingualism: Language and Cognition, 22*(3), 500–515.

De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning, 68*(4), 906–941.

DeKeyser, R. M. (1994). How implicit can adult second language learning be? *International Association of Applied Linguistics (AILA) Review, 11*, 83–96.

DeKeyser, R. M. (2013). Age effects in second language learning: Stepping stones toward better understanding. *Language Learning, 63*(1), 52–67.

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

Elgort, I., & Warren, P. (2014). L2 vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning, 64*(2), 365–414.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2017). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition, 40*, 341–366.

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

Elgort, I., Perfetti, C., Rickles, B., & Stafura, J. (2015). Contextual learning of L2 word meanings: Second language proficiency modulates behavioral and ERP indicators or learning. *Language, Cognition and Neuroscience, 30*(5), 506–528.

Elley, W. B., & Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly, 19*(1), 53–67.

Ellis, N. (1994). Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In N. Ellis (Ed.), *The implicit and explicit learning of languages* (pp. 211–282). Academic Press.

Ellis, N. C., & Beaton, A. (1993). Psycholinguistic determinants of foreign language vocabulary learning. *Language Learning, 43*(4), 559–617.

Feng, Y., & Webb, S. (2019). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition*, 1–25.https://doi.org/10.1017/S0272263119000494

Folse, K. S. (2006). The effect of type of written exercise on L2 vocabulary retention. *TESOL Quarterly, 40*(2), 273–293.

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology, 35*(4), 376–403.

Ghabanchi, Z., Davoudi, M., & Eskandari, Z. (2012). Vocabulary learning through input and output tasks: Investigating the involvement load hypothesis. *California Linguistic Notes, 37*(1), 2–18.

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H.-J. (2018). Incidental vocabulary learning in a natural reading context: an eye-tracking study. *Bilingualism: Language and Cognition, 21*(3), 563–584.

Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique feature analysis or involvement load hypothesis: Estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research, 47*, 859–869.

Horst, M. (2005). Learning L2 vocabulary through extensive reading: A measurement study. *The Canadian Modern Language Review, 61*(3), 355–382.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207–223.

Hu, M.H.-C. (2013). The effects of word frequency and contextual types on vocabulary acquisition from extensive reading: A case study. *Journal of Language Teaching and Research, 4*(3), 487–495.

Hu, M.H.-C., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System, 56*, 28–39.

Huang, S., Eslami, Z., & Willson, V. (2012). The effects of task involvement load on L2 incidental vocabulary learning: A meta-analytic study. *The Modern Language Journal, 96*(4), 544–557.

Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition, 21*, 181–193.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell Publishing.

Hulstijn, J. H. (2005). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell.

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., & Schmidt, R. (Eds.). (1994). Consciousness in second language learning. *International Association of Applied Linguistics (AILA) Review, 11*. Retrieved from: https://www.academia.edu/10199589/AILA11_consciousness_in_sla. Accessed March 04, 2022.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80*(3), 327–339.

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology, 82*(3), 472–481.

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and Syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behaviour, 12*, 471–480.

Jing, L., & Jianbin, H. (2009). An empirical study of the involvement load hypothesis in incidental vocabulary acquisition in EFL listening. *Polyglossia, 16*, 1–11.
Joe, A. (1998). What effects to task-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics, 19*(3), 357–377.
Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology, 166*, 190–211.
Keating, G. D. (2008). Task effectiveness and word learning in a second language: The involvement load hypothesis on trial. *Language Teaching Research, 12*(3), 365–385.
Keyvanfar, A., & Badraghi, A. H. (2011). Revisiting task-induced involvement load and vocabulary enhancement: Insights from the EFL setting in Iran. *Žmogus Ir Žodis, 13*(3), 56–66.
Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.
Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university. *Journal of Second Language Writing, 28*, 39–52.
Kohler, B. T. (2014). Perspectives on the involvement load hypothesis. *Enjoy Teaching Journal, 2*(3), 23–29.
Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning, 62*(2), 439–472.
Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Pergamon.
Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence by the input hypothesis. *The Modern Language Journal, 73*(4), 440–464.
Kweon, S., & Kim, R. (2008). Beyond raw frequency: Incidental vocabulary acquisition in extensive reading. *Reading in a Foreign Language, 20*(2), 191–215.
Lasagabaster, D. (2008). Foreign language competence in content and language integrated learning. *The Open Applied Linguistics Journal, 1*, 31–42.
Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review, 59*(4), 567–588.
Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1–26.
Laufer, B., & Rozovski-Roitblat, B. (2015). Retention of new words: Quantity of encounters, quality of task, and degree of knowledge. *Language Teaching Research*, *Special issue*, *19*(6), 687–711.
Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing, 21*, 202–226. https://doi.org/10.1191/0265532204lt277oa
Li, J. (2014). Effect of task-induced online learning behaviour on incidental vocabulary acquisition by Chinese learners—Revisiting the involvement load hypothesis. *Theory and Practice in Language Studies, 4*(7), 1385–1394.
Lubliner, S., & Hiebert, E. H. (2011). An analysis of English-Spanish cognates as a source of general academic language. *Bilingual Research Journal, 34*(1), 76–93.
Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22–60.
Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach.* John Benjamins B.V.
Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition, 40*, 651–675.
Manchón, R. M., & Williams, J. (2016). Introduction: SLA-L2 writing interfaces in historical perspective. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 567–586). De Gruyter.
Martínez-Fernández, A. (2008). Revisiting the involvement load hypothesis: Awareness, type of task and type of item. In *Selected Proceedings of the 2007 Second Language Research Forum*

(pp. 210–228). Retrieved from: http://www.lingref.com/cpp/slrf/2007/paper1746.pdf. Accessed February 16, 2019.
Mondria, J., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 79–100). John Benjamins Publishing.
Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics, 12*(3), 249–267.
Mulder, E., Van de Ven, M., Segers, E., & Verhoeven, L. (2019). Context, word, and student predictors in second language vocabulary learning. *Applied Psycholinguistics, 40*, 137–166.
Muñoz, C. (2006). The BAF project: Research on the effects of age on foreign language acquisition. *Linguistic Insights, 22*, 83–94.
Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts.* Oxford University Press.
Nagy, W. E., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91–108.
Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*(2), 233–253.
Nassaji, H., & Hu, H. M. (2012). The relationship between task-induced involvement load and learning new words from context. *International Review of Applied Linguistics, 50*, 69–86.
Nation, I. S. P. (2013) *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
Nation, I. S. P., & Meara, P. (2010). Vocabulary. In N. Schmitt (Ed.), *An introduction to applied linguistics* (2nd ed., pp. 34–52). Hodder Education.
Nation, I. S. P., & Wang, M. K. (1999). Graded readers and vocabulary. *Reading in Foreign Language, 12*(2), 355–380.
Nation, I. S. P., & Webb, S. (2011) *Researching and analysing vocabulary*. Heinle Cengage Learning.
Nesi, H., & Haill, R. (2014). A study of dictionary use by international students at a British university. *International Journal of Lexicography, 15*(4), 277–305.
Ortega, L. (2009). Sequences and processes in language learning. In M. H. Long & C. J. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 81–105). Wiley-Blackwell.
Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism, 22*(8), 974–991.
Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross linguistic similarity matter? *Language Learning,* 1–47.https://doi.org/10.1111/lang.12390
Paribakht, T. S., & Wesche, M. (1999). Reading and "incidental" L2 vocabulary acquisition: An introspective study of lexical inferencing. *Studies in Second Language Acquisition, 21*, 195–224.
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.
Peters, E. (2019). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). Routledge.
Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and sentence writing for second language vocabulary acquisition. *Applied Linguistics, 33*(1), 66–82.
Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition through extensive reading: A case study. *Reading in a Foreign Language, 18*(1), 1–28.
Puimège, E., & Peters, E. (2019). Learner's English vocabulary knowledge prior to formal instruction: The role of word-related and learner-related variables. *Language Learning, 69*(4), 943–977.
Qi, D. S., & Lapkin, S. (2001). Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing, 10*, 277–303.

Qian, D. D., & Lin, L. H. F. (2019). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 66–80). Routledge.
Qin, C., & Teng, F. (2017). Assessing the correlation between task-induced involvement load, word learning, and learners' regulatory ability. *Chinese Journal of Applied Linguistics, 40*(3), 261–280.
Randy, L., & Morris, R. K. (2017). Impact of contextual constraint on vocabulary acquisition in reading. *Journal of Cognitive Psychology, 29*(5), 551–569.
Reilly, J., & Kean, J. (2007). Formal distinctiveness of high- and low-imageability nouns: Analyses and theoretical implications. *Cognitive Science, 31*, 157–168.
Rice, C. A., & Tokowicz, N. (2020). State of the scholarship: A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition, 42*, 439–470.
Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition, 21*(4), 589–619.
Rott, S. (2005). Processing glosses: A qualitative exploration of how form-meaning connections are established and strengthened. *Reading in a Foreign Language, 17*(2), 95–124.
Rott, S., Williams, J., & Cameron, R. (2002). The effect of multiple-choice L1 glosses and input-output cycles on lexical acquisition and retention. *Language Teaching Research, 6*(3), 183–222.
Ruiz de Zarobe, Y. (2008). CLIL and foreign language learning: A longitudinal study in the Basque country. *International CLIL Research Journal, 1*(1), 2008.
Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing, 28*, 1–19.
Sadoski, M. (2005). A dual coding view of vocabulary learning. *Reading & Writing Quarterly, 21*(3), 221–238.
Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System, 6*(2), 72–78.
Sarani, A., Negari, G. M., & Ghaviniat, M. (2013). The role of task type in L2 vocabulary acquisition: A case of involvement load hypothesis. *Maringá, 35*(4), 377–386.
Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. In J. Hulstijn & R. Schmidt (Eds.), *Consciousness in second language learning* (Vol. 11, pp. 11–26). AILA Review. Retrieved from http://www.aila.info/download/publications/review/AILA11.pdf#page=11
Schmidt, R. (2010). Attention, awareness, and individual differences in language learning. In W. M. Chan, S. Chi, K. N. Cin, J. Istanto, M. Nagami, J. W. Sew, T. Suthiwan & I. Walker (Eds.), *Proceedings of CLaSIC 2010*, Singapore, December 2–4 (pp. 721–737). National University of Singapore, Centre for Language Studies.
Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158.
Schmidt, R. W., & Frota, S. N. (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). Newbury House.
Schmitt, N. (2007). Current trends in vocabulary learning and teaching. In J. Cummins & C. Davison (Eds.), *The international handbook of English language teaching* (Vol. 2). Springer.
Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. *Language Teaching Research, 12*(3), 329–363.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.
Silva, B. (2019). *Learning academic words through writing: Can cognitive load affect task involvement* [Unpublished MA thesis]. School of Education, University of Nottingham.
Silva, B., & Otwinowska, A. (2018). Vocabulary acquisition and young learners: Different tasks, similar involvement loads. *International Review of Applied Linguistics, 56*(2), 205–229.
Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604.

Soleimani, H., & Rahmanian, M. (2015). Visiting involvement load hypothesis and vocabulary acquisition in similar task types. *Theory and Practice in Language Studies, 5*(9), 1883–1889.
Sonbul, S., & Schmitt, N. (2010). Direct teaching of vocabulary after reading: Is it worth the effort? *ELT Journal, 64*(3), 253–260.
Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House Publishers.
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 Henry Widdowson* (pp. 125–144). Oxford University Press.
Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. Lantolf (Ed.), *Sociocultural theory*
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. (2002). Talking it through: Two French immersion learners' response to reformulation. *International Journal of Educational Research, 37*, 285–304.
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,* 1–10.https://doi.org/10.1177/2158244017730596
Tajeddin, Z., & Daraee, D. (2013). Vocabulary acquisition through written input: Effects of form-focused, message-oriented, and comprehension tasks. *Teaching English as a Second or Foreign Language, 16*(4), 1–19.
Teng, F. (2015). Involvement load in translation tasks and EFL vocabulary learning. *The New English Teacher, 9*(1), 83–101.
Teng, M. F. (2019). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *The Language Learning Journal, 47*(2), 145–158.
Teng, M. F. (2020). Retention of new words learned incidentally through reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research, 24*(6), 785–812. https://doi.org/10.1177/1362168819829026
Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 1–41.https://doi.org/10.1111/lang.12343
Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258.
Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130–163. http://nflrc.hawaii.edu/rfl/October2003/waring/waring.pdf
Watanabe, Y. (1997). Input, intake and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*, 287–307.
Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*, 33–52.
Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65.
Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language, 20*(2), 232–245.
Webb, S. (2019). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). Routledge.
Webb, S., & Chang, A.C.-S. (2015). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research, 19*(6), 667–686.
Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91–120.

Wesche, M., & Paribakht, T. M. (1996). Assessing vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13–40.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*(2), 87–95.

Yang, Y., & Cao, X. (2020). Effects of task involvement load on l2 vocabulary acquisition and their association with language aptitude. *Asia-Pacific Educational Research.* https://doi.org/10.1007/s40299-020-00528-8

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.

Chapter 3
The Writing Cycle and Cognitive Processes that May Affect Learning

3.1 Introduction

As briefly discussed in the previous chapter, text production, particularly the writing of longer, complex texts, may affect language processing and allow for different learning opportunities. This chapter will focus on the mechanisms underlying the writing of complex texts such as argumentative essays, which Hayes (2012) refers to as *formal writing*. The reason for this is that such types of essays pervade higher education, and their potential to afford the learning of academic words is therefore explored in the second and third studies (Chaps. 6 and 7). The first section of this chapter will focus on the recursivity of the writing process. The following section will then draw attention to how this process may affect a writer's cognitive capacity, and how this may affect writing performance and learning. Finally, the third section will highlight a few relevant hypotheses that may help researchers detect signs of increased cognitive load in writing and explain changes in lexical learning that may have stemmed from this increase in cognitive load.

3.2 The Writing Process

Few models devised to study the writing process in L1 contexts have influenced L2 writing research. Among them, the landmark model by Flower and Hayes (1981), Hayes and Flower (1980) is especially influential and will thus be discussed here in some detail. The model introduces a theory of cognitive processes that take place when composing. This theory hinges upon four crucial points which are better illustrated when quoted in their entirety (Flower & Hayes, 1981, p. 366):

1. The process of writing is best understood as a set of distinctive thinking processes which writers orchestrate or organize during the act of composing.
2. These processes have a hierarchical, highly embedded organization in which any given process can be embedded within any other.

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_3

3. The act of composing itself is a goal-directed thinking process, guided by the writer's own growing network of goals.
4. Writers create their own goals in two key ways: by generating both high-level goals and supporting sub-goals which embody the writer's developing sense of purpose, and then, at times, by changing major goals or even establishing entirely new ones based on what has been learned in the act of writing.

The first point posits that the thinking processes underlying writing do not consist of a linear series of stages that are independent and separated in time. This may seem evident, but it is not uncommon to conceptualize writing in a linear, sequential fashion: pre-writing (i.e., the planning stage), the writing per se, and the re-writing, when the text is reviewed and edited to its final product. Nevertheless, as rightly pointed out by Flower and Hayes (1981), this conceptualization of writing oversimplifies the process as it assumes independence between planning, writing, and reviewing. In Flower and Hayes' (1981) model, writing is a cyclical process comprised of the three main processes of *planning*, *translating* (or *formulating*; i.e., the writing process itself), and *reviewing*, which are controlled by the *monitor*. These three processes are co-dependent and recursive and may occur during any time during writing. Monitoring refers to writers' cognitive ability to control the writing process, to track their progress, and to decide when to move between planning, formulating, and reviewing.

Planning is a hierarchical system (see point 2 above) divided into the sub-processes of *generating ideas*, *organizing*, and *goal setting*. Writers need to retrieve information from long-term memory in order to generate the ideas they want to incorporate in the compositions. Sometimes these ideas are well structured and developed, especially among experienced writers, to the extent that the information retrieved almost equates to standard written English. At other times, these ideas may be disconnected, fragmented, and incoherent, thus necessitating higher levels of organization (Flower & Hayes, 1981). Organizing allows writers to structure the text into a coherent whole; also, the organizational process leads to the identification of categories and to the search for subordinate and superordinate ideas in order to expand a topic. Obviously, the way in which a text will be organized and its ideas expanded depends on the effect writers wish the final product to have on the audience; that is, organization is contingent upon the goals set by the writer (point 3 above).

Goal setting may be related to the structure of the text or to the ideas that make up the message thereof (see point 4 above). Writers often must make moment-to-moment structural decisions such as whether to alter the topic sentence, clarify the link between sentences, change the punctuation, or improve sentence clarity by simplifying its syntactic structure or lexical density. Also, and perhaps self-evidently, writers need to decide on the ideas to be included so the text achieves the intended effect on the reader (i.e., a high-level goal). More importantly, most of the writer's goals are created, expanded, and revised by the very same processes underlying idea generation and organization. Put differently, these three sub-processes of planning are constantly interacting in a rather complex, recursive process. Writers generate ideas that need to be organized depending on the goals set. Organization results in

revised goals which are then used to generate ideas. These ideas induce the need to re-evaluate goals and reorganize information, which in turn leads to further idea generation, organization, and goal setting (Flower & Hayes, 1981).

Formulating and reviewing are the two other processes underlying writing. Formulating is fundamentally the process of putting ideas originated during the planning stage into visible language. It is during formulating that writers need to tackle the demands of written English, such as grammar, spelling, vocabulary, and coherence (Cumming, 1990). During formulating, any stress in a writer's cognitive capacity may interfere with planning, thus affecting the whole writing process (see below). Reviewing is divided into *evaluating* and *revising*. Reviewing may be a planned conscious process, such as when a writer deliberately reads what has been written, and often results in further planning and formulating. It may also be unplanned and typically occurs during formulation. This is when writers evaluate the text produced or their yet-unwritten planning (e.g., their thoughts) and may as a result set new goals, generate, and organize new ideas, even before writing.

Clearly, this cyclical process comprised of planning, formulating, and reviewing is hierarchical (point 2) and goal oriented (points 3 and 4 above; Flower & Hayes, 1981). First, it is hierarchical since each process may be "embedded within another process or even within another instance of itself" (Flower & Hayes, 1981, p. 375). An example of a process within another process is the fact that reviewing may occur during planning or formulating; an example of a process embedded within itself is the three subprocesses of planning—namely, generating ideas, organizing, and goal setting—which influence each other and often occur several times during one planning process (see above). Second, this cyclical process underlying writing is goal oriented as it is a writer's goals that guide the writing itself. A text may have one or more main goals, but a sentence, a word, or even a punctuation mark may be employed as a consequence of sub-goals that are guided by and aim to achieve the higher-level goals. I may choose a word to attain irony, for instance, which helps me build the criticism necessary to set the tone of my main message (i.e., my main goal). Furthermore, as outlined above, goals are being constantly re-evaluated and changed, which in turn may result in different ideas that need to be organized, translated, and reviewed.

In short, writing is a recursive, cyclical process comprised of complex co-dependent operations that may occur at any time during written composition (Flower & Hayes, 1981). On the one hand, the recursivity of writing enhances processing. Recursivity is an intense linguistic problem-solving process (Hyland, 2016; Manchón, 2014) that has been perceptively described by Cumming (1990, p. 491) as "reasoning about linguistic choices". This writing recursivity, combined with the time availability that is inherent to writing tasks, may facilitate language learning (Byrnes & Manchón, 2014; Casanave, 2016; Lee, 2016; Manchón, 2014, 2016; Manchón & Williams, 2016; Ruiz-Funes, 2015), including lexical learning, as discussed in Chap. 2. On the other hand, the complexity underlying writing processes may hinder vocabulary acquisition. This is because this complexity may overwhelm learners' cognitive capacity, impeding the allocation of sufficient mental resources to certain aspects of the text, and thus possibly inhibiting learning (Kellogg et al., 2013;

Stevenson et al., 2006). This may be true especially among learners below a certain L2 threshold and/or inexperienced writers (Ruiz-Funes, 2014; Schoonen et al., 2009), and/or writers under time pressure (Kormos & Trebits, 2012). The next sections will focus on some learner- and task-related aspects that may influence writers' ability to allocate attentional resources to linguistic matters when composing complex texts.

3.3 The Allocation of Attentional Resources in Formal Writing

Writing in a foreign language is a cognitive-intensive process that drains resources from learners' *working memory* (WM). Biedroń and Szczepaniak (2012, p. 290) define WM as "the temporary storage and manipulation of information that is necessary for the performance of a wide range of cognitive tasks" and explain that WM consists of mechanisms for the storage of information and its executive control. According to a more precise definition, WM refers to the ability "to control attention with a view to actively maintaining and processing a limited amount of stimulus under conditions of interference" (Alptekin et al., 2014, p. 537). Put differently, multiple stimuli (both auditory and visual) may be stored, controlled, and manipulated in WM at the same time (Bohn-Gettler & Kendeou, 2014), but only temporarily and to a limited extent (Biedroń & Szczepaniak, 2012). This all means that WM has a limited capacity, so the amount of information manipulated and processed at the same time is restricted. Therefore, when too much information is processed too intensively, WM may be overloaded. As further proposed by the *cognitive load theory* (Paas & Van Merriënboer, 1994) when WM is overloaded by too much information, learners' ability to allocate attention to a task is significantly hamstrung. Conversely, as WM becomes less stressed, task performance and possibly learning may increase (Klepsch et al., 2017; Lee, 2019; Paas et al., 2003).

Research has consistently shown that a larger WM capability, which allows for freer WM performance on task, is positively correlated with lexical learning: Learners with better WM tend to learn more words (e.g., Elgort et al., 2018; Juffs & Harrington, 2011; Linderholm & Van den Broek, 2002; Speciale et al., 2004; Varol & Erçetin, 2016). However, studies exploring the role of WM with L2 writing tasks are very limited (Juffs & Harrington, 2011). One case in point is Elgort et al.'s (2018) study with Chinese and Dutch learners of English (see Sect. 2.3). Here, learners either saw words in sentences and wrote them down (word-writing condition) or were required to infer the meaning of the keywords read in the sentences. In both conditions and with participants of both L1s, WM capacity correlated positively with knowledge of word form. Of importance, participants with larger WM benefited from this advantage more in the word-writing condition than in the inferencing condition (Elgort et al., 2018, p. 657). This may mean that WM capacity modulates learning through writing more than it modulates learning through reading, possibly owing to the higher cognitive demands of writing tasks.

It is indeed possible that the complexity underlying the writing process may place a large burden on learners' WM, undermining the allocation of sufficient cognitive resources to certain aspects of the text (Kellogg et al., 2013; Stevenson et al., 2006). In fact, researchers have long argued that WM may be overloaded by the writing process. Hayes and Flower (1980) model of L1 writing has postulated that learners' WM capacity is overloaded by the need to juggle planning, formulating, and reviewing (see above). Later, Kellogg's (1990) *overload hypothesis*, also dealing with L1 writing, posited that the act of composing typically overloads attentional capacity, often resulting in a decrease in writing performance. As explained by Kellogg et al. (2013), linguistic encoding is multi-faceted. Learners must retrieve from memory syntactic information, phonological information (i.e., the covert inner speech that accompanies writing), and orthographic forms, not to mention the lexical, stylistic, and pragmatic choices learners must make. Even the motor processes of handwriting or typing, if not sufficiently automated, may increase cognitive load. The result is that even the writing of isolated sentences may place a substantial burden on WM capacity (Kellogg et al., 2013).

3.3.1 The Role of L2 Proficiency in Lexical Learning Through Writing

Bearing in mind the specific demands of L2 writing, more recently, Stevenson et al.'s (2006) *inhibition hypothesis* also predicted that the linguistic demands of formal writing overloads cognitive resources, inhibiting attention to content elaboration and text organization. Such heavy cognitive demands of writing make the L2 proficiency of writers essential when trying to predict the vocabulary learning potential of such tasks (Gánem-Gutierrez & Gilmore, 2018; Manchón, 2014; Manchón & Roca de Larios, 2007; Ortega, 2012). This is because, as Stevenson et al. (2006) suggest, attentional resources are more easily overwhelmed among inexperienced, usually lower-proficiency, writers. Indeed, research has shown that lower-level L2 learners have problems coping with the cognitive demands of formal writing.

Ruiz-Funes (2015) asked English speakers of intermediate and advanced proficiency in L2 Spanish to write essays of varying complexity. The researcher concluded that lower-proficiency writers were unable to allocate sufficient attentional resources to both the formal demands of text (i.e., linguistic aspects, including vocabulary) and conceptual demands (i.e., content and organization) of argumentative-essay writing. Instead, learners focused solely on the formal side of their writing. A similar conclusion was reached by Manchón et al. (2009) in their study with 21 Spanish pre-intermediate, intermediate, and advanced EFL learners. The researchers required participants to compose argumentative and narrative writing tasks under time pressure and found that advanced writers focused more on the quality of ideas and coherence whilst pre-intermediate learners dealt mostly with linguistic matters. These findings resonate well with Schoonen et al. and's (2009, p. 86) assertion that below

a certain L2 threshold, "writers can become so absorbed in linguistic processing, that is, in searching for the right words, the right sentence structures and the right spelling, that they have little eye for conceptual processing, that is, for the global content and structure of the text" (see Ortega, 2012; Roca de Larios et al., 2016 for a similar argument). This may mean that when composing complex texts, lower-proficiency learners may process lexical items more elaborately than more advanced writers because they focus mostly, if not solely, on linguistic processing, rather than on content elaboration. Therefore, this enhanced linguistic processing may result in higher levels of vocabulary learning.

A different way proficiency may affect lexical acquisition through writing may be connected to how learners compose complex texts, namely, to the recursive process of planning, formulating, and reviewing. As suggested by Flower and Hayes model of L1 writing (1981; see Sect. 3.1), the ideas generated during the planning process of formal writing may be devised in two principal ways. Writer's ideas may be generated as already well-structured and well developed, and then they require little subsequent organization. Alternatively, they may be messy, disconnected, fragmented, and incoherent, therefore necessitating more organization and significant textual restructuring. Flower and Hayes (1981) have asserted that often only more experienced L1 writers, typically more proficient writers, may be able to generate well-organized ideas. Conversely, the fragmented ideas devised by inexperienced L1 writers will need a significant amount of organization, which will depend on and will lead to further goal setting, idea generation, and organization. All this extra restructuring is likely enhance linguistic processing and as a result to facilitate linguistic learning (e.g., the learning of novel words).

The study conducted by Manchón et al. (2009; see above) with L2 writers appears to reinforce Flower and Hayes's (1981) assertion. The results demonstrated that higher-proficiency learners (also experienced writers) spent more time planning and less time writing (i.e., formulating) than lower-proficiency participants. Manchón and Roca de Larios (2007) came to a similar conclusion: pre-intermediate, intermediate, and advanced participants allocated around 2%, 7%, and 10%, respectively, of the total writing time to planning. A reasonable conclusion, then, is that higher-proficiency L2 writers generate well-structured ideas during planning, as suggested by Flower and Hayes (1981), also because they spend more time planning. By the same token, these higher-level learners need not allocate so much time to formulation (Manchón et al., 2009) because the ideas are already well-structured during planning.

However, how do these findings affect lexical learning? Let us assume that writers are required to incorporate in their texts some keywords that are provided in a glossary (as is the case in this book), or whose meaning they have to check in a dictionary. Learners of different L2 proficiency levels will process these keywords differently. Higher-proficiency learners may process these keywords during the planning stage more than lower-proficiency writers. However, lower-proficiency L2 learners will focus on the formulation stage of writing more than their higher-level counterparts (Manchón et al., 2009), therefore paying more attention to the *use* of these keywords. This will be done mentally and mechanically (i.e., the writing per se), quantitatively

and qualitatively, and such increase in output processing may result in further lexical involvement, thus enhancing vocabulary learning (see Sect. 2.3). For instance, during formulation, lower-proficiency learners may need to consult the dictionary or the glossary more often than higher-proficiency writers just to ensure that the reorganized ideas and re-set goals can incorporate the lexical items properly. It is also possible, and perhaps likely, that these learners will need to re-write sentences with keywords more often than more experienced writers, further enhancing lexical processing. This may be particularly true when composing argumentative essays, as in the studies reported in Chaps. 6 and 7, which elicit a high level of processing (Ruiz-Funes, 2014, 2015).

In short, L2 proficiency may affect the lexical learning potential of formal writing in at least two ways. First, lower-proficiency L2 writers are more easily overwhelmed by the cognitive demands of complex writing. Because they are unable to focus on the formal and conceptual demands of the writing task simultaneously, these learners will typically opt to allocate more attention to form, including lexical use. This enhanced attention to form—added to the recursivity of the writing process, which also increases word processing—may result in higher levels of lexical acquisition. Second, lower-proficiency learners tend to spend more time in the formulation stage of the writing process than higher-level learners. This writing and re-writing with keywords may enhance lexical processing and thus boost lexical learning.

3.3.2 The Influence of Multitasking in Lexical Learning Through Writing

Learners' cognitive resources may be further stressed if they are required to perform a secondary task while composing a text. As mentioned above, let us consider a situation where, for learning or research purposes, students are provided with a glossary with keywords that they need to incorporate in their writing. The *secondary task* of using a glossary is performed simultaneously with writing, which is the *primary task* (as in two of the studies reported in Chaps. 6 and 7). The rationale behind using glossaries containing keywords is as follows. When writing in the L2, learners use the words they know. If they do not know a word in the L2, but they intend to use that word in writing, they are likely to consult an outside source, e.g., a dictionary. Thus, by providing students with glossaries in a classroom task or a research task, I simulate monolingual dictionary use while still controlling for the number and quality of keywords to be learnt by students within a given writing task. Providing learners with glossaries may take place in the classroom, but this will also add to the complexity of performing the primary task, i.e., writing a coherent text, and may further strain learners' cognitive recourses.

One obvious reason for this increase in cognitive stress is the necessary shift between tasks (i.e., text composition and keyword use), which warrants extra time and effort and may easily place an excessive burden on one's limited cognitive resources (Kellogg, 1990; Kellogg et al., 2013; Olive, 2004, 2011; see also Skehan, 2009, 2014).

This is because when performing these two tasks simultaneously, learners need to shift their attentional resources in at least a few ways. They will need to focus on (a) the many intricacies of textual composition while (b) consulting the glossary to (c) understand the words meanings and use and then (d) retain the information in memory long enough to (e) devise ways to properly use these keywords in the text.

Additionally, multitasking may be even more cognitively demanding when the tasks necessitate similar codes of processing (i.e., spatial or verbal/linguistic) and similar modalities (auditory or visual), hence competing for resources (Wickens, 1981, 2008). Often, psychologists measure the WM memory demands of a task by asking participants to perform two tasks simultaneously, a primary (e.g., writing) and a secondary one, typically a task tapping into verbal, spatial, auditory, or visual WM resources. This dual-task technique assumes that performance in the primary task will hamper performance in the secondary task (or both) provided that both tasks draw from a similar pool of WM resources. For instance, if learners perform a primary task concurrently with a secondary task that relies mainly on visual WM, and performance in this secondary task suffers, it is believed that the primary task also utilizes visual WM resources. This appears to be the case with writing and keyword use, as laid out below.

Research of this type has demonstrated that writing requires mainly visual and verbal WM (e.g., Kellogg et al., 2007; Olive & Passerault, 2012; Olive et al., 2008). In two experiments, Olive et al. (2008) subjected 132 speakers of French to a dual-task experiment whereby participants were required to write a 30-min argumentative essay while performing secondary tasks that placed a burden on different components of participants' WM. The researchers utilized secondary tasks designed to tap into writers' verbal, visual, and spatial WM. The primary and secondary tasks were performed first in isolation (the control), then concurrently (the dual-task condition), and their results were compared. Each secondary task required participants to decide whether stimuli that was visually presented on a screen matched the stimuli presented 15–45 s earlier. In other words, participants needed to keep the most recent information in memory while trying to decide whether the stimulus was a match. Participants were instructed to answer as quickly as possible, and their response times (RTs) were measured in milliseconds. Longer RTs, that is, slower responses, as well as inaccurate responses, were then interpreted as signs of cognitive interference.

The results showed that writing fluency was impaired in the verbal, visual, and spatial conditions, but mostly in the verbal condition. Accuracy in the secondary task dropped significantly in the three dual-task conditions, relative to the control, thus indicating cognitive interference from multitasking; however, this drop was significantly higher in the verbal and visual conditions. Similarly, RTs were longer in the dual-task conditions than in the control, with responses to the spatial stimuli being faster than to the verbal and visual stimuli, which suggests a higher cognitive interference from the verbal and visual conditions. Olive et al. (2008) concluded that writing argumentative essays relied mostly upon the verbal and visual, not upon the spatial, WM. As a result, when learners are composing such types of texts and the secondary task is also a verbal and visual task—as is the use of pre-specified keywords provided in a glossary—performance on either or both tasks may suffer. In this case,

even higher-proficiency learners, usually better at multitasking, may struggle to write argumentative essays with keywords, especially under time pressure.

So far, this chapter has overall drawn conclusions based on theories on L1 writing (although see Stevenson et al., 2006 inhibition hypothesis presented in Sect. 3.2.1). Considering mostly Flower and Hayes' (1981) L1 writing model and Kellogg's (1990), Kellogg et al. (2013) overload hypothesis, this chapter has explained that writing is a highly cognitively demanding process that may be less or more conducive to lexical learning depending on writers' proficiency. It has also focused on L1 research on multitasking to highlight that the need to incorporate pre-specified keywords during composition may stress attentional resources even further. By contrast, L2 writing researchers have turned to task-based SLA hypotheses to account for an increase in cognitive load and explain task-performance, possibly paving the way to a better understanding of the lexical learning potential of writing tasks. The next section will outline two of these hypotheses: Skehan's (1998, 2003) *trade-off* or *limited capacity hypothesis* and Robinson's (2001, 2005) *cognition hypothesis*.

3.4 Task Characteristics and Their Effect on Task Performance and Lexical Learning

3.4.1 The Notion of Complexity and Complexity Measures

The performance of a task has often been measured based on its *complexity*, *accuracy*, and *fluency*. According to Skehan (2009), complexity refers to how advanced the language produced is. Accuracy is related to learners' ability to avoid errors, and fluency concerns the capacity to produce language at a normal pace, without interruptions. Complexity may be related to syntactic complexity, for example, mean number of clauses per t-unit, mean length of clause, complex nominals per clause (see Biber et al., 2004; Norris & Ortega, 2009 for more comprehensive lists). Of note, a t-unit is defined as "one main clause plus any subordinate clause or non-clausal structure that it is attached to or is embedded within it" (Mazgutova & Kormos, 2015, p. 6). Thus, the number of clauses per t-unit is considered a measure of subordination (Ortega, 2015). Complexity may also refer to lexical complexity, and its measures will be used in this book (see Chaps. 4, 6 and 7). Although it is debatable whether lexical complexity should be included within complexity or whether lexical and structural complexity should be considered separately (Skehan, 2009), here, following Skehan (2009), lexical complexity will be considered part of the overarching group of complexity.

As for measures of lexical complexity (e.g., Daller et al., 2003; Durán et al., 2004; McCarthy & Jarvis, 2007), research often differentiates between lexical variation (also called lexical richness or lexical diversity) and lexical sophistication measures. Lexical variation is usually measured through some sort of type/token ratio. That is, the number of individual word forms divided by the total number of words.

Nevertheless, longer texts are likely to yield lower lexical variation scores, so some correction must be applied to correct for bias in text length. Currently, it appears that the measures D (Malvern & Richards, 2002) and MTLD (McCarthy & Jarvis, 2010) are the most reliable (e.g., Crossley et al., 2015; Fergadiotis et al., 2015). Lexical sophistication focuses on word difficulty, that is, how rare the words used in production are in the language (Crossley et al., 2013; Daller et al., 2003; Vermeer, 2000). Lexical variation and lexical sophistication measures have been used in the literature, for instance, to successfully differentiate between proficiency levels in spoken (Crossley et al., 2011a) and written production (Crossley et al., 2011b, 2011c, 2015). For this reason, in this book these measures will be used to ascertain similar proficiency between the treatment conditions (see Chaps. 6 and 7).

3.4.2 The Trade-Off or Limited Capacity Hypothesis

Obviously, task complexity will affect task performance, and possibly lexical learning. Skehan's (1998, 2003, 2014) trade-off or limited capacity hypothesis posits, as the name suggests, that attention and working memory are limited and that more cognitively demanding L2 tasks will necessitate the employment of more attentional resources. This is very similar to the L1 theories discussed in Sect. 3.2, but this hypothesis focuses on how different task characteristics affect task performance. The limited capacity hypothesis suggests that changes in task characteristics (e.g., pre-task planning, task repetition, or task familiarity) will affect the complexity, accuracy, and fluency of production. More specifically, there is a tension between form (i.e., complexity and accuracy) and fluency; also, within form, complexity and accuracy will compete for resources. As stated by Skehan (2009, p. 511), "one could portray these tensions in the form of a Trade-off Hypothesis, which would predict that committing attention to one area [e.g., complexity], other things being equal, might cause lower performance in others [e.g., accuracy and fluency]". This could be the case with the writing tasks explored in this book. Composing argumentative essays without using pre-specified keywords should be less complex than writing these essays with keywords provided in glossaries. Therefore, essays without pre-specified keywords should be written faster (i.e., higher fluency) and more accurately than essays where specific words must be used (see Chap. 4 for an overview of the second and third studies reported in this book). This being the case, obligatory keyword use may be interpreted as a factor that increases task complexity, therefore demanding more attentional resources from learners. This is in line with what has been explained about multitasking in Sect. 3.2.2 above.

At times, measures of lexical variation have been employed to investigate the tenets of Skehan's (1998, 2009) limited capacity hypothesis. One relevant finding is that when L2 learners attempt to make more complex lexical choices, accuracy and syntactic complexity suffer (Skehan, 2009). Again, the need to incorporate novel lexical items in writing (thus increasing complexity) may reduce accuracy,

as explained above. Furthermore, it appears that the provision of pre-task and in-task planning time (i.e., the planning that occurs during formulating and reviewing; see Sect. 3.1) allows native speakers and L2 learners to use less frequent words in production (i.e., more advanced vocabulary; Skehan, 2009, 2014). Indeed, Skehan (2003) has demonstrated that giving learners time for planning improved fluency, accuracy, and complexity, including lexical complexity (Foster & Skehan, 1996; Skehan, 2003). Similarly, Kellogg (1990) has found that the availability of planning time prior to writing frees attentional resources and thus improves the quality of written compositions (see also Sect. 3.3.3 below). In the words of Skehan (2014, p. 237), "easing tasks and conditions [by, for example, providing planning time,] could create space for attention to focus on form and, if not pushed to handle greater complexity, then to achieve higher levels of accuracy".

3.4.3 The Cognition Hypothesis

The effect of planning time and other task characteristics on performance have also been investigated in light of Robinson's (2001, 2005, 2007, 2011) cognition hypothesis. Differently from Skehan's (2003) trade-off hypothesis, which considers attention and WM limited, the cognition hypothesis views attention as consisting of multiple resources. As such, during task performance, learners may draw on different pools of attentional resources simultaneously. The hypothesis states that as long as the task does not demand attentional resources from a similar pool, task performance in not compromised. The cognition hypothesis describes task complexity along two dimensions (see Table 3.1).

Manipulating task characteristics along the *resource-directing* dimension directs L2 learners' attention to features of the language necessary to deal with the complexities of a task, therefore increasing complexity and accuracy, but decreasing fluency. On the other hand, increasing task complexity along the *resource-dispersing* dimension, which diverts learners' attention from language production, will negatively affect complexity, accuracy, and fluency. According to Robinson (2001), this would

Table 3.1 Resource-directing and resource-dispersing features of cognitive task complexity (Johnson, 2017, p. 14, referring to ideas presented by Robinson, 2011)

Resource-directing features	Resource-dispersing features
± Here-and-now	± Planning time
± Few elements	± Prior knowledge
± Spatial reasoning	± Single task
± Causal reasoning	± Task structure
± Intentional reasoning	± Few steps
± Perspective taking	± Interdependency of steps

happen, for example, when a secondary task (e.g., keyword use) is added to the primary task or when no planning time is available.

At least two findings from research drawing on the cognition hypothesis are relevant to the current discussion. First, Johnson (2017) conducted a literature review and meta-analytic study and found that generally, an increase in reasoning demands (a resource-directing task feature) decreased accuracy in production. This runs counter to the predictions of the cognition hypothesis (Robinson, 2011), whereby an increase in resource-directing features should improve accuracy and complexity. Still, as Johnson (2017) explains, defining resource-directing and resource-dispersing features is far from clear cut. Thus, it may be that the way studies have proceduralised reasoning demands may have indeed tapped into the "± single task" resource-dispersion feature, in which case a decrease in accuracy and complexity is expected. For instance, Frear and Bitchener (2015) study with 34 English learners in New Zealand manipulated reasoning demands by investigating performance in three tasks, from lower to higher complexity. The first (low complexity) task required participants to write a letter to a friend who was coming to New Zealand to tell him/her about the country. The second task, the medium complexity task, asked participants to (1) write the same letter and (2) recommend two restaurants where they could eat together. The third and more complex task required learners to (1) write the same letter, (2) recommend three restaurants, and (3) include the information of two more friends who were supposed to join them in the restaurants. Clearly, it is also possible to interpret these tasks' increase in complexity as an increase in multitasking (i.e., ± single task): one task at first (i.e., write a letter), then two tasks (writer a letter and recommend two restaurants), then three tasks (write a letter, recommend three restaurants, and describe two friends). Indeed, Frear and Bitchener (2015) found that the tasks of higher complexity resulted in a decrease in syntactic complexity, which is in line with the predictions of resource-dispersing features. Again, evidence suggests that multitasking decreases task performance, which supports Skehan's (2003) trade-off hypothesis and findings from dual-task research in L1 (see Sect. 3.2.2 for more information).

Studies drawing on the cognition hypothesis have also investigated the effects of providing planning time on task performance. The results have found that allowing writers to plan generally leads to an increase in syntactic complexity, lexical complexity, and accuracy (Johnson, 2017). For example, Ellis and Yuan (2004) conducted a study with 42 Chinese speakers of L2 English and found that the provision of unpressured in-task planning improved accuracy in production. These findings are in clear support for Robinson's (2011) cognition hypothesis, and as explained above, Skehan's (2003) trade-off hypothesis and Kellogg's (1990) overload hypothesis. Finally, Johnson (2020) has put together a very convenient "timeline" review of literature on the role of planning in task performance. There, the author lists in chronological order important studies in the field, each with a quick summary of the main findings.

3.5 The Various Factors at Play: Connecting SLA and L2 Writing Research

Chapters 2 and 3 have provided the necessary background for Studies 2 and 3 reported in Chaps. 6 and 7. Chapter 2 focused on incidental lexical learning, especially through writing, and the involvement load hypothesis (ILH). This chapter has emphasized L2 writing research and the possibility of lexical learning trough writing, especially when students are provided with keywords to be included in their texts. I have shown that learner- and task-related factors may influence vocabulary acquisition via writing tasks, although research is yet to measure such learning following formal writing tasks.

As explained above, high-proficiency participants, focusing on the primary task, may be able to generate well organized ideas during planning (Flower & Hayes, 1981), therefore reducing the need for subsequent organization, goal setting, formulating, and reviewing. This would likely decrease processing of the keywords, which, together with learners' preference for treating the writing proper as the primary task, may reduce lexical learning. If this happens, one possibility is that timed essays (i.e., writing with a time limit), despite the recursivity of the writing process, may generate the same amount of vocabulary learning as, for example, writing sentences with keywords. Such finding would corroborate the predictions of the ILH (Laufer & Hulstijn, 2001). It is also possible that the complexities inherent in the writing of timed essays may be highly cognitively demanding, with the result that lexical learning will be even lower than that yielded by sentence writing. Eventually, performance in and the lexical learning potential of formal writing tasks may be predicated on whether learners treat either the writing or keyword use as their primary task, thus allocating more attentional resources to one or the other. The investigation of this is attempted in Study 2 reported in Chap. 6.

As further discussed in this chapter, the provision of in-task and/or pre-task planning time frees up attentional resources, hence improving task-performance (e.g., Kellogg, 1990; Robinson, 2011; Skehan, 2009). These findings may have direct consequences to Study 3, reported in Chap. 7. This study compares task performance and lexical learning following the writing of timed and untimed essays (i.e., writing without a time limit). In neither task are students obliged to plan their essay prior to writing, but it stands to reason that the untimed task will allow for more pre-task and in-task planning. One possibility is that the untimed essays will free attentional resources (e.g., Kellogg, 1990) and give participants the opportunity to focus on form, including lexical items, which should enhance vocabulary learning without undermining accuracy. Thus, it is possible that untimed essays may be more accurate and yield more lexical learning than timed essays. Another possibility is that participants writing without time constraints will decide to allocate their freer cognitive resources to the development of the primary task (i.e., the writing proper), not to the secondary task (i.e., keyword use). This is typically the case among higher-proficiency L2 writers (e.g., Ruiz-Funes, 2015; see also Sect. 3.2.1), such as the ones investigated in Studies 2 and 3. In this case, participants may not allocate extra

attention to the keywords (the secondary task) and may feel pushed to use the time available to improve text complexity, thus decreasing learning and possibly accuracy (Skehan, 2014).

Unfortunately, the complex interplay of learner-related factors (e.g., proficiency, WM) and task-related factors (e.g., multitasking, planning time) discussed in this chapter makes accurate predictions about the interaction between task performance and lexical learning almost impossible. Existing research provides very few clues. SLA and psycholinguistic studies (see Sects. 2.3 and 2.5) measuring learning through writing (e.g., Elgort et al., 2018; Pichette et al., 2012) focused on the writing of words, sentences, or very short texts, or they disregarded the cognitive processes underlying the composition of more complex texts, and the effects these processes may have on language learning. On the other hand, L2 writing research has focused on the writing process and on task performance whilst ignoring the vocabulary learning potential of writing (Manchón & Williams, 2016). For example, researchers have investigated the effects of different levels of task complexity on the production of written output (e.g., Ruiz-Funes, 2015). Other L2 writing studies have measured the amount of attention allocated to different writing stages (e.g., Gánem-Gutierrez & Gilmore, 2018; Manchón & Roca de Larios, 2007; Schoonen et al., 2009). And yet other studies have focused on the relationship between the allocation of attentional resources in writing and an increase in cognitive load (Kellogg et al., 2013; Kormos & Trebits, 2012; Stevenson et al., 2006). In other words, it appears that the fields of SLA and L2 writing have so far lacked an interdisciplinary dialogue (Ortega, 2012).

This book attempts to address this issue by bringing together measures that are common to both fields. The studies will utilize tests and techniques often used in SLA and related fields to measure the amount of lexical learning yielded by different writing tasks (i.e., sentence writing, timed essays, and untimed essays). The studies will also employ textual measures that are typical of L2 writing research to assess writing performance, namely measures of lexical complexity, accuracy, and fluency. This interdisciplinary dialogue is expected to shed further light on the lexical learning potential of writing tasks. Because the entire research project presented here is rather complex, the next chapter will outline the studies reported in the book.

3.6 Conclusion

This chapter has focused mainly on L1 and L2 writing research that investigates the writing process and its demands. The chapter has explained that formal writing is a complex, cyclical process that forces writers to allocate many attentional resources to the task, at times to the point that writers may be cognitively overloaded. Less experienced writers, for instance, typically less proficient writers, are often overwhelmed by the cognitive demands of writing and may be forced to focus on formal aspects of the text (e.g., vocabulary use) to the detriment of contextual aspects (e.g., coherence). Then, the chapter discussed how multitasking when writing may affect writing performance and any learning that may be yielded by the writing task.

Regarding lexical learning, two hypotheses are relevant to Studies 2 and 3 (Chaps. 6 and 7) reported in this book. On the one hand, the recursivity of formal writing may enhance processing of the keywords, resulting in high levels of vocabulary acquisition; on the other, the cognitive demands of writing, especially when multitasking (e.g., having to incorporate keywords in the text), may overload participants, hence reducing learning. Another possibility, addressed in Study 3, is that the provision of extra planning time during the writing task may free attentional resources, thus increasing lexical learning.

References

Alptekin, C., Özemir, O., & Erçetin, G. (2014). Effects of variations in reading span task design on the relationship between working memory capacity and second language reading. *The Modern Language Journal, 9*(2), 536–552.

Biber, D., Conrad, S., & Cortes, S. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Biedroń, A., & Szczepaniak, A. (2012). Working memory and short-term memory abilities in accomplished bilinguals. *The Modern Language Journal, 96*(2), 290–306. https://doi.org/10.1111/j.1540-4781.2012.01332.x

Bohn-Gettler, C. M., & Kendeou, P. (2014). The interplay of reader goals, working memory, and text structure during reading. *Contemporary Educational Psychology, 39*, 206–219. https://doi.org/10.1016/j.cedpsych.2014.05.003

Byrnes, H., & Manchón, R. M. (2014). Task, task performance and writing development: Advancing the constructs and the research agenda. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 267–299). John Benjamins.

Casanave, C. P. (2016). Qualitative inquiry in L2 writing. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 497–541). De Gruyter.

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System, 41*, 965–981.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). 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. (2011c). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243–263.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*(1), 182–193.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*(4), 561–580.

Cumming, A. (1990). Metalinguistic and ideational thinking in second language composing. *Written Communication, 7*(4), 482–511.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197–222.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242.

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

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84. 10+10170S0272263104261034

Fergadiotis, G., Wright, H. F., & Green, S. B. (2015). Psychometric evaluation of lexical diversity indices: Assessing length effects. *Journal of Speech, Language, and Hearing Research, 58*, 840–852.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299–323.

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing, 30*, 45–57.

Gánem-Gutierrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning, 68*(2), 469–506.

Hayes, J. R. (2012). Modelling and remodelling writing. *Written Communication, 29*(3), 369–388.

Hayes, J. R., & Flower, L. S. (1980). *Identifying the organization of writing processes.* Retrieved from: https://www.researchgate.net/publication/200772468_Identifying_the_organization_of_writing_processes. Accessed September 6, 2020.

Hyland, K. (2016). *Teaching and researching writing* (3rd ed.). Routledge.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing, 37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

Johnson, M. D. (2020). Planning in L1 and L2 writing: Working memory, process, and product. *Language Teaching, 53*, 433–445.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44*(2), 137–166. https://doi.org/10.1017/S0261444810000509

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands. *The American Journal of Psychology, 103*(3), 327–342.

Kellogg, R. T., Olive, T., & Piolat, A. (2007). Verbal, visual, and spatial working memory in written language production. *Acta Psychologica, 124*, 382–397.

Kellogg, R. T., Whiteford, A. P., Turner, C. E., Cahil, M., & Mertens, A. (2013). Working memory in written composition: An evaluation of the 1996 model. *Journal of Writing Research, 5*(2), 159–190.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1–18. https://doi.org/10.3389/fpsyg.2017.01997

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning, 62*(2), 439–472.

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

Lee, I. (2016). EFL writing in schools. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 113–139). De Gruyter.

Lee, J. (2019). Time-on-task as a measure of cognitive load in TBLT. *The Journal of Asia TEFL, 16*(3), 958–969. https://doi.org/10.18823/asiatefl.2019.16.3.12.958

Linderholm, L., & Van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology, 94*(4), 778–784. https://doi.org/10.1037//0022-0663.94.4.778

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Manchón, R. M. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 27–52). John Benjamins.

Manchón, R. M. (2016). Quantitative enquiry in L2 writing. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 519–540). De Gruyter.

Manchón, R. M., & Roca de Larios, J. (2007). On the temporal nature of planning in L1 and L2 composing. *Language Learning, 57*(4), 549–593.

Manchón, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal and problem-solving nature of foreign language composing processes: Implications for theory. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 102–129). Multilingual Matters.

Manchón, R. M., & Williams, J. (2016). Introduction: SLA-L2 writing interfaces in historical perspective. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 567–586). De Gruyter.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English academic purposes programme. *Journal of Second Language Writing, 29*, 3–15.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488.

McCarthy, P. 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.

Norris, J. M., & Ortega, L. (2009). Towards and organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Olive, T. (2004). Working memory in writing: Empirical evidence from the dual-ask technique. *European Psychologist, 9*(1), 32–42.

Olive, T. (2011). Working memory in writing. In V. W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 485–504). Psychology Press.

Olive, T., Kellogg, R. T., & Piolat, A. (2008). Verbal, visual, and spatial working memory demands during text composition. *Applied Psycholinguistics, 29*, 669–687.

Olive, T., & Passerault, J. M. (2012). The visuospatial dimension of writing. *Written Communication, 29*, 326–344.

Ortega, L. (2012). Epilogue: Exploring L2 writing-SLA interfaces. *Journal of Second Language Writing, 21*, 404–415.

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing, 29*, 82–94.

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

Paas, F., & Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 51–71.

Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and sentence writing for second language vocabulary acquisition. *Applied Linguistics, 33*(1), 66–82.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*, 1–32.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*, 193–213.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–37). John Benjamins. https://doi.org/10.1075/tblt.2

Roca de Larios, J., Nicolás-Conesa, F., & Coyle, Y. (2016). Focus on writers: Processes and strategies. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 267–286). De Gruyter.

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163–191). John Benjamins.

Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing, 28*, 1–19.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77–101). Multilingual Matters.

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

Skehan, P. (2003). A cognitive approach to language learning (2nd ed.). Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510–532.

Skehan, P. (2014). Limited attentional capacity, second language performance and task-based pedagogy. In S. Peter (Ed.), *Processing perspectives on task performance* (pp. 211–260). John Benjamins.

Speciale, G., Ellis, N. C., & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics, 25*, 293–321. https://doi.org/10.1017/S0142716404001146

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multidimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing, 15*, 201–233.

Varol, B., & Erçetin, G. (2016). Effects of working memory and gloss type on L2 text comprehension and incidental vocabulary learning in computer-based reading. *Procedia: Social and Behavioural Sciences, 232*, 759–768. https://doi.org/10.1016/j.sbspro.2016.10.103

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83.

Wickens, C. D. (1981). *Processing resources in attention, dual task performance, and workload assessment*. Technical report. Engineering-psychology Research Laboratory. University of Illinois.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 449–455.

Chapter 4
Overview of the Research Project: Methodology and Statistical Analyses

4.1 Introduction

This book has highlighted the importance of lexical knowledge to attain higher levels of proficiency in an L2, particularly English, and to succeed at English-medium universities. Hopefully, it is clear by now that achieving desirable levels of vocabulary knowledge is as daunting a task for learners as it is essential. The previous chapters have explained that university students may need help to improve their knowledge of academic vocabulary, and that such help may come in the form of a more explicit attention to the teaching and learning of these words. One option, this I have argued, is to embed academic words in writing tasks, including argumentative essays. Students would then perform tasks that are typically used to assess their knowledge of content (the essays) to improve their vocabulary knowledge, thus effectively integrating content and language. This chapter will first briefly review the content of the previous three (theoretical) chapters and then quickly present an outline of the three studies reported in this book. The readers may thus use this chapter as a quick reference guide for the theoretical (Chaps. 1–3) and empirical (Chaps. 5–8) parts of the book.

4.2 Theoretical Assumptions Behind the Studies

In Chap. 1, I showed that learners need to know thousands of word families to succeed in academia. I explained that these words must be known in breadth and in depth, receptively and productively, and not only individual units, but also multiword items such as phrasal verbs and other fixed phrases. Still, I argued that this process of lexical acquisition may be facilitated. Having mastered the essential most common 2000 families in English, university students could focus on mastering academic words, rather than low-frequency or technical vocabulary. Nevertheless, academic words are usually unlikely to be learnt incidentally through exposure alone, even after

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_4

years studying at universities in English-speaking countries. This is because they are morphologically complex, abstract, and they lack salience in discourse. Due to this difficulty in academic vocabulary learning, it is often necessary to assess university learners' knowledge of academic words at the commencement of their studies to decide, for instance, if students need remedial language lessons. The problem, I argued, is that the presence of cognates (often words of Latin/Greek origin) in tests distorts results, and researchers do not agree on a solution to this problem.

Chapter 2 discussed incidental vocabulary learning through reading and writing, and the involvement load hypothesis (ILH). The first section explored some issues behind the definition of incidental learning to provide theoretical support for the definition adopted in this book. As a reminder, I consider learning to be incidental when participants perform a primary task involving the processing of some information (i.e., the academic keywords) without being aware of the true purpose of the experiment and without being told in advance that they will be tested afterwards on their recall of that information. Sections 2.2 and 2.3 briefly discussed incidental learning through reading and writing, respectively. These sections emphasized the need to promote incidental learning of academic words through writing (i.e., producing output) because it is more effective than learning through reading (written input). Chapter 2 then introduced the ILH and discussed in depth four ILH studies in which students were presented with keywords to include in their writing. The four studies compared the effectiveness of writing sentences (SW) and compositions (CW) with keywords for lexical learning. These tasks will also be explored in this book. The chapter concluded by stating that when writing compositions with keywords, learners may focus both on content (i.e., the message and structure) and language or mostly on language (e.g., keyword use), thus enhancing learning. Still, to test this hypothesis, it is necessary to draw on L2 writing research, which was the topic of Chap. 3.

The main purpose of Chap. 3 was to discuss the stages of formal writing, the cognitive processes underlying writing, and how these processes may facilitate or impede incidental lexical learning. We saw that formal writing is a cyclical process made up of co-dependent operations that enhance linguistic processing while at the same time overwhelming L2 writers' attentional resources. The chapter explained that L2 writers' individual differences (such as their linguistic proficiency and working memory), the need to multitask while writing (e.g., consulting a glossary and incorporating the keywords in writing), as well as the amount of planning time, may all interact to enhance or hamper incidental lexical learning through complex writing. These issues have been tackled by reviewing both studies into the writing process and SLA research. We argued that, so far, neither of the two lines of research have managed to account convincingly for and to explain the process of learning L2 words while writing in the L2. Whereas L2 writing research appears to have ignored lexical learning, SLA studies on learning through production have failed to consider the writing process and how it modulates learning. Thus, the research reported in this book aims to fill in this important research gap.

4.3 The Studies Reported in Chaps. 5, 6 and 7

The research reported in this book consists of three consecutive studies. Participants of these studies were first- and second-year BA-level students at the Institute of English Studies, University of Warsaw. Their proficiency in English averaged advanced (i.e., B2+) according to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). Almost all data were collected during students' regular writing classes, but the participation in the studies was voluntary: The student had the right to withdraw their data from the research pool.

Study 1, reported in Chap. 5, is based on the research background provided in Chap. 1. Adopting a cross-sectional design, it examined participants' knowledge of academic vocabulary and the usefulness of two tests for assessing their lexical knowledge. Studies 2 and Study 3, reported in Chaps. 6 and 7, are both complex longitudinal quasi-experiments, the theoretical support for which was provided in Chaps. 2 and 3. The following three sections will provide a more detailed overview of each study that may be useful as a quick reminder of the design of the studies. Each of the next three sections will deal with one study separately, and the essential information is presented in tables for ease of reference.

4.3.1 Study 1 (Chap. 5): VST as a Reliable Academic Placement Tool Despite Cognate Inflation Effects (Silva & Otwinowska, 2019)

As explained above, it is crucial to know whether Polish learners of English at the academic level need extra help with academic vocabulary. Hence, this study combines the Vocabulary Size Test (VST; Nation & Beglar, 2007; see Sect. 1.5)—a readily-available, easy to administer meaning-recognition test—and a tailor-made Academic Vocabulary Test (AVT) to identify a threshold in the VST that may be utilized for academic placement purposes with English majors who are native speakers of Polish. This way, I aimed to identify a manner of assessment that is practicable and reliable. The secondary aim of the study was to assess cognate inflation effects in tests (cognate guessing) and avoid them thanks to the statistical methods used, thus circumventing the disagreement among researchers regarding what to do with cognates in tests of proficiency (see Chap. 1). For details on the research methodology see Table 4.1. This study had three research questions:

- RQ1. Do undergraduate English majors at a Polish university achieve similar scores on cognates and noncognates as measured by an academic-vocabulary meaning-recall test (AVT) and the VST?
- RQ2. If cognate inflation is detected, is this this effect similar across the 14 levels of the VST?

Table 4.1 Overall structure of Study 1

Topic	Description	
Participants	106 speakers of Polish (61 first-year and 45 s-year learners; 77 females). Proficiency in English B1 or above	
Instruments	VST	• Multiple-choice, receptive general vocabulary lexical test • 14 levels, 10 words per level, totaling 140 items • Score × 100 = total score. For example, 89 items = 8900 (assumed learners' vocabulary size)
	AVT	• Tailor-made receptive academic vocabulary test • Checklist format: learners need to tick the words they know • Three compatible versions. Each version contained 105 noncognates, 35 cognates, and 65 nonwords to control for guessing
Quantitative analyses	VST	• Comparing scores, in all 14 levels, between cognates and noncognates (to answer RQ1). Statistical analyses used: Wilcoxon signed-rank tests • Checking if word frequency in English predicts learning of cognates and noncognates (to answer RQ1). Statistical analysis used: Simple linear regression
	AVT	• Comparing scores of cognates and noncognates (to answer RQ1). Statistical analysis used: Paired-samples t-tests
	Clustering	• Combining scores in VST and AVT to find independent clusters • Goal: find a threshold to divide students into two groups, lower and higher levels (to answer RQ3)
Qualitative analyses	• Analyses of the cognates answered correctly by most participants in the VST to identify word-related and participant-related patterns that may distort scores even further (to answer RQ2) • Goals: (1) verify the effectiveness of the threshold established (To answer RQ3); (2) identify types of cognates that inflate scores considerably in order to avoid these cognates in future tests	

- RQ3. Can the VST predict learners' academic vocabulary knowledge? If so, is there a threshold on the VST that differentiates learners with regards to their academic vocabulary knowledge?

4.3.2 Study 2 (Chap. 6): Incidental Learning of Academic Words Through Writing Sentences and Timed Essays: Can an Increase in Cognitive Load Affect Acquisition? (Silva et al., 2021)

In this study, I compared incidental lexical learning of academic words following writing sentences (SW) and timed compositions (Timed CW). Since the learning following CW may be affected by time pressure, multitasking, and learners' proficiency, these variables were taken into consideration when designing the study.

I hypothesized that the multitasking inherent in the timed CW task (i.e., writing argumentative essays under time pressure with keywords provided in glossaries) may overwhelm learners' cognitive resources. To detect signs of increased cognitive load, I compared participants' control argumentative essays (i.e., without the need to incorporate keywords) to their treatment essays (with keywords) on textual measures of lexical complexity, accuracy, and fluency (see Chap. 3). Importantly, I only investigated higher proficiency learners. Thus, our participants are similar to Kim's (2008), who found similar lexical gains following CW and SW, but more proficient than Zou's (2017), who found higher gains after CW (see Chap. 2). Consequently, the general aims of Study 2 were to compare learning through Timed CW and SW and to investigate whether an increase in cognitive load in CW may be the culprit for the reduction in the amount of lexical learning yielded by the task (see Table 4.2 for more details). This could equate the learning following CW to the learning following SW, thus corroborating Kim's (2008) findings and the ILH. Considering the above, this study has two research questions:

- RQ1. Do Polish EFL learners acquire and retain academic words to a similar degree after writing sentences and timed argumentative essays?
- RQ2. Does the need to use pre-specified keywords in the CW task affect overall quality, accuracy, and fluency of students' writing as compared to the control essay, thus indicating an increase in cognitive load?

4.3.3 Study 3 (Chap. 7): Incidental Learning of Academic Words Through Writing Sentences, Timed Essays, and Untimed Essays

Here, I compared three different groups: SW, Timed CW, and Untimed CW. For details on the research methodology see Table 4.3. I wanted to find out whether freeing L2 writers from time constraints, thus providing them with plentiful planning and composing time, may free attentional resources and hence enhance lexical learning when writing with keywords (see Chap. 3 for the relevant theoretical background). I utilized similar textual measures to those in Study 2 to detect signs of increased cognitive load. I hypothesized that (1) writing untimed essays with keywords may generate more lexical learning than writing timed essays; or, (2) writing timed and untimed essays with keywords may yield similar learning because writers may decide to use the extra time to focus on content, not form. Put differently, in Study 3 (Chap. 7) untimed essays may be equally or more conducive to learning than timed essays, not less. The research questions follow:

- RQ1. Do Polish EFL learners acquire and retain academic words to a similar degree after writing sentences, timed, and untimed argumentative essays?
- RQ2. Does writing untimed argumentative essays with pre-specified keywords reduce L2 writers' cognitive stress when compared to timed essays?

Table 4.2 Overall structure of Study 2

Topic	Description	
Participants and treatments	39 first-year Polish English majors (30 females). Proficiency in English B1 or above. Timed CW ($n = 17$), SW ($n = 22$)	
Measures of proficiency	Receptive	• LexTALE (Lemhöfer & Broersma, 2012)
	Productive	• Control essay holistic score: overall text quality • D: lexical variation (Malvern & Richards, 2002) • $SUBTLEX_{us}$: word frequency (sophistication; Brysbaert & New, 2009) • Number of errors: normed to account for text size
Measure of cognitive load (CL)	Comparing essays	• Variables: holistic essay scores, normalized errors, words per minute (WPM)
Instruments	Keywords	• 20 academic keywords divided into two compatible sets (Sets A and B) matched for concreteness, frequency, length, and part of speech
	Glossary	• Definition (in English) and 2 examples provided for each keyword
	Tests of learning	• Vocabulary knowledge scale (VKS; Wesche & Paribakht, 1996): measures of breadth and depth of learning • Free association test: measure of depth of learning
	Questionnaire	• To ensure the study was truly incidental
Quantitative analyses	CW versus SW (to answer RQ1.)	• VKS_6 variable: compares learning in SW and timed CW using the full VKS scale. Statistical test used: generalized linear mixed models • VKS_3 variable: same comparison with combined scores. 1 = no knowledge or recognition; 2 = receptive knowledge; 3 = productive knowledge. Statistical test used: generalized linear mixed models • Association scores: compares SW and timed CW in depth of knowledge, only. Statistical test used: generalized linear mixed models
	Task cognitive load (to answer RQ2.)	• Scores, errors, WPM: textual performance measures to detect and compare CL between control, timed, and untimed essays. Statistical test used: paired-samples t-tests

Table 4.3 Overall structure of Study 3

Topic	Description	
Participants and treatments	90 first year Polish English majors (68 females). Proficiency in English B1 or above. SW ($n = 33$), Timed CW ($n = 33$), Untimed CW ($n = 24$)	
Measures of proficiency	Receptive	• LexTALE (Lemhöfer & Broersma, 2012)
	Productive	• Control essay holistic score: overall text quality • D: lexical variation (Malvern & Richards, 2002) • $SUBTLEX_{us}$: word frequency (sophistication; Brysbaert & New, 2009) • Number of errors: normalized to account for text size
Measure of working memory (WM)	Digit span	• Average of forward and backward digit span tasks (Brzeziński et al., 2004) • Goal: to control for the influence of participants' WM on lexical learning
Measure of cognitive load (CL)	Comparing essays	• Holistic essay scores, normalized errors, WPM
	Self-rating scale of CL	• Questionnaire to assess task difficulty, mental effort, and stress
Instruments	Keywords	• 20 academic keywords divided into two compatible sets (Sets A and B) matched for concreteness, frequency, length, and part of speech
	Glossary	• Definition (in English) and 2 examples provided for each keyword
	Tests of learning	• VKS (Wesche & Paribakht, 1996): measures of breadth and depth of learning • Free association test: measure of depth of learning
	Questionnaire	• To ensure the study was truly incidental
Quantitative analyses	CW versus SW (to answer RQ1.)	• VKS_6: compares learning in SW, Timed and Untimed CW using the full VKS scale. Statistical test used: generalized linear mixed models • VKS_3: same comparison with combined scores. 1 = no knowledge or recognition; 2 = receptive knowledge; 3 = productive knowledge. Statistical test used: generalized linear mixed models • Association scores: compares SW, Timed and Untimed CW in depth of knowledge, only. Statistical test used: generalized linear mixed models

(continued)

Table 4.3 (continued)

Topic	Description	
	Task cognitive load (to answer RQ2.)	• Scores, errors, WPM: textual performance measures to detect and compare CL between control, timed, and untimed essays. Statistical tests used: paired-samples t-tests and ANOVAs • Self-rating scale: to compare learner perceived CL between the three tasks. Statistical tests used: one-sample t-tests and ANOVAs

4.4 Some Considerations on Statistical Analyses

This section has two main aims. First, it will provide general yet essential information on many of the statistical analyses used in the three studies reported in Chaps. 5, 6, and 7. Specific aspects of separate analyses will be briefly laid out in their specific chapters. Then, this section will explain the basics of using linear mixed-effects models in order to justify their use in Studies 2 and 3.

4.4.1 Basic Inferential Statistics Used

All statistical analyses in the three studies reported in this book were conducted using IBM SPSS statistics version 26, and the alpha level was set at 0.05. The figures in Study 1 were created in Minitab (Minitab 19 Statistical Software, 2019); the figures in Studies 2 and 3 were created in RStudio (RStudio Team, 2020). The raw data employed in Studies 1 and 2 can be found in online repositories which will be indicated when reporting the studies proper. The raw data for Study 3 is not yet available online, but may be shared upon request.

In all studies, I ran statistical tests typically used to compare differences between independent samples (e.g., scores following CW and SW) and dependent samples (e.g., scores in the pretest and posttest). Some examples include t-tests and analysis of variance (ANOVA). In Study 1, regression analyses and cluster analyses were also conducted. Simple linear regressions were run to verify the predictive value of one variable over another (the outcome/dependent variable). Hierarchical and K-means cluster analyses were run to statistically divide the participants into two separate proficiency groups. I have also reported effect sizes and 95% confidence intervals where appropriate.

In all within-group, between-group, and correlational analyses, parametric tests were preferred to nonparametric tests because they have more power, that is, they are better able to find statistical significance ($p < 0.05$) where one exists (Field, 2017; Howell, 2010; Perry, 2011; Salkind, 2011). However, to use these tests, I needed to

ensure that they met certain assumptions including normality of distribution, homogeneity of variance, and independence of observations. The data were considered normal after inspecting histograms and boxplots, and if the *z* score for skewness was <1.96.[1] Shapiro–Wilk tests are often used to assess normality, but they may be inaccurate (Field, 2017). Levene tests were utilized to assess the assumption of homogeneity of variance, together with data visualization in graphs such as box plots and Q-Q plots. Independence of observations was controlled by design. Other less common assumptions will be discussed separately when reporting the analyses in each study. When the data failed one or more of the assumptions of parametric tests, data transformations were employed (e.g., log, square root). If transformations did not solve the problem, nonparametric tests were used.

Apart from the more typical inferential statistical methods outlined above, Studies 2 and 3 also utilized more reliable, but also more complex statistical methods, called linear mixed-effects models. Because they are still a novelty and are less frequently used in SLA studies than other inferential statistics, the sections below provide a brief introduction to the method.

4.4.2 Linear Mixed Models: An Introduction

Linear mixed models (LMMs) are also known as multilevel models, random coefficient models, mixed-effect models, mixed linear models, or hierarchical linear models (Field, 2017). Mixed-linear models seems to be the term preferred in the psycholinguistic literature (Carson & Beeson, 2013); however, here, I opt for LMM simply because this is what it is called in IBM SPSS, and the *generalized* extension to this model (see below), which I will use, adopts this terminology. LMMs are becoming ever more popular in recent years, especially in the field of psycholinguistics (Meteyard & Davies, 2020). Unfortunately, in applied linguistics and SLA, LMMs are still relatively unpopular (Cunnings & Finlayson, 2015). This subsection aims to briefly explain some key characteristics of LMMs; however, the subject is rather complex, and much disagreement remains (Hajduk, 2019).

Multilevel Data. In SLA, data is typically analyzed at a single level, the exception being repeated-measures designs (e.g., pretest–posttest designs). For instance, study participants (the single level) undergo some type of treatment and the results are measured. Nevertheless, these participants may come from different classes in a school or even from different schools, each with various classes. In this case, the data is said to be multilevel, with participants being *nested* within classes, and classes nested within schools (Carson & Beeson, 2013). Participants are then considered level 1, classes level 2, and schools level 3. It could be that classes are nested within teachers, or schools could be nested within different districts, and even within

[1] To obtain this value in SPSS, it is necessary to divide the skewness statistic by its standard error. For example, skewness of 0.538 and standard error of 0.094 yields a *z* score of 5.72. As it is higher than 1.96, the data fails the assumption of normality (see Mayers, 2013; Trevethan, undated).

different countries. Similarly, in medical sciences, patients may be nested within doctors, who are in turn nested within hospitals and so forth (Hedeker, 2003). In addition to nested effects, LMMs may also include *crossed* effects. As clearly explained by Cunnings and Finlayson (2015, p. 161), "in language research, the subjects sampled are tested on a series of linguistic items, and the same linguistic items are tested across subjects. In this way, subjects and items are crossed at the same level of sampling". This is exactly the case in Studies 2 and 3 reported in this book (see Tables 4.2 and 4.3). My subjects are nested within classes, which are nested within teachers, all in the same school (the Institute of English Studies); also, participants and keywords (linguistic items) are crossed in that different learners are tested on the same linguistic items.

All this is important because nested data is not truly independent, and therefore the assumption of independence of observations is not met (Field, 2017; Hajduk, 2019). It is expected, for example, that responses from the same participant will be correlated across linguistic items, or that answers among participants within the same teacher will be correlated, and the results from different teachers correlated within the same school. It is not difficult to imagine, for example, that learners with the same teacher may behave more similarly than learners with different teachers, or that classes that take place early morning may have different characteristics than classes that take place in the evening, or in different schools or districts. By controlling for these contextual variables (e.g., teachers, classes, schools), the problem with non-independence of observations is overcome (Field, 2017; Hedeker, 2003), and the statistical model will generate more accurate measurements with increased power, leading to more reliable results (Meteyard & Davies, 2020).

Traditional ANOVAs or regression analyses do not account for the variation in scores between participants and items. These tests utilize mean scores per participant per test. As a result, they disregard the variation in scores for each individual test item and the variation between participants' scores, both of which almost always exist and may well influence the results. LMMs, by contrast, compute the data for each individual score for each item in a test. For instance, in Study 2 each participant took two pretests and two posttests. Each test had 10 lexical items (i.e., 10 scores). Consequently, in the LMM analyses, each participant had 40 data points (i.e., 40 rows in the data frame). In an ANOVA, it could be possible to include participants and items (and their interaction) as independent variables; however, this would drastically increase the use of degrees of freedom of the test: one per item, another per participant, and another per each participant-item interaction. This would lead to an enormous (and unacceptable) decrease in power (Field, 2017).

Fixed and Random Effects. LMMs are called "mixed" because they may have *fixed* and *random effects*. Fixed effects are the effects of primary interest to the researcher, which would be used again if the experiment were to be replicated (Seltman, 2018). They are commonly called independent, predictor, or explanatory variables and are used to explain the outcome (dependent) variable (Hajduk, 2019; Heck et al., 2012). The results of fixed effects can only be generalized to situations that are similar to the experiment (Field, 2017). For instance, in Studies 2 and 3 in this book, two

fixed effects are Group (e.g., SW and CW) and Time (i.e., pretest and one-week delayed posttest). The results can only be generalized when comparing the learning of these two groups one week after the experiment with comparable conditions (e.g., participants with similar L1s and similar proficiency in English). Fixed effects may also be included not as predictors, but as variables used to control for certain variance in the data, as long as this variable is continuous (Hajduk, 2019; Singmann & Kellen, 2019). For example, participants' scores in a standardized test (e.g., 1–100 scores) may be entered as a fixed effect to control for the effect of proficiency on the results. When continuous variables are entered as fixed effects, they are called *covariates*, just as they are called in Analysis of Covariance (ANCOVA) tests.

Effects are considered random when they are not of primary interest but are "rather thought of as a random selection from a much larger set of levels" (Seltman, 2018, p. 358). These effects can be generalized beyond the conditions of the experiment as long as they are representative of the whole population (Field, 2017). In my quasi-experiments, Participants and Items (in my case, the keywords), or Classes, would be some random effects. If I find that the different keywords explain some of the variance of the results in the posttest—i.e., that different keywords yield different levels of learning, which is expected—this finding is generalizable to the wider context (provided that our keywords are representative of academic words). If researchers decide to replicate our studies, they will need similar groups and tests (i.e., the fixed effects), but would be able to use their own participants and keywords, although these would need to be comparable. Importantly, only factors (i.e., categorical variables) may be entered to LMMs as random effects (Hajduk, 2019; Singmann & Kellen, 2019). If the variable is continuous, it should be entered as a covariate (see above).

Fitting Linear Mixed Models. There may be a random intercept model, a random slope model, or a random intercept and slope model (Carson & Beeson, 2013; Cunnings & Finlayson, 2015; Hajduk, 2019; Harrison et al., 2018; Meteyard & Davies, 2020; Singmann & Kellen, 2019). These will be briefly introduced below.

A random intercept model for the random effect Participants, for example, takes into account how the average score of each participant varies randomly across the data (Cunnings & Finlayson, 2015). For example, my random intercept model would measure whether the mean scores in the VKS (my dependent variable) vary randomly across Participants. Figure 4.1 depicts an example of such model (the data are similar but not obtained from my studies). One can see that instead of one regression line for the whole dataset, there is one regression line per participant. Each participant has a different starting point (i.e., Time 0), hence a different intercept. A problem is that random intercept models assume that all participants would have the same slope (here, the same rate of learning, from Time 0 to Time 3; see Fig. 4.1), which is highly unlikely.

A random slopes model assumes that intercepts are fixed, but the slopes vary. Typically, intercepts of random effects are assumed to be zero (Harrison et al, 2018). In this book, such model would assume that all participants have similar scores in the pretest but would learn at different rates (as measured by the posttest; see a simulated

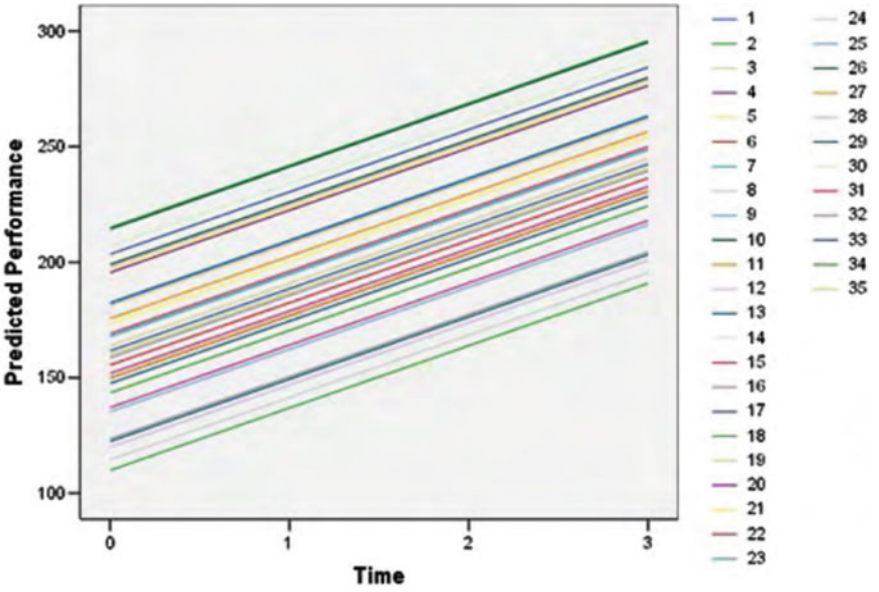

Fig. 4.1 A random intercept model. The right side of the figure depicts the 35 participants in this simulated model. *Source* IBM SPSS (2005, p. 22)

model in Fig. 4.2). A random intercept and slope model is a more realistic model (Carson & Beeson, 2013) and assumes different intercepts and slopes for a random effect. Here, it is likely that intercepts and slopes are correlated, and the model can account for this: It is possible, for example, that participants with higher intercepts (higher scores in the pretest) will learn less vocabulary (lower slope), and vice versa. Figure 4.3 illustrates this with another simulated model.

There is no agreement among researchers on exactly how linear mixed models may be created, or *fit* (Field, 2017; Hajduk, 2019; Meteyard & Davies, 2020). Generally, models may be fit in a "maximal-to-minimal that converges process" or in a "minimal to maximal-that-improves-fit process" (Meteyard & Davies, 2020, p. 17). Put differently, in the maximal-to-minimal model, researchers may prefer to fit all relevant fixed and random effects at once, then eliminate variables, one by one, until the statistical software is able to compute the results (i.e., the model converges). Barr et al. (2013), Hajduk (2019), and Singmann and Kellen (2019) recommend this. Other researchers may prefer to start with the maximal model just as explained above; however, after convergence is achieved, the statistician may keep on simplifying the model to reach *parsimony*: remove non-significant random slopes and intercept-slope correlations to improve fit (see below). Bates et al. (2018) recommend this method. Yet other researchers may choose the minimal-to-maximal modelling strategy, starting from a simple model, either from the fixed or random effects, until obtaining the best fit (e.g., Carson & Beeson, 2013; Cunnings & Finlayson, 2015; Field, 2017; Heck et al., 2012). The quality of fit is usually assessed by comparing

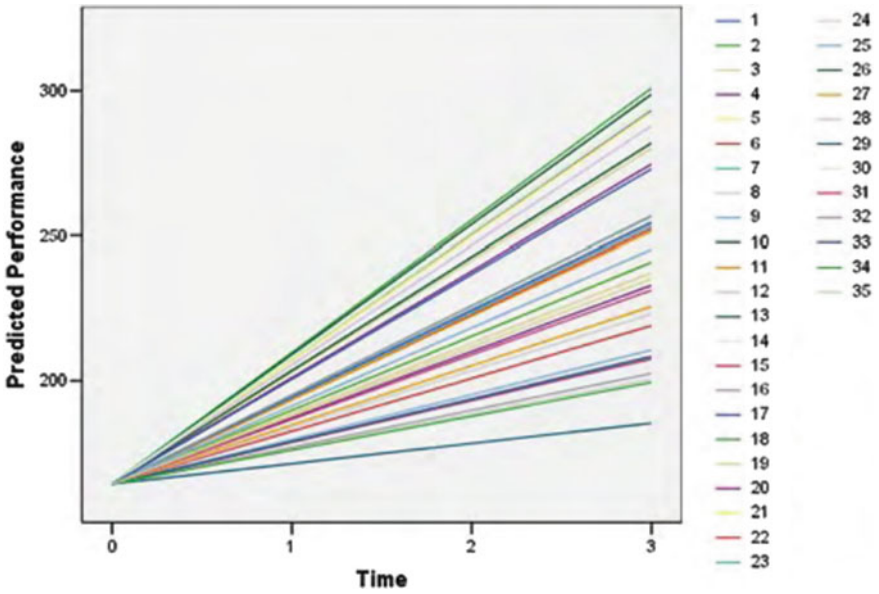

Fig. 4.2 A random slope model. *Source* IBM SPSS (2005, p. 23)

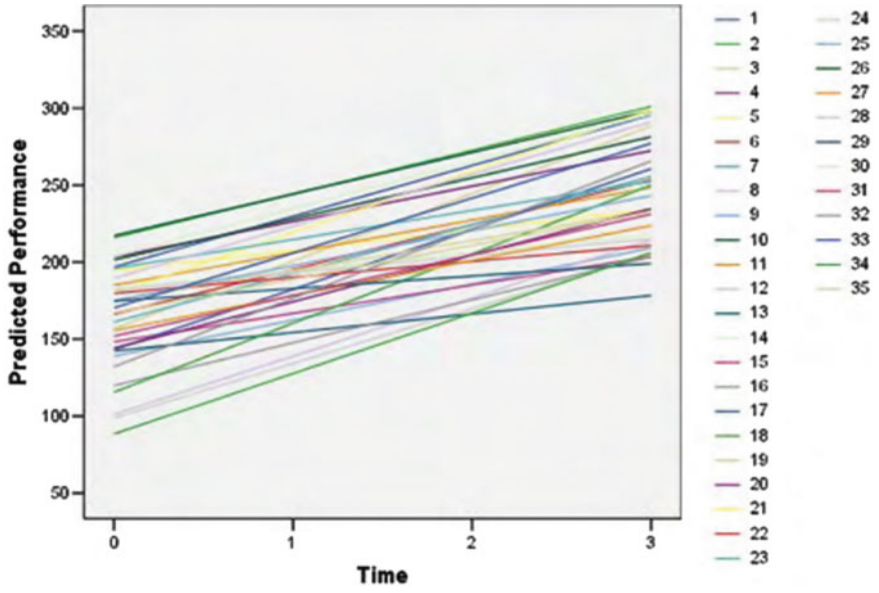

Fig. 4.3 A random intercept and slope model. *Source* IBM SPSS (2005, p. 25)

measures of deviance across models, usually *Akaike's information criterion* (AIC) or the *Schwarz's Bayesian criterion* (BIC), with lower scores meaning better fit models (Field, 2017; Heck et al., 2012).

Since this statistical method is relatively new and it is being constantly developed, in order to create a useful best-practice guide, Meteyard and Davies (2020) interviewed 163 psycholinguistics researchers who are familiar with mixed models and carried out a systematic review of 400 published papers using LMMs. When hypotheses are set, and the study is confirmatory, the researchers recommend following a maximal-to-minimal process using BIC to compare models. When the research is exploratory, the minimal-to-maximal process, with AIC, should be adopted.

4.4.3 Fitting Linear Mixed Models in the Current Book

In this book, Study 2 is exploratory, and thus a minimal-to-maximal approach will be used. Study 3, conversely, seeks to confirm the results from Study 2, so it is confirmatory, while exploring learning through untimed argumentative essay writing (thus exploratory). Despite this mix, I will follow the minimal-to-maximal process recommended for exploratory studies. This is because I cannot conduct ordinary LMMs since my dependent variables are not continuous, an assumption of LMMs. In my case, the variables VKS_6, VKS_3, and Association (see Tables 4.2 and 4.3) are ordinal (VKS) and count (Association) variables. Due to this, I will need to run *generalized* linear mixed models (GLMMs), which can cope with different types of data (Heck et al., 2012; IBM SPSS, 2020). GLMMs are even more computationally intensive than LMMs, and therefore, researchers recommend starting with more basic models (Heck et al., 2012). The consequence of treating both studies as exploratory is that the results will need new data to be confirmed (i.e., more so than usually; see Winter, 2020).

In Studies 2 and 3, when fitting minimal-to-maximal GLMMs, I will mainly follow recommendations set out by Bates et al. (2018) and Meteyard and Davies (2020). I will start with random effects in a stepwise fashion, beginning with intercepts, then slopes, then intercepts and slopes, and finally covariances (i.e., correlations between intercepts and slopes). If certain random effects do not improve fit (i.e., lower the AIC criterion), are deemed redundant, or there are errors in computation (e.g., the model does not converge), they will be eliminated from the model. Once the random effects have been established (i.e., best fit), the fixed effects will be added. Then, to test whether model fit may be improved, I will again start adding random effects that were previously deemed non-redundant in a similar stepwise fashion as above. The final models (i.e., the ones reported in this book) will be the models with the lowest AIC (best fit) that did not return any error. Importantly, however, fixed effects that are essential to the model (e.g., to answer the research questions) will not be removed, even if they are non-significant and do not improve fit.

In both studies, the principal fixed effects will be Group (SW × CW, Study 2; or SW × Timed CW × Untimed CW, Study 3) and Time (pretest and posttest). The main random effects will be Participants and Items (the keywords). Teachers will not be included as a random effect in Studies 2 and 3 because participants had only two teachers and a random effect needs to have at least five levels (here, teachers) to be reliable (Bolker, 2020; Hajduk, 2019; Harrison et al., 2018). Likewise, Sets A and B (the keywords were divided into two compatible sets of 10 words each; see Tables 4.2 and 4.3), thus two levels, cannot be included (although the random effect for Items will account for the variance in keywords). Classes had four levels in Study 2 and nine levels in Study 3, and hence will be included as a random effect only in Study 3. More details on the fixed and random effects and dependent variables specific to each study will be discussed in their respective chapters.

4.5 Conclusion

This chapter was divided in two main parts. The first part has provided an overview of the three studies reported in this book, with tables summarizing the research design of each study. The second part tackled some issues related to statistical analyses. I first outlined aspects of inferential statistics that are germane to the statistical analyses conducted in this book. Then, this chapter provided a more detailed overview of linear mixed models (LMMs), which I hope will be useful for readers. The next chapter, Chap. 5, reports on Study 1 and fully discusses its findings. Chapters 6 and 7 will report on Studies 2 and 3, respectively, and each will bring a very brief discussion of the results. Chapter 8 will then fully discuss the results of Studies 2 and 3, which are complementary.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2018). *Parsimonious mixed models*. Retrieved from: https://arxiv.org/pdf/1506.04967.pdf. Accessed November 26, 2020.

Bolker, B. (2020). GLMM FAQ. Retrieved from: http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#reml-for-glmms. Accessed November 27, 2020.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Brzeziński, J., Gaul, M., Hornowska, E., Jaworowska, A., Machowski, A., & Zakrzewska, M. (2004). SkalaInteligencjiD. Wechsleradladorosłych. Wersjazrewidowana–Renormalizacja,WAIS-R(PL). Pracownia Testów Psychologicznych PTP.

Carson, R. J., & Beeson, C. M. L. (2013). Crossing language barriers: Using crossed random effects modelling in psycholinguistics research. *Tutorials in Quantitative Methods for Psychology, 9*(1), 25–41.

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

Cunnings, I., & Finlayson, I. (2015). Mixed effects modelling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 159–181). Routledge.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Hajduk, G. K. (2019). *Introduction to linear mixed models*. Retrieved from: https://ourcodingclub.github.io/tutorials/mixed-models/#crossed. Accessed November 27, 2020.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ,* 1–32.https://doi.org/10.7717/peerj.4794

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2012). *Multilevel modelling of categorical outcomes using IBM SPSS*. Routledge.

Hedeker, D. (2003). A mixed-effects multinomial logistic regression model. *Statistics in Medicine, 22*, 1433–1446. https://doi.org/10.1002/sim.1522

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Cengage Learning.

IBM SPSS. (2005). Linear mixed-effects modeling in SPSS: An introduction to the mixed procedure. Retrieved from: https://www.spss.ch/upload/1126184451_Linear%20Mixed%20Effects%20Modeling%20in%20SPSS.pdf. Accessed November 26, 2020.

IBM SPSS. (2020). IBM SPSS advanced statistics 26. Retrieved from: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/26.0/en/client/Manuals/IBM_SPSS_Advanced_Statistics.pdf. Accessed November 26, 2020.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Pearson Education Limited.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 1–22. https://doi.org/10.1016/j.jml.2020.104092

Minitab 19 Statistical Software. (2019). *Computer software*. Minitab, Inc. Retrieved from: www.minitab.com

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Perry, F. L. (2011). *Research in applied linguistics: Becoming a discerning consumer* (3rd ed.). Routledge.

RStudio Team. (2020). *Computer software*. Integrated Development for R. RStudio, PBC. Retrieved from: http://www.rstudio.com/

Salkind, N. J. (2011). *Statistics for people who (think they) hate statistics* (4th ed.). Sage Publications.

Seltman, H. J. (2018). Experimental design and analysis. Retrieved from: http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf. Accessed November 26, 2020.

Silva, B., & Otwinowska, A. (2019). VST as a reliable academic placement tool despite cognate inflation effects. *English for Specific Purposes, 54*, 35–49.

Silva, B. B., Kutyłowska, K., & Otwinowska, A. (2021). Learning academic words through writing sentences and compositions: Any signs of an increase in cognitive load? Language Teaching Research, 1–33. https://doi.org/10.1177/13621688211020421

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. H. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (pp. 4–31). Psychology Press.

Wesche, M., & Paribakht, T. M. (1996). Assessing vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13–40.

Winter, B. (2020). *Statistics for linguists: An introduction to R*. Routledge.

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.

Chapter 5
Study 1—The Assessment of Academic Vocabulary: Developing a Reliable Academic Placement Tool

5.1 Introduction

This chapter reports on a study that has been published in the journal *English for Specific Purposes* (Silva & Otwinowska, 2019). The chapter gives an extended account of the study, with some editions and inclusions, and a more extensive discussion. The reproduction of the paper in this book complies with the journal's copyright rules for authors' rights (Elsevier, 2020).

The study reported here measures the general and academic vocabulary size of undergraduate English majors at a large university in Poland. In doing so, it uses quantitative and qualitative measures. The study seeks detect signs of inflation in test scores due to the presence of cognates and to suggest how a general vocabulary size test may be used for university placement purposes with higher education students.

5.2 Method

5.2.1 Aims and Research Questions

As discussed in Chap. 1, knowledge of academic vocabulary is essential for academic success. However, academic words are often abstract and morphologically complex (Corson, 1997), and are therefore difficult to learn incidentally, even after learners have spent years studying at English-medium universities in English speaking countries (Knoch et al., 2015). As a result, a more explicit focus on academic words is needed in order to provide university students with sufficient knowledge of academic vocabulary. First, nevertheless, it is necessary to identify learners who need extra practice with academic words.

One potential problem when trying to identify these learners is which test to use. This is because academic words, being mostly of Greek and Latin origin (see Sect. 1.2.1), contain a large proportion of cognates, which are easier to recognize

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_5

and learn than noncognates (see Sect. 1.3). As a result, the presence of cognates in standardized international vocabulary tests, designed for all learners and often utilized to assess level of proficiency, distorts scores. The reason is that learners whose L1 borrow heavily from Latin or Greek may obtain artificially inflated test results, hence having an advantage over learners whose L1 has fewer Greek- or Latin-based words (see Sect. 1.5). Because of this, it is possible, for example, that speakers of Spanish may have higher scores in standardized vocabulary tests than speakers of Polish, and thus be erroneously deemed more proficient. What follows is that the possibility of cognate inflation must be considered when assessing learners' knowledge of academic vocabulary.

Researchers have suggested a few solutions to this problem. Gyllstad et al. (2015) favor keeping cognates in tests, but this does not solve the problem with the inflation of scores. Petrescu et al. (2017) suggest scoring cognates and noncognates separately but do not explain when one or the other would be more suitable. Other researchers (e.g., Elgort, 2013) have recommended keeping in the test the same proportion of cognates as the proportion in learners' L1. The problem here is that the exact proportion of cognates is usually unknown; also, this is not a solution when the test is developed to be used with learners from different backgrounds.

Considering the above, the purpose of this study is two-fold. First, to assess knowledge of academic vocabulary in the case of BA-level English majors at a Polish university utilizing multiple measurements of receptive lexical knowledge. When doing so, this study assesses a much larger number of general and academic lexical items than previous studies did. Second, by combining the VST (Nation & Beglar, 2007)—a meaning-recognition multiple-choice test assessing receptive vocabulary size—and an academic-vocabulary meaning-recall test (AVT), tailor-made for the present study, I aim to verify the practicability and reliability of using the VST for academic placement purposes with English majors. When doing this, I attempt to find a manner of assessment that is less sensitive to cognate inflation effects and reflective of the test-takers' vocabulary knowledge, thus circumventing the disagreement concerning what to do with cognates. Considering my goals, the research questions were as follows:

- RQ1. Do undergraduate English majors at a Polish university achieve similar scores on cognates and noncognates as measured by an academic-vocabulary meaning-recall test (AVT) and the VST?
- RQ2. If cognate inflation is detected, is this effect similar across the 14 levels of the VST?
- RQ3. Can the VST predict learners' academic vocabulary knowledge? If so, is there a threshold on the VST that differentiates learners with regards to their academic vocabulary knowledge?

5.2.2 Participants

Data were collected from 116 speakers of Polish, undergraduate English majors (68 first-year and 48 s-year students) at a large university in Poland. The participants' English proficiency at university entrance was measured by an external national exam at the B2 level according to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The minimum score for acceptance was 87%, so participants' level can be estimated at B2 or above. In the study, the level was additionally measured by participants' performance on the VST ($M = 9877$, $SD = 1428$, Range: 6600–12,800).

Data were analyzed from 106 students: 61 first-year (42 females, $M_{age} = 19.84$, $SD = 2.59$) and 45 s-year (35 females, $M_{age} = 20.16$, $SD = 1.11$) participants. Four first-year and three second-year learners were eliminated from the analyses due to excessive guessing in the AVT (see Sect. 5.2); one first-year student was excluded as an outlier, and two for failing to complete the test.

Importantly, apart from Polish native speakers, there were 6 Russians, 4 Ukrainians, 2 Arabic speakers and 1 Spanish speaker. However, they were not excluded from the analysis for the following reasons. First, this mixed-nationality sample reflects the reality of the institution, which must be considered when devising any placement test; second, none of these students showed any signs of being outliers. Finally, their L1s different from Polish could have mattered in the case of vocabulary cognate with Polish. However, this was also controlled for in the study. As confirmed by native speakers of the languages involved (other than the students themselves), of all Polish-English cognates in the study, 95.36% were also cognates or partial cognates in Ukrainian and Spanish, and 96.03% in Russian. Moreover, 76.92% of these foreign participants self-reported their proficiency in Polish as between B2 and C2, including the two Arabic speakers. This means that learners were highly likely to recognize Polish-English cognates in the receptive test measurements.

5.2.3 Instruments

The Vocabulary Size Test (VST). The online version of Nation and Beglar's (2007) Vocabulary Size Test (VST) was chosen as the receptive general vocabulary measure of proficiency (available at www.vocabularysize.com). The VST draws on the 14 frequency-ordered lists of one thousand word families from the BNC. The test contains 10 items from each frequency band, totaling 140 items. Each item appears in isolation and as part of a short, non-defining sentence, and the examinees need to choose the correct meaning from a set of four options. The VST is thus a multiple-choice, meaning-recognition test (see Fig. 5.1 and Appendix B for the whole test). The final score is obtained by multiplying the result by 100, so the maximum score is 14,000, meaning the examinee has a written receptive vocabulary of approximately 14,000 word families.

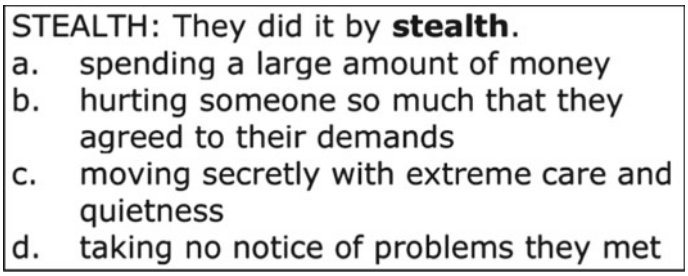

Fig. 5.1 Example item from the seventh band of the VST

Following suggestions by Nation (2012, 2013), Nation and Beglar (2007) and previous research (Bundgaard-Nielsen et al., 2011; Elgort, 2013; Nguyen & Nation, 2011), it was decided that the participants should sit the whole test. When validating the VST, Beglar (2010) recommended the examinees sit at least two levels above their current proficiency level. As the participants were estimated to be at the English B2 level or higher, using the whole test was considered suitable.

The Academic Vocabulary Test (AVT). The Academic Vocabulary Test (AVT), tailor-made for the study, consisted of 405 items present in the AWL. It is a checklist test, also called a *Yes/No* test and first introduced by Meara and Buxton (1987). In this test, learners are instructed to tick only known words. According to Schmitt (2010a), it could be considered a meaning-recall test, although no translation is required. Yes/No tests are regarded as reliable and valid measurements (Milton et al., 2010; Mochida & Harrington, 2006; Nation & Webb, 2011) and are widely used in SLA and psycholinguistics (e.g., Lemhöfer et al., 2008; Vidal, 2011). To avoid different interpretations of "knowing a word", as recommended by Schmitt (2010a), the task script instructed participants to tick only the words they were sure they could translate or define in any language they knew (see Fig. 5.2 for part of the test and Appendix C for the AVT in its entirety).

Creating a Corpus of Texts. The AVT was built from a corpus of applied linguistics texts of 168,598 tokens. The corpus comprised two kinds of texts: 12 book chapters and 12 research papers. The 12 chapters were selected from *An Introduction to Applied Linguistics* (Schmitt, 2010b), included because such introductory texts were deemed relevant to undergraduate participants (Lei & Liu, 2016). The 12 research

adair		precise		automatic		text	
retrogradient		distribute		contrivial		bance	
detect		strategy		chapter		cottonwool	
creative		final		acknowledge		research	
stimulcrate		detailoring		awareness		community	

Fig. 5.2 An example of the presentation of items in the AVT

Table 5.1 Book chapters and papers used in the creation of the corpus

Book chapter	Length of chapter (tokens)	Research articles (RAs)	Length of RA (tokens)	Total length
2—Grammar	7451	Spada and Lightbown (2008)	8507	15,958
3—Vocabulary	8459	Newton (2013)	9099	17,558
5—Pragmatics	8335	Angouri (2012)	7401	15,736
6—Corpus linguistics	6767	Khani and Tazik (2013)	4760	11,527
8—Psycholinguistics	8635	Silva and Otwinowska (2018a)	7918	16,553
9—Sociolinguistics	6494	Guy (2013)	4135	10,629
10—Styles, strategies and motivation	7392	Dörnyei (2009)	5887	13,279
11—Listening	6379	Field (2008)	7132	13,511
12—Speaking and pronunciation	7187	Evison et al. (2007)	5228	12,415
13—Reading	6827	Renandya (2007)	4764	11,591
14—Writing	6439	Hartshorn et al. (2010)	6903	13,342
15—Assessment	8457	Min (2016)	8042	16,499
Total	88,822		79,776	168,598

articles were of similar length, were published between 2007 and 2017, and each tackled a topic equivalent to a chapter of the abovementioned book. Thus, a total of 12 topics (i.e., sub-corpora) were used to create the test. Table 5.1 shows the chapters and research articles used with their respective lengths provided.

Selecting the Words for the Test. The selection of words for the AVT was conducted as follows. Based on Coxhead's (2000) criteria and using "Range for Texts V3" and "Text Lex Compare", both available at *Lextutor* (Cobb, 2020), words that occurred at least 5 times in the whole corpus (frequency) and across 5 of the 12 subcorpora (range) were selected. Of those, Text Lex Compare found that 583 types overlapped with the 3100 types that comprise the AWL; however, 10 words were eliminated as Polish-English false cognates. Given the Yes/No test format, I decided to exclude all false cognates, as it would be impossible to tell whether the ones marked as known were really known by my participants.

The remaining 573 types were then transformed into *lemmas*, i.e., "words with a common stem, related by inflection only, and coming from the same part of speech" (Gardner & Davies, 2013, p. 308). Lemmas were used instead of types to reduce

the number of keywords (e.g., the types "acquire", "acquired" and "acquiring" were replaced by the lemma "acquire"); also, they were used instead of families so that parts of speech could be represented (e.g., I opted to keep the three lemmas "accuracy", "accurate" and "accurately" instead of solely the family member "accuracy"). This is important when seeking to detect cognates and false cognates in that cognateness might be manifested in some but not other parts of speech (cf. English-Polish translations for nouns and adjectives: "coordinator" and "koordynator" vs. "coordinated" and "skoordynowany"). Furthermore, using lemmas instead of families helped avoid the possibility of scores being inflated by the assumption that learners know the entire family of a lexical item, which research has demonstrated to be unlikely, even receptively (Ward & Chuenjundaeng, 2009). After lemmatization, the final list consisted of 405 items (300 noncognates and 105 Polish-English cognates).

In order to compare the coverage of my corpus and wordlist to previous studies, which utilized families instead of lemmas, I submitted my 405 items to "VocabProfile" in Lextutor (Cobb, 2020) and found they corresponded to 308 AWL families. The analyses with Text Lex Compare (Cobb, 2020) revealed that the 570 AWL families covered 12.77% of my corpus of applied linguistics texts, thus similar to the average of 12.21% reported in previous studies (e.g., Chung & Nation, 2003; Cobb & Horst, 2004; Khani & Tazik, 2013; Vongpumivitch et al., 2009). By comparison, my 308 families provided a coverage of 11.03%, or only 1.75% less than the whole AWL. Consequently, my corpus was comparable to others in the field, and the keywords seemed suitable for my test as they represent the AWL items most relevant to students in the field of applied linguistics. Table 5.2 provides details of the coverage of the

Table 5.2 Coverage of the target academic words in my study compared to coverage of the AWL

Corpora topics	Coverage lexical items in study (%)	Coverage of the whole AWL[a] (%)	Difference (%)
1	8.72	10.17	−1.45
2	9.97	11.76	−1.79
3	10.94	12.56	−1.62
4	12.33	13.92	−1.59
5	10.00	11.54	−1.54
6	12.81	14.38	−1.57
7	9.68	11.60	−1.92
8	10.51	12.88	−2.37
9	11.42	12.75	−1.33
10	14.68	16.56	−1.88
11	10.17	12.12	−1.95
12	11.07	13.01	−1.94
Average	11.03	12.77	−1.75

Note The percentages given represent the text coverage of running words (tokens)
[a]The 3100 types were used for analysis, representing the 570 word families of the AWL

keywords per topic of the corpus.

Creating Test Versions . Since a test with all 405 keywords might be rather long, and fatigue effects might occur, three equivalent versions of the AVT were created. The 300 noncognates and 105 cognates were first roughly divided into three versions. Then, the items within the three noncognate and cognate groups were matched for parts of speech, their raw frequency of occurrence in the academic section of the Corpus of Contemporary American English (COCA; Davies, 2012) and concreteness (ratings on a scale: 1—abstract to 5—concrete; Brysbaert et al., 2014). This was controlled because concrete words are learned better and faster than abstract words, which also pertains to cognates (see e.g., De Groot, 2011 for discussion).

Table 5.3 shows the descriptive statistics of each version of the test (see also Appendix C for the three versions of the AVT). One-way ANOVAs[1] revealed no significant differences between the three test versions in frequency ($p = 0.520$) and concreteness ($p = 0.895$) for noncognates; for cognates, the differences were equally non-significant for frequency ($p = 0.922$), and concreteness ($p = 0.574$). Consequently, I was confident the three versions of the test, each containing 100 noncognates and 35 cognates, were fully equivalent.

In the final step, 65 *plausible nonwords* were added, the same words for each AVT version. Nonwords are non-existing words that follow the orthographic and phonological conventions of the English language (De Groot, 2011). The nonwords, taken from Paul Meara's list available at Lextutor (Cobb, 2020), were included to control for participants' guessing. In each AVT version they constituted about 33% of the test items, as recommended in other studies (Lemhöfer et al., 2008; Vidal, 2011). The instruction alerted learners to the inclusion of nonwords; it urged participants not to guess and informed them that ticking too many nonwords would imply excluding their data from the analysis. Three scores were derived from the AVT: (1) a percentage for all items, (2) a percentage for cognates and (3) a percentage for noncognates (see Sect. 5.3).

5.2.4 Procedure

Piloting the Tests. The AVT test and the online VST were piloted on a group of 10 graduate students who also served as a focus group to comment on the AVT layout and construction. They drew my attention to a lack of space in the original layout of the test, but this was improved in the final version.

Main Data Collection. The data for the main study were collected at the beginning of the academic year from 5 first-year and 6 s-year intact groups of students. The testing took place during normal class time over a period of two weeks. No time limit was set for either test, but the AVT and the VST took between 7–15 and 30–60 min

[1] Here, the data failed that assumption of normality and was therefore log-transformed. See Sect. 5.3 for more details.

Table 5.3 Descriptive statistics per test version for test items

Version	Frequency			Concreteness			No. Items per part of speech			
	M	*SD*	*Mdn*	*M*	*SD*	*Mdn*	Noun	Verb	Adj	Adv
Noncognates (n = 100)										
1	11,767.35	13,076.77	8391	2.36	0.67	2.17	39	24	12	25
2	11,656.84	9382.03	9101	2.41	0.69	2.27	41	24	12	23
3	11,811.03	10,043.08	9124	2.40	0.69	2.32	39	22	15	24
Cognates (n = 35)										
1	15,167.43	12,401.43	13,866	2.61	0.78	2.54	20	4	0	11
2	14,883.20	11,225.29	10,521	2.52	0.73	2.44	20	4	0	11
3	15,025.29	11,600.80	11,559	2.71	0.78	2.44	19	4	0	12

Table 5.4 Percentages of correct AVT items ($N = 135 + 65$ nonwords) per test version

Groups	Test version (No. of participants)	*M*	*SD*	*Mdn*
Year 1	1 (19)	89.55	9.33	92.59
	2 (22)	87.51	14.33	90.74
	3 (20)	88.37	10.11	92.59
Year 2	1 (15)	90.27	10.52	94.81
	2 (17)	90.81	9.95	96.30
	3 (13)	87.41	10.10	88.89

to complete, respectively. The instructions were given orally, and participants were given time to read the test scripts (including the instructions) and ask questions. The participants were assured their results would remain anonymous and were urged not to guess.

The three versions of the AVT were evenly distributed among the participants (see Table 5.4). Also, to avoid test-practice effects, the order of testing was counterbalanced (52.46% first- and 53.33% second-year students sat the AVT before the VST). Learners taking the AVT first remained in their classrooms with their classroom teacher; those sitting the VST went to the computer lab with the first author. In the lab, participants read the test instruction on the screen and were orally discouraged from making wild guesses. Although it was impossible to progress in this online test without selecting an item, and there was no "I don't know" option, participants were asked to "think carefully and make the best guess possible". By doing so, learners were better equipped to draw on subconscious knowledge (Nation, 2012).

5.3 Analysis

5.3.1 Variables Derived from the AVT

The participants who ticked 10% or more of the nonwords on the AVT test were excluded from the analysis in order to reduce the influence of guessing on the scores, as suggested by Schmitt et al. (2011). For the remaining participants, the correctly indicated AVT items were converted into a percentage (i.e., the score). Three dependent variables were derived from the test: AVT_Total (all 135 items), AVT_Noncognates (100 items) and AVT_Cognates (35 items). Following Read (1988), I considered a minimum AVT_Total score of 88.89% indicative of receptive mastery of the AVT words. Of note, other scores have been suggested in the literature, e.g., 86.66% (Schmitt et al., 2001) and 80% (Lin & Morrison, 2010). However, 86.66% was only a rough approximation of Read's (1988) original suggestion, and 80% was considered too liberal for a receptive test for academic placement purposes.

5.3.2 *Variables Derived from the VST*

Results in the VST yielded three variables: VST_Score (the total score for performance), VST_Cognates (the proportion [percentage] of participants who answered cognates correctly), and VST_Noncognates (the proportion of learners answering noncognates accurately). To do so, I identified all the Polish-English cognates (46) in the 14 frequency bands of the VST. The number of items per band differed, ranging from one to six words. Then, I searched the bands for the same number of noncognates that could be matched with cognates for concreteness (Brysbaert et al., 2014) and raw frequency in SUBTLEXus (Brysbaert & New, 2009). Since band 13K has 6 cognates, and each band consists of a total of 10 words, only 4 noncognates could be matched, and hence only 44 noncognates were chosen for comparison (see Appendix D for the list of 46 cognates and 44 noncognates).

Also, as several bands had only 1–3 items of each word type, counting participants' mean score per band would result in very limited scores. For instance, the possible scores for band 3K, with only one item, would be 0% or 100%. Instead, for each lexical item, I calculated the percentage of students in each of the 11 intact classes who answered the item correctly. For example, band 1K has three items, so the data for this band comprises 33 (11 × 3) cases (i.e., scores) per word type. Thus, effectively, VST_Cognates and VST_Noncognates are the proportion of correct answers for each band (and not the mean score per VST band).

All data used for the analyses are available online (see Silva & Otwinowska, 2018b). Because most of the variables were heavily skewed, all data were normalized via a log-transformation (Ln) before being submitted to parametric analyses, unless otherwise stated. The three variables derived from the AVT, and VST_Cognates, were also reflected[2] as their skew was negative. This means that larger values became smaller, and vice-versa, so the resulting trend corresponds, effectively, to its inverse. To ascertain that the three versions of the AVT were equivalent before answering the research questions, a 2 × 3 ANOVA was run on AVT_Total searching for the main effects of Year (first and second year) and Test version (Versions 1, 2, and 3). Table 5.4 shows the scores per group in each test version of the AVT. The results revealed no effect for Year, $F(1, 100) = 0.732$, $p = 0.394$, Test, $F(2, 100) = 1.066$, $p = 0.348$, and the Year × Test interaction, $F(2, 100) = 0.594$, $p = 0.554$. Consequently, the data from first- and second-year students were collapsed for further analyses.

[2] The reflected variable was calculated by the following formula: New X = Ln ((K + 1) – X), where K = maximum raw score.

Table 5.5 Percentage of correct cognates and noncognates in the AVT

	M	SD	Mdn
AVT_Total	88.94	10.85	92.59
AVT_Noncognates	86.76	12.86	91.00
AVT_Cognates	95.53	6.97	97.14

5.4 Results

5.4.1 Comparison of Test Performance for Cognates and Noncognates

The first research question pertained to differences in the students' scores on cognates and noncognates in the VST and the AVT. The students' mean performance on the VST was $M = 9877$ ($SD = 1428$, Range: 6600–12,800); the descriptive statistics for the students' performance on the AVT are presented in Table 5.5. To compare the performance on the AVT for cognates and noncognates, the two variables of AVT_Noncognates and AVT_Cognates were submitted to a paired-samples t-tests. The results showed that the scores for cognates were significantly higher than the scores for noncognates, with a large effect size: $t(105) = 14.059$, $p < 0.001$, $r = 0.81$ (Plonsky & Oswald, 2014).

Research question 2 asked whether cognate inflation has a similar effect in the 14 bands of the VST. What needs to be answered here is whether the difference in scores between cognates and noncognates is wider in the lower-frequency bands than in the higher frequency bands. To answer this question, I was able to calculate the scores per frequency level and compare the variables VST_Cognates and VST_Noncognates per band. Figure 5.3 displays the comparison of cognates and noncognates; Table 5.6 lists the results of the Wilcoxon signed-rank tests.[3] Clearly, more participants knew cognates better than noncognates in almost all frequency bands, with medium ($r = 0.40$) to very large ($r = 0.87$) effect sizes (Plonsky & Oswald, 2014). Curiously, participants found noncognates easier than cognates in band 1K. I will discuss and qualitatively explain these findings in the next section.

Next, I ran two simple linear regressions to find out whether the word frequency (as per SUBTLEXus) for cognates and noncognates predicted the proportion of the correct answers for VST_Cognates and VST_Noncognates. Results show that while the frequency of cognates did not explain any variance (Adjusted $R^2 = 0.022$, $F(1, 44) = 1.991$, $p = 0.165$), the frequency of noncognates explained 28.1% of the variance (Adjusted $R^2 = 0.281$, $F(1, 43) = 17.772$, $p < 0.001$) and significantly predicted VST_Noncognates ($\beta = 0.545$, $p < 0.001$). These results demonstrate that frequency of occurrence in English only plays a role in predicting correct answers

[3] Here, even after transformation, several of the variables still failed the assumption of normality; also, several bands had a low number of data points (e. g., band 3K), which further justified using this non-parametric paired-samples t-test.

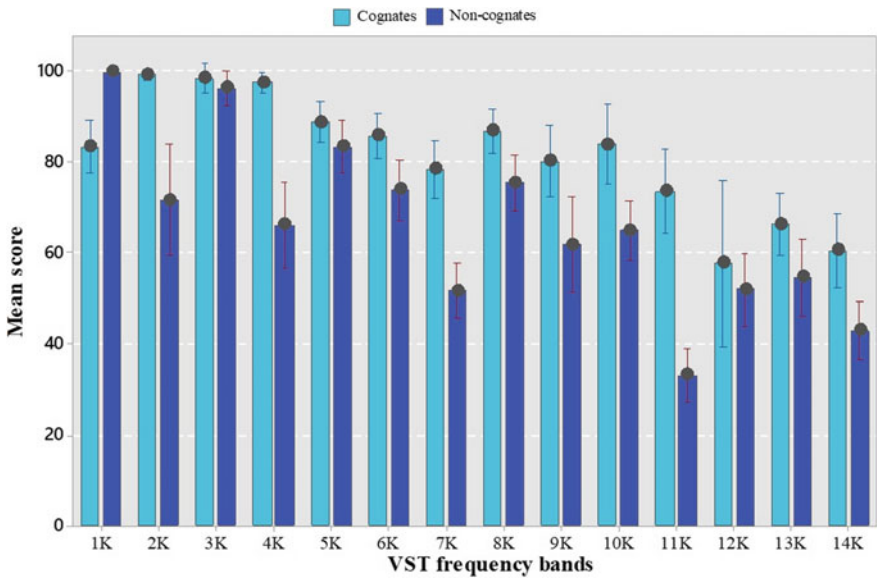

Fig. 5.3 Comparison of percentage of learners who answered cognates and noncognates correctly in each band of the VST. Higher-number bands contain lower-frequency words

Table 5.6 Results of Holm-Bonferroni-corrected Wilcoxon tests comparing 46 cognates to 44 noncognates in the VST

VST band	*Z*	No. of cases	Effect size (*r*)[a]
1K	4.043**	33	0.70
2K	−3.186*	22	0.68
3K	−0.813	11	0.25
4K	−4.542**	33	0.79
5K	−1.900	44	0.29
6K	−2.588*	33	0.45
7K	−5.176**	55	0.70
8K	−3.388**	44	0.51
9K	−3.38**	33	0.59
10K	−2.997*	22	0.64
11K	−5.744**	44	0.87
12K	−0.634	22	0.14
13K	−2.712*	44	0.41
14K	−2.68*	44	0.40

Note *$p < 0.01$; **$p < 0.001$

[a]According to Plonsky and Oswald's (2014) recent guidelines for effect sizes in the field of applied linguistics, an effect may be considered small when $r = 0.25$–0.39, medium from $r = 0.40$, and large from $r = 0.60$

for noncognates, with lower-frequency words having lower results. However, participants' knowledge of cognates is not predicated on how frequent these words are in English. In other words, it may be that even very low-frequency cognates may be well known by participants, a clear sign of cognate inflation effect. Furthermore, based on the results of the regression analyses, it appears that cognate inflation is more pronounced at lower-frequency bands, which answers my second research question. This is because while the score for noncognates decreases with frequency, the scores for cognates do not, likely resulting in an ever-increasing disparity in the scores for cognates and noncognates. Put differently, cognate inflation will be larger as words become less frequent, probably until the cognates themselves become rare in the learners' L1, here Polish.

5.4.2 *Using the VST to Predict Academic Vocabulary Knowledge*

The third research question asked whether the VST could predict scores on the AVT, and whether there was a threshold in the VST that could help predict learners' academic vocabulary (as measured by the AVT). To answer this question, I first conducted a simple linear regression analysis with the VST_Score as the predictor and the AVT_Total as the outcome variable. The results indicated that VST_Score explained 38.5% of the variance (Adjusted $R^2 = 0.385$, $F(1, 104) = 66.637$, $p < 0.001$) and significantly predicted AVT_Total ($\beta = 0.625$, $p < 0.001$). Figure 5.4 displays the regression model.

Then, cluster analyses were run with the variables AVT_Total and VST_Score in the following way. First, I manually grouped VST_Score into High (equal or above 10,000, the approximated mean) and Low (below 10,000). Then, I ran a hierarchical

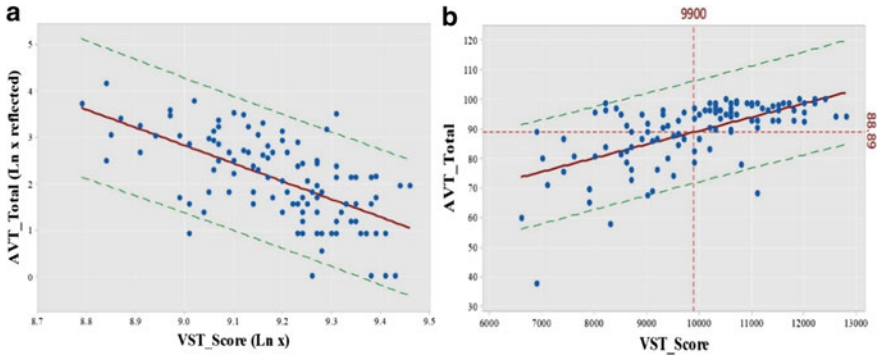

Fig. 5.4 The relationship between VST_Score and AVT_Total. *Note* **a** shows the regression reported with AVT_Total reflected, thus the negative trend; **b** presents the regression with the untransformed data

analysis using Ward's method and Euclidean Distance as the measure of association and standardizing the data into z scores. Figure 5.5 illustrates the division into two main clusters. I subsequently ran a K-means cluster analysis on the same two variables to optimize the results. The analysis identified a VST score of 9900 as the High-Low threshold, confirming my intuitive division. Figure 5.4b shows how this threshold and the score of 88.89% in the AVT (the minimum score needed to determine receptive mastery) intercept on the regression line. Fifty-eight participants were classified at or above the threshold in the VST; 48 learners fell below the threshold, as presented in Table 5.7.

Independent-samples *t*-tests were used to compare the scores between the High and Low groups. I found significant differences between the High and Low students' scores for VST_Score, $t(104) = 14.439$, $p < 0.001$, $r = 0.82$, and AVT_Total, $t(104) = -7.352$, $p < 0.001$, $r = 0.58$, both representing large-sized effects (Plonsky & Oswald, 2014). Among the 58 participants at or above the 9900 threshold, 52 also showed scores of 88.89% or higher in the AVT. The six other participants had a

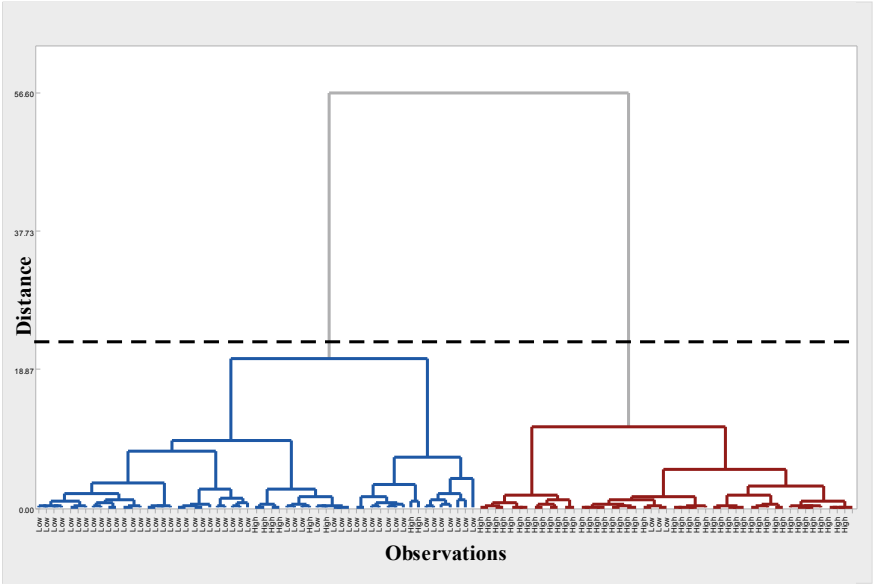

Fig. 5.5 Dendrogram from hierarchical analysis showing the two clusters

Table 5.7 Comparison of the high and low VST scorers revealed in cluster analysis

	VST		AVT	
	M	*SD*	*M*	*SD*
High scorers ($n = 58$)	10,955	759.31	94.15	6.10
Low scorers ($n = 48$)	8575	831.43	82.65	12.18

mean AVT score of 79.38% ($SD = 6.38$). In answering RQ3, because 89.65% of participants who scored at or higher the 9900 VST threshold also succeeded in the AVT, it seems that although not perfect, the threshold is a good indicator of academic vocabulary mastery for Polish learners of English.

5.5 Discussion

Utilizing the entire Academic Word List (AWL; Coxhead, 2000) for assessment, with its 3100 types, is obviously not practicable. On the other hand, existing academic-vocabulary test measurements such as the academic section of the Vocabulary Levels Test (VLT; Schmitt et al., 2001) target only 30 lexical items, which is far too few items to produce reliable results. Furthermore, the proportion of cognates in the academic section of the VLT is likely to differ from the proportion of English cognates in Polish, thus distorting students' results (see Allen, 2018; Elgort, 2013; Laufer & Mclean, 2016). Consequently, this study explored a novel way to estimate Polish learners' receptive academic vocabulary size in English, so the estimation could be used for placement purposes in the academic context. To this aim, I combined the readily-available and easy to administer VST (Nation & Beglar, 2007) and a custom-made English academic vocabulary meaning-recall test (AVT) to examine whether the VST score could predict the academic vocabulary knowledge of English majors. In doing so, I attempted to find a manner of vocabulary assessment that would be less sensitive to cognate inflation effects than either of the tests conducted on their own. During the process, I addressed some of the limitations of previous studies. First, I utilized multiple measurements of receptive lexical knowledge, previously done only with general, but not with academic, vocabulary (e.g., Laufer & Mclean, 2016). Second, I assessed a much larger number of general (140) and academic words (405) than previous studies did (e.g., Jordan, 2012; Petrescu et al., 2017). The findings were as follows.

5.5.1 Cognate Inflation in the VST and AVT

Overall, scores in the VST ($M = 9877$, Range: 6600–12,800) showed that participants possessed a high level of general lexical knowledge. This was evinced by comparing my mean to average scores in previous research, that is, between 5922 and 8000 words (Beglar, 2010; Bundgaard-Nielsen et al., 2011; Elgort, 2013; Nguyen & Nation, 2011). Additionally, such scores confirmed my expectations that the English majors investigated here were at level B2 or higher. However, my analyses revealed a clear cognate inflation effect: Scores were considerably higher for cognates than noncognates on the AVT and on 10 of 14 frequency bands of the VST. This lends support to previous findings, which attested to cognate inflation effects even when

participants' L1 was not written in the Latin script, hence not sharing the same script as English (Allen, 2018; Elgort, 2013; Jordan, 2012; Laufer & Mclean, 2016).

When higher orthographic similarity exists between the L1 and the L2 words, there is less impediment to cognate recognition (Berthele, 2011; Lemhöfer et al., 2008; Otwinowska & Szewczyk, 2019). Then, learners might benefit even more from the presence of cognates, whether they exist in the language due to linguistic genealogy or cross-linguistic borrowing (Otwinowska, 2015). For instance, Petrescu et al. (2017) found that, in the academic context, speakers of Romanian, a Romance language rich in Latin lexis, outperformed speakers of Vietnamese largely due to a cognate advantage. Similarly, Leśniewska et al. (2018), in a study measuring the receptive vocabulary knowledge of French and Polish learners, demonstrated higher cognate inflation for the speakers of French, a Romance language. This was true even though English-Polish cognates could also be found in the test, which benefited speakers of Polish, a Slavic language, to some extent. Obviously, knowledge of cognates is part of vocabulary knowledge in any L2. Still, in this study, the proportion of English-Polish cognates in the VST was 32% (46 cognates out of 140 words), which is a much larger proportion than the one estimated to exist in Polish (below 9%, as discussed in Sect. 1.6). Therefore, the cognate advantage registered in the VST translates into a score overestimation.

Overall, similar to previous findings (e.g., Elgort, 2013; Jordan, 2012; Petrescu et al., 2017), the current study provides evidence of a more pronounced cognate advantage at lower frequency bands. As the regression analyses showed, scores on noncognates decreased with frequency, but this effect was not found for cognates. Put differently, the model predicted a larger gap in scores at lower frequencies since while scores for cognates remained relatively stable, scores for less frequent noncognates decreased significantly. On the other hand, a closer look at the individual VST bands revealed a different picture as scores for some cognates and noncognates were rather unpredictable. This may be because I did not control for the varying degrees of formal and semantic overlap of cognates in the analysis, and such variation may have affected recognition, as explained in Sect. 1.4. Although this may be seen as a limitation of my study, it made my results comparable to those of previous studies, which also did not take account of such variation (e.g., Elgort, 2013; Petrescu et al., 2017). It also allowed me to better understand how Polish learners deal with cognateness, which will be explained qualitatively below. In fact, this appears to be the first study measuring cognate inflation effects in SLA to qualitatively analyze individual cognates.

5.5.2 Explaining Cognate Inflation for Polish VST Test-Takers

The first surprising finding concerns the 1K band in the VST, representing the 1000 most frequent words in English, wherein noncognates were known better than cognates. The reason was that the cognates "figure" (Pl.: "figura") and "basis" (Pl.:

"baza") proved problematic to learners. This is because the meaning of "figura" in Polish is included in the meaning of the English "figure" but is narrower. In Polish it means "geometrical figure" (e.g., a triangle), or is used to describe people (as in "she has a good figure"); however, the key in the VST was "number", which is not used in Polish (the Polish equivalent would be "liczba"). Therefore, as per the options provided in the VST, it could not be treated as a Polish-English cognate. This shows the importance of considering the polysemy of words when including cognateness as a variable in a study. Moreover, "basis" in the VST contained two confusing options: "reason" (the key) and "main part", both of which could be considered correct. In fact, "basis" has been highlighted by Beglar (2010) as a potentially problematic item, and indeed, many of my participants chose the option "main part".

Furthermore, participants' scores for VST bands 3K, 5K, and 12K did not show significant differences between cognates and noncognates. This may be owing to the fact that one or more of the distractors provided in the test (i.e., the three options in the multiple-choice item other than the key) confused learners. For example, in band 12K, comprised of two cognates, "refectory" and "caffeine", "refectory" (Pl.: "refektarz") was known by only 24.44% of learners. One probable reason for this is the low orthographic similarity of the Polish word "refektarz" to its English counterpart, making it more difficult for learners to recognize it as a cognate (Comesaña et al., 2015; De Groot, 2011; Dijkstra et al., 2010; Duyck et al., 2007; Mulder et al., 2015; Otwinowska & Szewczyk, 2019). The meaning of "refektarz" (used only to refer to the dining hall in a convent) in Polish is also narrower than that in English. Additionally, one of the distractors, "office where legal papers can be signed", may have elicited the Polish word "referat", meaning "bureau", thus misleading learners into selecting this distractor.

On the other hand, in band 11K, the mean for cognates surged to 85.02%, whereas the mean for noncognates was predictably low ($M = 33.19\%$). The reason may be that all cognates in this band ("yoga", "emir", "aperitif", "puma") were the so-called *exotics* (Otwinowska, 2015), i.e., words that are borrowed independently by various languages and are rarely adapted orthographically between languages sharing the same script. Thus, their spelling and meaning were in fact the same as in Polish ("joga", "emir", "aperitif", "puma"). This begs the question whether these exotics should be used in proficiency tests at all, especially whether they should be incorporated as test items in lower-frequency bands. If they are, such as in the VST, they are bound to inflate test-takers' scores across many languages. As an illustration, Table 5.8 demonstrates how little these exotics change across related and unrelated languages.

Summing up, the cognate inflation evidenced in the overall scores in my study is contingent on participants' L1 (Polish), as the qualitative analysis shows. The inflation effect would likely be exacerbated with speakers of Romance languages (Schmitt et al., 2001), and any languages that borrowed heavily from Latin and Greek. This is problematic for tests designed for all learners, such as the VST. Even though bilingual tests may help solve this problem (see Elgort, 2013; Nguyen & Nation, 2011), they need to be developed; and when developed, such tests must contain a percentage of cognates that reflects the percentage in participants' L1

Table 5.8 The exotics "yoga", "emir", "aperitif", "puma" across languages

Language	Exotics
Spanish	"yoga", "emir", "aperitivo", "puma"
Icelandic	"jóga", "emir", "aperitif", "puma"
Hungarian	"jóga", "emír", "aperitif", "puma"
Albanian	"yoga", "emir", "aperitif", "pumë"
Swedish	"yoga", "emir", "aperitif", "puma"
Vietnamese	"yoga", "emir", "Khai vị", "puma"
German	"yoga", "emir", "aperitif", "puma"
Malay	"yoga", "emir", "minuman beralkohol", puma
Finnish	"jooga", "emiiri", "aperitiivi", "puma"
Turkish	"yoga", "emir", "aperatif", "puma"
Czech	"jóga", "emir", "aperitiv", "puma"
Swahili	"yoga", "emir", "aperitif", "puma"
Hawaiian	"yoga", "emir", "aperitif", "puma"
Samoan	"yoga", "emira", "aperitifa", "puma"

Source Google Translator

(Laufer & Mclean, 2016), but such proportion may be unknown in many contexts. Also, although the VST rightly includes cognates, the inclusion of exotics in the test is indicative of a *monolingual bias*, i.e., the lack of insight that the words borrowed into English, such as "yoga", might exist in the same form across diverse languages. Obviously, it is impossible to create tests that will accurately cater for learners of all backgrounds; nevertheless, the VST band 11K seems to be seriously flawed in its exaggerated use of exotics.

5.5.3 *The VST Results Predict Academic Vocabulary Knowledge*

In this study, I further asked whether learners' academic vocabulary knowledge could be predicted based on the VST, and whether there was a threshold in the VST identifying learners in need of academic instruction. To this end, I first found that an increase in the VST scores significantly predicted 38.5% of the increase in the total score of the AVT. The proportion of the variance explained may not appear high, considering that vocabulary can predict between 48 and 64% of variance in comprehension scores of academic reading (Laufer & Ravenhorst-Kalovski, 2010; Milton et al., 2010; Mochida & Harrington, 2006). Still, I believe this can be partly explained in at least three ways. First, the VST measures meaning recognition whereas the AVT measures meaning recall. Second, while the VST focused on general vocabulary, the AVT only measured knowledge of academic words. Last, the tests have different

formats: The VST is a multiple-choice test while the AVT is a Yes/No test. This may mean that learners approach the tests differently, being for example, more inclined to guess in one format or the other, thus distorting scores (see more on guessing below). Given the above, 38.5% of explained variance may be quite large, and indeed, it is considered a large effect size (Field, 2017).

It could also be argued that since the average vocabulary size of the participants was high (9877), such knowledge is bound to include knowledge of academic words, rendering the attempt to measure academic vocabulary and find a threshold unnecessary. The AWL comprises lexical items (families) between the 2K and 8K bands, at least according to corpora available when the list was created 20 years ago. This means that a vocabulary size of 8000–9000 words should include knowledge of all academic words. This is reinforced by the fact that knowledge of 8000–9000 word families is typically deemed sufficient for unassisted reading (Nation, 2006). However, there are several arguments against this stance that support my study. First, according to Webb and Paribakht (2015), knowledge of 9000 academic word families provides a mean coverage of 96.69% of academic texts, which is significantly lower than the 98% required to understand texts (Hu & Nation, 2000). Second, my results suggest that the VST score is inflated, and that the 9900 threshold obtained by combining the VST and AVT helps control for this inflation. Finally, as argued previously, academic words may be more difficult to learn than low-frequency general vocabulary (Corson, 1997; Lubliner & Hiebert, 2011; Nagy & Townsend, 2012; Vidal, 2011), so learners' general vocabulary size of 9000 word families may not mean they are familiar with all academic word families in the AWL.

For example, an analysis of the 3100 AWL types in the VocabProfile tool in Lextutor (Cobb, 2020), using very recent corpora (BNC-COCA 1-25K), showed that the first 8K bands encompass only 98.86% of the AWL types, with items found among the 15K band. Furthermore, of the 22 participants in my study who scored between 8000 and 9000 in the VST ($M = 8563$), only 10 scored at or above 88.89% in the AVT, with a mean of 85.11%. In other words, if one were to rely solely on VST scores, students' receptive knowledge would have been considered sufficient to understand most, if not all, academic words. This is not true. It is with this belief, reinforced by the results of the cluster analyses, that I suggest the VST threshold may be used for placement purposes with Polish English majors.

The cluster analyses pointed towards a threshold of 9900, at or above which 89.65% of participants also demonstrated mastery in receptive academic vocabulary. Even so, several learners scored at or above the threshold but failed the AVT, whilst others scored lower than 9900 in the VST but succeeded in the academic test. This disparity might be explained by the format of the academic test: A checklist test with plausible nonwords added to control for guessing.

5.5.4 *Explaining the AVT Results and Interpreting the VST Results*

An additional qualitative analysis of the data showed that the six learners who failed the AVT despite scoring at or above the 9900 VST threshold were rather conservative when selecting words in the academic test. These learners selected, on average, only 0.77% ($SD = 1.29$) of nonwords, or fewer than one item. This is much below the overall mean of 2.89% ($SD = 2.63$). Moreover, these participants often ticked one member of a word family (e.g., "consistent", "benefit"), while ignoring others (e.g., "consistently", "beneficial"). Thus, these learners appear to have been rather rigorous, likely out of insecurity, which may have underestimated their score in the AVT.

By contrast, 13 learners adopted a less conservative approach to the AVT and thereby succeeded in this academic test while scoring below the VST threshold. These participants were also multilingual, or fluent in at least one other language in addition to English and Polish (e.g., German, Spanish, Russian), so they probably exploited cross-linguistic similarities more liberally (Berthele, 2011; Otwinowska, 2015). This is corroborated by their selecting of a higher proportion of nonwords ($M = 4.5$, $SD = 2.84$), a sign of increased guessing rates, than the average participant in the study. That is, some participants were clearly less conservative than others in the AVT, which impacted scores. Consequently, cognate inflation effects and examinee variability, an inherent characteristic of the checklist format (Mochida & Harrington, 2006; Schmitt, 2010a), may distort scores to an unacceptable extent, rendering them misleading when used for placement purposes.

Effectively, cognate inflation may have been exacerbated by participants' multilingualism and high proficiency in both tests. As discussed previously, learners' high proficiency in English enhances their ability to benefit from cognates of discrepant orthography (Otwinowska & Szewczyk, 2019). Similarly, proficient multilinguals are more likely to take advantage of cognates and correctly guess the meaning of unknown words based on their cross-linguistic similarity (Berthele, 2011). In my sample, over 92% (98 learners) spoke languages other than English and Polish; and over 20% (22 learners) knew their other languages at B2 level or higher, which means they were highly multilingual. Studies devoted to cognate inflation in testing do not report on participants' multilingualism, which I believe will be an important future path to follow in explaining learner variability in test taking.

One may argue that the score distortions imparted by cognate inflation and examinee variability in both tests may render the 9900 VST threshold invalid. However, the purpose of the study was not to measure learners' vocabulary size, and therefore, the effects of such distortions in score may be of little consequence. By creating and employing the AVT, I managed to establish a reliable threshold in the VST. As a result, I argue that the 9900 VST threshold could be established for Polish learners of English as the cut-off point below which university students might need additional teaching of receptive academic vocabulary.

5.5.5 *Limitations of the Study*

Importantly, scores in the AVT and VST cannot be interpreted to fully predict proficiency in listening or any productive skill, which is a limitation of the assessment method suggested here. In fact, the scores are only indicative of reading ability to a limited extent (Beglar, 2010; Nation, 2012; Nation & Beglar, 2007), as vocabulary is not the only element affecting reading comprehension (see Schmitt et al., 2011, p. 29; Webb & Paribakht, 2015, p. 36 for other factors influencing comprehension). Even so, lexical knowledge has been shown to correlate positively and very highly with performance in the four basic skills, including academic reading and writing (Laufer & Ravenhorst-Kalovski, 2010; Milton et al., 2010; Paribakht & Webb, 2016). In other words, while receptive lexical knowledge should not be the only measure used when conducting placement in higher education, it does play a fundamental role in helping estimate learners' academic reading and writing proficiency, and possibly speaking and listening skills. Therefore, I feel comfortable suggesting the VST threshold as a significant part of any assessment-for-placement program with English majors in Poland. It could then be complemented by, for example, a composition writing task to assess learners' productive vocabulary, syntactic and lexical accuracy, and ability to discuss a topic cohesively, clearly, and persuasively.

Another limitation of my study is that learners were tested on single-word lexical items only. There is a considerable body of research that indicates that multiword items are an important part of academic discourse (e.g., Byrd & Coxhead, 2010) and that learners' use of these formulaic sequences is an effective predictor of lexical proficiency (e.g., Crossley et al., 2015). It would be interesting to conduct a similar study in the future investigating formulaic sequences in addition to (or rather than) single-word lexical items. A final limitation refers to my lack of control of the formal similarity between Polish-English cognates. The literature shows that more similar cognates are typically recognized faster (Comesaña et al., 2015; Dijkstra et al., 2010; Duyck et al., 2007; Mulder et al., 2015) and possibly learnt better and retained longer than less similar cognates (e.g., De Groot, 2011; Otwinowska & Szewczyk, 2019; Otwinowska et al., 2020). Had I controlled for this similarity, this extra variable could be entered in a statistical model to increase its power, hence yielding more reliable results.

5.6 Conclusion

In line with previous research, I have demonstrated that cognate inflation effects and examinee language background may misrepresent learners' true lexical knowledge. A novel contribution of this study lies in the investigation of the combination of the tailor-made AVT and the VST, which appears to have ameliorated the effects of such score distortions. In practice, student placement can be facilitated and its reliability increased thanks to establishing a VST threshold below which students are in the

need of academic vocabulary instruction. The VST is freely available online, can be administered in paper or electronic form and is easily scored. Once a threshold has been found, the VST may be used with little concern regarding distortions in score. Nevertheless, if teachers or researchers decide not to identify and implement a threshold, I recommend the following: If possible and practicable, one may ascertain that the number of cognate items in vocabulary tests is proportional to the number of cognates found in learners' L1. If not, cognates should be kept in the test, as without them "it would be impossible to produce valid vocabulary size estimates" (Elgort, 2013, p. 269). This being the case, one must remember cognates may be over- or under-represented in the test and that the resulting distortion in scores should be considered when interpreting the results. I would go further and argue that the misuse of exotics (e.g., "puma") indicates a monolingual bias in test construction that should also be avoided.

Another implication, this time more specific to the Polish context, concerns the number of Polish first- and second-year English majors who scored below the VST threshold. As the cluster analyses have showed, 48 out of 106 learners (45.28%) scored below the 9900 threshold, averaging a score of 8575 in the VST. This means that almost half of the first- and second-year BA-level English majors at a large Polish university would benefit from extra practice with academic vocabulary. This may be also true for third, fourth, or even fifth year students for at least three reasons. First, the institute does not offer any courses geared towards the explicit instruction and practice of academic vocabulary. Second, my (rather limited) teaching experience at the BA-level and anecdotal evidence from colleagues and students indicate that lecturers of content (i.e., most lecturers) do not focus on language (e.g., academic vocabulary), whereas teachers of language do not emphasize academic words. Third, and perhaps most importantly, research has shown that academic words are not learnt incidentally (e.g., in lectures or during reading assignments) even after three years attending classes at English-medium universities in English-speaking countries (e.g., Knoch et al., 2015). As a result, if the first point is true and the second point is to be believed, there is a high likelihood that at least one third of students majoring in English studies obtain their masters' degree without sufficient knowledge of academic vocabulary. Obviously, this is just reasonable speculation at this point, so more research is needed. Still, it is an argument that is sensible enough to warrant further investigation.

In the future, it is also worth investigating learners' productive knowledge of academic vocabulary and comprehension skills in the same context to make sure they can cope with the production and interpretation of academic texts. I would also welcome more studies measuring knowledge of academic words longitudinally, to ensure that current pedagogical practice in higher education facilitates lexical learning. Finally, it is paramount to explore the effectiveness of different tasks in the teaching of academic words as, to my knowledge, evidence in this area is still scarce. This would provide further insight into efficient lexical instruction. Some of the tasks investigated could be those that already permeate higher education, such as the writing of argumentative essays. It was to this aim that I conducted Studies 2 and 3, reported in the following two chapters of this book (Chaps. 6 and 7). They will compare the learning of academic words facilitated by the writing of sentences,

timed argumentative essays (simulating in-class writing under time restrictions), and untimed argumentative essays (simulating out-of-class written assignments).

Funding This work was partly supported by the National Science Centre Poland, under grant number 2016/21/B/HS6/01129 awarded to Agnieszka Otwinowska-Kasztelanic.

References

Allen, D. (2018). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism*, 1–16.https://doi.org/10.1177/1367006918781063

Angouri, J. (2012). Managing disagreement in problem solving meeting talk. *Journal of Pragmatics, 44*, 1565–1579.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118.

Berthele, R. (2011). The influence of code-mixing and speaker information on perception and assessment of foreign language proficiency: An experimental study. *International Journal of Bilingualism, 16*(4), 453–466.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Studies in Second Language Acquisition, 33*, 433–461.

Byrd, P., & Coxhead, A. (2010). *On the other hand: Lexical bundles in academic writing and in the teaching of EAP* (Vol. 5, pp. 31–64). University of Sydney Papers in TESOL.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language, 15*(2), 103–116.

Cobb, T. (2020). *Lextutor [computer software]*. Retrieved from: https://lextutor.ca/. Accessed November 24, 2020.

Cobb, T., & Horst, M. (2004). Is there room for an academic word list in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in second language: Selection, acquisition, and testing* (pp. 15–38). John Benjamins.

Comesaña, M., Soares, A. P., Ferré, P., Romero, J., Guasch, M., & García-Chico, T. (2015). Facilitative effect of cognate words vanishes when reducing the orthographic overlap: The role of stimuli list composition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 614–635.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718.

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570–590.

Davies, M. (2012). *Corpus of contemporary American English* (1990–2012). Retrieved from: https://www.wordfrequency.info/. Accessed November 06, 2020.

De Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction*. Psychology Press.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-linguistic similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*, 284–301.

Dörnyei, Z. (2009). The 2010s communicative language teaching in the 21st century: The 'principled communicative approach.' *Perspectives, 36*(2), 33–43.

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 663–679.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*(2), 253–272.

Elsevier. (2020). *Copyright: Journal author rights.* Retrieved from: https://www.elsevier.com/about/policies/copyright. Accessed November 27, 2020.

Evison, J., McCarthy, M., & O'Keeffe, A. (2007). 'Looking out for love and all the rest of it': Vague category markers and shared social space. In J. Cutting (Ed.), *Vague language explored* (pp. 138–157). Palgrave Macmillan.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Field, J. (2008). Revising segmentation hypotheses in first and second language listening. *System, 36*, 35–51.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327.

Guy, G. R. (2013). The cognitive coherence of sociolects: How do speakers handle multiple sociolinguistic variables? *Journal of Pragmatics, 52*, 63–71.

Gyllstad, H., Vilkaite, 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. https://doi.org/10.1075/itl.166.2.04gyl

Hartshorn, K. J., Evans, N. W., Merrill, P. F., Sudweeks, R. R., Strong-Krause, D., & Anderson, N. J. (2010). Effects of dynamic corrective feedback on ESL writing accuracy. *TESOL Quarterly, 44*(1), 84–109.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 23*(1), 403–430.

Jordan, E. (2012). Cognates in vocabulary size testing—a distorting influence? *Language Testing in Asia, 2*(3), 5–17.

Khani, R., & Tazik, K. (2013). Towards the development of an academic word list for applied linguistics research articles. *RELC Journal, 44*(2), 209–232.

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university. *Journal of Second Language Writing, 28*, 39–52.

Laufer, B., & Mclean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly, 13*(3), 202–217.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. Retrieved in March 2018 from http://files.eric.ed.gov/fulltext/EJ887873.pdf

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes, 22*, 42–53.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 12–31.

Leśniewska, J., Pichette, F., & Béland, S. (2018). First language test bias? Comparing French-speaking and Polish-speaking participants' performance on the Peabody picture vocabulary test. *Canadian Modern Language Review*.https://doi.org/10.3138/cmlr.3670

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*, 255–266.

Lubliner, S., & Hiebert, E. H. (2011). An analysis of English-Spanish cognates as a source of general academic language. *Bilingual Research Journal, 34*(1), 76–93.
Meara, P., & Buxton, B. (1987). An alternative multiple-choice vocabulary test. *Language Testing, 4*, 142–145.
Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. D. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–97). Multilingual Matters.
Min, H.-T. (2016). Effect of teacher modelling and feedback on EFL students' peer review skills in peer review training. *Journal of Second Language Writing, 31*, 43–57.
Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73–98.
Mulder, K., Dijkstra, T., & Baayen, R. H. (2015). Cross-language activation of morphological relatives in cognates: The role of orthographic overlap and task-related processing. *Frontiers in Human Neuroscience, 9*(16), 1–18.
Nagy, W. E., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91–108.
Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–81.
Nation, I. S. P. (2012). *The vocabulary size test: Information and specifications.* Retrieved in December 2017 from http://www.victoria.ac.nz/lals/about/staff/publications/paul-nation/Vocabulary-Size-Test-information-and-specifications.pdf
Nation, I. S. P. (2013) *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.
Nation, I. S. P., & Webb, S. (2011) *Researching and analysing vocabulary*. Heinle Cengage Learning.
Newton, J. (2013). Incidental vocabulary learning in classroom communication tasks. *Language Teaching Research, 17*(2), 164–187.
Nguyen, L. T. C., & Nation, P. (2011). A bilingual vocabulary size test for English for Vietnamese learners. *RELC Journal, 42*(1), 86–99.
Otwinowska, A. (2015). *Cognate vocabulary in language acquisition and use: Attitudes, awareness, activation.* Multilingual Matters.
Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism, 22*(8), 974–991.
Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross linguistic similarity matter? *Language Learning,* 1–47.https://doi.org/10.1111/lang.12390
Paribakht, T. S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes, 21*, 121–132.
Petrescu, M. C., Helms-Park, R., & Dronjic, V. (2017). The impact of frequency and register on cognate facilitation: Comparing Romanian and Vietnamese speakers on the vocabulary levels test. *English for Specific Purposes, 47*, 15–25.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079
Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19*, 12–25.
Renandya, W. A. (2007). The power of extensive reading. *RELC Journal, 38*(2), 133–149.
Schmitt, N. (2010a). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.
Schmitt, N. (Ed.) (2010b). *An introduction to applied linguistics* (2nd ed.). Hodder-Viva Edition.
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.

Silva, B., & Otwinowska, A. (2018a). Vocabulary acquisition and young learners: Different tasks, similar involvement loads. *International Review of Applied Linguistics, 56*(2), 205–229.

Silva, B., & Otwinowska, A. (2018b). Data for paper entitled "VST as a reliable academic placement tool despite cognate inflation effects." *Figshare.* https://doi.org/10.6084/m9.figshare.6047183

Silva, B., & Otwinowska, A. (2019). VST as a reliable academic placement tool despite cognate inflation effects. *English for Specific Purposes, 54*, 35–49.

Spada, N., & Lightbown, P. M. (2008). Form-focused instruction: Isolated or integrated? *TESOL Quarterly, 42*(2), 181–207.

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258.

Vongpumivitch, V., Huang, J.-Y., & Chang, Y.-C. (2009). Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes, 28*, 33–41.

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System, 37*, 461–469.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test. *English for Specific Purposes, 38*, 34–43.

Chapter 6
Study 2—Incidental Lexical Learning Through Writing Sentences and Timed Compositions: Is Learning Affected by Task-Induced Cognitive Load?

6.1 Introduction

The study reported in the current chapter has been published in the journal *Language Teaching Research* (Silva et al., 2021) with some alterations, especially in the Results section. This is the first of two studies reported in this book that explores the learning of academic vocabulary through writing sentences (SW) and argumentative essays (CW). Here, I compare the learning following SW to the learning yielded by Timed CW (i.e., writing essays under timed restrictions). The next study (Chap. 7) expands this comparison to include a third condition: Untimed CW.

6.2 Method

6.2.1 Aims and Research Questions

It is clear by now that Polish university students majoring in English may need assistance with learning academic words (see Chap. 5). It is also clear, based on previous research (e.g., Knoch et al., 2015), that exposure through academic reading and listening alone will not suffice to provide learners with enough knowledge of academic vocabulary. Therefore, a more explicit approach to the teaching and learning of academic words is needed. To achieve this, Chaps. 6 and 7 explore how different writing tasks may be conducive to the learning of academic words.

Chapter 2 has demonstrated that output production facilitates lexical learning. However, most studies investigating learning through writing were conducted with short sentences or, when texts were used, they were short pieces with low complexity. As Chap. 3 has explained, long, well-structured writing is a complex recursive process that demands much cognitive resources from learners. Therefore, any learning yielded by such complex writing may be contingent on the level of complexity of the task and on learners' ability and willingness to allocate attentional resources to

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_6

certain processes of the writing cycle. Because of this, little is known regarding the lexical learning potential of complex writing (Byrnes & Manchón, 2014; Manchón & Williams, 2016; Pichette et al., 2012).

To address this gap, the current quasi-experiment focuses on the incidental acquisition of academic vocabulary through writing sentences and argumentative essays in L2 English in a Polish academic context. Of note, incidental learning is defined in this book as learning that takes place when participants perform a primary task (i.e., writing) involving the processing of some information (i.e., novel academic words) without being aware of the true purpose of the experiment and without being told in advance that they will be tested afterwards on their recall of that information (see Sects. 2.2 and 8.6 for a thorough discussion on incidental learning). Importantly, incidental learning is not the same as implicit learning. The latter refers to learning that occurs unconsciously (Schmidt, 1994) whereas incidental learning may be either implicit or explicit (Laufer & Hulstijn, 2001).

In this study, I measure the acquisition of academic keywords, provided to students in glossaries, through writing sentences and argumentative essays. The rationale behind using the glossaries with keywords is as follows. When writing in the L2, learners use the words they know. If they do not know a word in the L2, but they intend to use that word in writing, they are likely to consult an outside source, e.g., a monolingual or bilingual dictionary. Thus, by providing students with glossaries, I simulate monolingual dictionary use (adding to the ecological validity of the study), but still control for the quality of the academic keywords to be learned. To our knowledge, the use of academic keywords is unique among research of this type. Another advantage of this study is that, thanks to its design, I utilize a larger number of keywords than the typical 10 words used in previous studies (i.e., Kim, 2008; Zou, 2017).

This study is based on the following rationale. It is uncertain whether the cyclical, complex process underlying the writing of argumentative essays may (a) enhance the processing of the keywords—therefore boosting vocabulary acquisition—or, at least when the use of keywords is obligatory, (b) increase the cognitive load of the task to the point of hindering learning. I explore these possibilities in two ways. First, I compare the vocabulary learning induced by sentence writing (SW; effectively, the control group) to the learning following timed argumentative essay (Timed CW) tasks. Then, I compare the control essay (without keywords) to the Timed CW essays (each with 10 keywords) regarding lexical complexity, writing accuracy, and fluency in production. This is because a decrease in one or more of the scores of these measures may indicate an increase in cognitive load (Klepsch et al., 2017; Paas et al., 2003; Robinson, 2001; Skehan, 2014). These textual measures are often used in L2 writing research; by contrast, lexical-learning measurements are typical of SLA and related fields. Thus, this study is an attempt to promote the much-needed interdisciplinary dialogue between the fields (as suggested by Ortega, 2012). By employing types of measures typical of both research traditions, I seek to systematically address the possibility "that the difficulty of semantic elaboration tasks [here, SW vs. Timed CW] may play a key role in their effectiveness" to learning (Rice & Tokowicz, 2020, p. 26).

Considering the above, the research questions are as follows:

- RQ1. Do Polish EFL learners acquire and retain academic words to a similar degree after writing sentences and timed argumentative essays?
- RQ2. Does the need to use pre-specified keywords in the CW task affect overall quality, accuracy, and fluency of students' writing as compared to the control essay, thus indicating an increase in cognitive load?

6.2.2 Participants

Data were originally collected from 55 first-year students majoring in English at the Institute of English Studies, University of Warsaw. However, the number of participants was significantly reduced after they completed a questionnaire at the end of the study (see Appendix E). As suggested by De Vos et al. (2018), the questionnaire was included to ascertain that the study truly measured incidental lexical learning since it is impossible to control what participants do after the treatment, including consulting a dictionary (see also Rice & Tokowicz, 2020). To this aim, the questionnaire asked (a) whether participants had studied any of the keywords post-treatment, and before the posttests, and (b) whether they suspected of the purpose of the quasi-experiment. Nine participants reported having studied at least one of the keywords, mostly by consulting a dictionary after the treatments. These learners were eliminated from further analyses. Seven other learners reported suspecting of the true aim of the study and were therefore also excluded from the analyses. Consequently, 39 learners remained.

The learners were all selected from four Writing Practice intact classes, all following the same syllabus. Two classes were taught by the author of this book and two by another teacher from the same institution. The two classes taught by the author became the composition-writing group (Timed CW), with 17 participants (4 males, $M_{age} = 19.18$, $SD = 0.73$); the two other classes constituted the sentence-writing group (SW), with 22 participants (5 males; $M_{age} = 19.45$, $SD = 1.44$).

All learners were Polish native speakers ($n = 35$) or speakers of Slavic languages ($n = 4$). Of the latter, three were Ukrainian and one was Russian, and all self-reported their proficiency in Polish as B1 or B2 according to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). Regarding English proficiency, only students at the B2 level or higher were accepted for enrolment (as assessed by the university's admission criteria). Additionally, five different proficiency measures were used (see the next subsection) to confirm that participants in the CW and SW groups had similar writing skills and comparable receptive and productive lexical proficiency in English.

Table 6.1 Descriptive statistics for proficiency measures

Group	Receptive lexical knowledge		Textual measures							
			Essay score		Productive lexical knowledge				Productive accuracy	
	LexTALE				D		Frequency		Normed errors[a]	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Timed CW	72.65	13.66	3.71	0.97	90.77	14.19	3.91	0.21	70.03	33.55
SW	80.81	10.08	3.93	0.90	97.16	15.26	3.83	0.17	65.99	34.99

[a](Number of errors/number of words in the text) * 1000. (See Biber et al., 2013; Friginal & Weigle, 2014; Jarvis et al., 2003 for a similar normalization of errors)

6.2.3 Measures of Participant Proficiency

The proficiency measures used in the study included a receptive lexical knowledge test and four measures derived from a control essay written before the experiment. Independent samples *t*-tests were run for all comparisons and showed no significant differences between the SW and the Timed CW groups in any test. Table 6.1 shows descriptive statistics for all five measures used.

Receptive Lexical Knowledge. The online version of Lemhöfer and Broersma's (2012) LexTALE was used (see www.lextale.com). The test is a lexical decision task: Test-takers are presented with strings of letters on a screen and must decide if the string is a real word or a nonword. There are 60 strings in the test: 40 words and 20 nonwords. The final score consists of the percentage of correct answers corrected for the unequal proportion of words and nonwords, and the scores range from 0 to 100 (see Appendix F for the 60 strings and the scoring formula). The test is typically administered to assess overall language proficiency (e.g., De Vos et al., 2018) because its results correlate significantly with the four skills (Lemhöfer & Broersma, 2012). In the current project, Participants' LexTALE scores (range = 52.50–95) confirmed that they were at level B2 or higher (above 70.7 indicates advanced proficiency), with almost no difference between the Timed CW and SW groups, $t(28.457) = -2.067$, $p = 0.048$, $r = 0.36$, representing a small effect size (Plonsky & Oswald, 2014).[1]

[1] This statistically significant difference in receptive lexical knowledge, although minor, may indicate a difference between the groups that might affect the results. I do not believe this is the case for the following reasons. First, the difference is minor. Second, the LexTALE measures receptive knowledge and my study focused on production. Third, the other four measures of proficiency adopted, all of which measure skills in language production, were far from significant (the lowest alpha value was $p = 0.189$ for D). Fourth, the participants were entered in the main analyses as random effects, which may have controlled for the variation in performance between participants, at least partially. Fifth, I obtained a second measure of receptive lexical knowledge, i.e., the scores in the Vocabulary Size Test (VST; Nation & Beglar, 2007), which has also been shown to correlate with overall proficiency (e.g., Beglar, 2010). When I compare the groups in VST scores, the difference is far from significant: $t(37) = -0.553$, $p = 0.584$. I chose to report the LexTALE instead of the VST simply because the LexTALE appears to be more commonly used in psycholinguistic

Textual Measures. Three textual measures were used, namely control essay scores, measures of productive lexical knowledge, and normed errors. These are discussed below.

Control Essay Score. Participants in both groups (Timed CW and SW) wrote a 300- to 400-word argumentative essay that served as a control essay and as a measure of writing proficiency. Participants were allowed 60 min to complete the task. Its design and scoring were parallel to the essays used in the treatments, but without incorporating any pre-specified keywords (see Appendix G for the control and treatment essays). Based on the TOEFL independent-essay criteria (ETS, 2020), the essays were scored holistically, from 1 (worst) to 5 (best) (see Leńko-Szymańska, 2020, pp. 73–77 for a discussion on holistic scoring and on the TOEFL scoring system). The essays were scored by two trained raters, but if scores had differed by more than one point (e.g., scores of 3 and 5), a third scorer would have acted as an adjudicator. However, this was not necessary. The interrater reliability was Pearson's $r = 0.771$, $p < 0.001$. This rate is comparable to the rate reported by Crossley et al. (2016), i.e., $r = 0.79$, and higher than the rates reported by Kellogg (1990), namely $r = 0.41$ and $r = 0.46$. Each rater, the author of this book and a PhD candidate in applied linguistics, has over 12 years of experience teaching English, including the teaching of writing and exam preparation courses. Of note, the raters were familiar with the research questions. The TOEFL criteria was studied and discussed, and several similar TOEFL independent essays, written by participants of similar proficiency, were scored separately and discussed by the raters prior to the experiment proper. The final score was the average of both scores and indicated no difference between the Timed CW and SW treatments, $t(37) = -0.750$, $p = 0.458$.

Productive Lexical Knowledge. Before deriving the productive lexical knowledge measures from the control essays, participants' compositions needed to be formatted. Following Meara and Miralpeix (2017) and Miralpeix (2006), learners' essays were edited in the following ways. Spelling was normalized to American English, minor spelling mistakes were corrected, and lexical inventions were removed. Also, proper names and hyphenated words were joined so the software could count them as one word (e.g., Abraham_Lincoln; e-mail → email). Finally, punctuation marks were deleted, non-complete words were completed (e.g., lab → laboratory), and contractions were written in full (e.g., don't → do not).

Two measures of productive lexical knowledge were derived from the control compositions: D and word frequency. The first one, D (Malvern & Richards, 2002), measures lexical diversity in production, with higher scores indicating larger variation in the lexical items produced by participants. D has been shown to correlate significantly with linguistic proficiency (e.g., Crossley et al., 2011a, 2011b; Jarvis, 2002; Yu, 2009). D was obtained using McNamara et al. (2014) Coh-Metrix (www.cohmetrix.com). The D score measures type/token ratio while controlling for effects of text length (see McCarthy & Jarvis, 2007). D takes 15 samples of 100 words from the text, content and function words, and computes a mean type/token ratio for

research than the VST. Given the above, I believe that the minor statistical difference in LexTALE scores between the groups may be inconsequential to the final results reported.

each sample. The samples start with 35 and gradually increase to 50 words. Because the samples differ in each calculation, D scores may differ slightly per computation (see Meara & Miralpeix, 2018 for more details). Consequently, and following recommendation by Malvern and Richards (2002), the final D score was the mean of three D scores. The second measure, word frequency, was calculated with the Tool for the Automatic Analysis of Lexical Sophistication (TAALES version 2.2.; Kyle & Crossley, 2015) using SUBTLEXus (Brysbaert & New, 2009). Unlike D, frequency measures focus on word difficulty (Daller et al., 2003; Vermeer, 2000), and lower scores indicate the use of less frequent and hence of more advanced vocabulary. Frequency was calculated by TAALES as the sum of frequency scores of each content word (log-transformed to facilitate statistical calculations) divided by the total number of content words in the essay. Word frequency has also been found to accurately differentiate between proficiency levels (e.g., Crossley et al., 2011a, 2011b; Zareva et al., 2005). Neither D scores, $t(37) = -1.337$, $p = 0.189$, nor frequency scores, $t(37) = 1.181$, $p = 0.245$, evinced any significant difference between the Timed CW and SW groups.

Normed Errors. The fourth textual measure assessed learners' productive accuracy in the control essay (and other essays written by the Timed CW group; see below). Every instance of error increased a participant's error score by one point, and the sum of points was the participant's score. All types of error were counted and given the same weight. This marking scheme follows Knoch et al. (2014, 2015), Ruiz-Funes (2015), and Ellis and Yuan (2004) with the following exceptions: unlike in Knoch et al. (2014, 2015), spelling errors were included here, except for those that were clearly typos; also, differently from Ruiz-Funes (2015) and Ellis and Yuan (2004), errors in punctuation (e.g., comma use) were also counted. The score was the average from the scores obtained by two independent raters (interrater reliability, Pearson's $r = 0.877$, $p < 0.001$). Both groups produced a similar number of errors, $t(37) = 0.364$, $p = 0.718$, again evincing no difference in proficiency between Timed CW and SW.

6.2.4 Task-Performance Measures of Cognitive Load

In research, cognitive load has been measured in several ways, including task-performance measurements (Klepsch et al., 2017; Paas et al., 2003). In this study, three such measurements were adopted—i.e., Scores, Errors, and words per minute (WPM)—to compare the control, treatment essays and to detect signs of increased cognitive load (i.e., to answer RQ2). It is believed that a decrease in one or more of these measures may indicate an increase in cognitive load (e.g., Klepsch et al., 2017; Lee, 2018, 2019; Paas et al., 2003; Robinson, 2001; Ruiz-Funes, 2014; Skehan, 2003, 2009, 2014). The first two variables, holistic essay scores (Scores) and normed errors (Errors), were calculated for the treatment essays following a similar procedure to the one used for the control essay (see Sect. 6.2.3). The third variable, the number of

words per minute produced by participants (WPM), a measure of fluency, was also calculated for all essays.

6.2.5 *Instruments*

Keywords. The 20 keywords to be learnt by the students were all selected from Coxhead's (2000) Academic Word List (AWL). More specifically, the keywords were chosen based on Silva and Otwinowska's (2019) research on academic vocabulary (reported in Chap. 5). In this research, the knowledge of 405 AWL items was measured among 106 Polish learners majoring in English. The participants in the current study were also Polish, also English majors of similar proficiency in English, so using these already-tested keywords was advisable. The process of selection of the keywords was as follows.

In a random sample of 30 tests from Silva and Otwinowska's (2019) participants, the number of mistakes per keyword was counted and the average obtained. Only the words with a number of mistakes around this average were selected. This helped avoid keywords that were too easy or difficult, thus excluding the possibility of obtaining a ceiling or floor effect, respectively. Then, two Polish-English bilingual judges and one Ukrainian-Russian-English trilingual judge ensured that none of the words were Polish-English, Ukrainian-English, or Russian-English cognates and that all words were lexicalised into the L2. This means that only words that described an existing Polish/Ukrainian/Russian concept and were thus directly translatable into these languages were chosen (see Chen & Truscott, 2010; Pellicer-Sánchez & Schmitt, 2010). Twenty keywords were then selected and divided into two sets of 10 words (see Table 6.2 for the keyword characteristics).

The comparability of words in Sets A and B was first ensured by matching the words for frequency in the academic section of the Corpus of Contemporary American English (COCA; Davies, 2012), concreteness (Brysbaert et al., 2014), part of speech and length (number of letters). Independent samples t-tests were run on the scores for each criterion to ensure similarity between the sets (see Table 6.2). Also, care was taken to divide equally between the two sets words with morphological clues to their meaning: These were "ongoing" (Set A) and "reinforce" (Set B). Finally, to

Table 6.2 Descriptive statistics and t-tests for keywords

Set	Frequency		Concreteness		Length		No. of items per part of speech		
	M	*SD*	*M*	*SD*	*M*	*SD*	Nouns	Verbs	Adjectives
A	8.57	0.47	2.04	0.43	9.3	2.0	4	3	3
B	8.66	0.69	1.89	0.26	9.1	1.52	4	3	3
$t(18)$ p value	−0.337 0.740		0.965 0.347		0.251 0.804				

( ) **Insight** (noun): (The ability to have) a clear, deep, and sometimes sudden understanding of a complicated problem or situation:

- *It was an interesting book, full of fascinating insights **into** human relationships.*
- *His book offers some fresh insights **into** the events leading up to the war.*

Fig. 6.1 Example of keyword from glossary (Set A)

ensure that the difference between Sets A and B did not distort results, the lexical items were included in the main statistical analyses (see Analysis) as random effects. This is because the different keywords are likely to influence learning differently; therefore, including the lexical items as a random effect should decrease overall error variance and increase power (Hajduk, 2019; Meteyard & Davies, 2020). The final sets were as follows: Set A: "apparent", "implicit", "ongoing", "constraint", "acquisition", "insight", "paradigm", "incorporate", "constitute", and "differentiate"; Set B: "qualitative", "inherent", "affective", "assessment", "variability", "phenomenon", "validity", "reinforce", "derive" and "facilitate".

Glossary. The keywords were presented to all participants in the form of glossaries to assist with task performance. All participants in CW and SW were given the glossary (see Appendix H) right before performing each of their writing tasks. All information in the glossary was taken from dictionaries for advanced learners (Cambridge University Press, 2020; Pearson, 2020). One definition and two examples were given for each keyword. Three keywords in each vocabulary set were given two definitions, with one example provided for each different meaning. Common prepositions following keywords were boldened, so learners had better access to word usage. This is one example of an entry in one of the glossaries (Fig. 6.1).

Test of Lexical Learning (Pretest and Posttest). The tests of lexical learning consisted of an adapted version of Wesche and Paribakht's (1996) Vocabulary Knowledge Scale (VKS). The VKS adopts a measurement scale from no knowledge (option I) to the ability to recognize a word (option II), through learners' ability to recall a word's meaning with different levels of confidence (options III and IV) to productive knowledge (option V). Put differently, the VKS measures the quality of lexical knowledge in a receptive-productive continuum, or depth of knowledge (see Schmitt, 2014; Zhang & Koda, 2017). There were two versions of the same VKS test, one for each set of 10 keywords. Figure 6.2. shows one test item (see Appendix I for the whole test). The last item (VI) is a free association test; it is effectively a separate measure and was scored separately. Number of associations also assesses depth (see Zareva et al., 2005 for a similar design); they tap into depth of knowledge by measuring the development of learners' lexical network as it is restructured to accommodate novel words (Read, 2004).

In the VKS, each lexical item received a maximum of 6 points. Participants choosing options I or II obtained scores of 1 or 2, respectively; selecting options III or IV gave scores 3 or 4, if the answer was correct, or 2 if it was wrong. Option V could yield scores of 2, 4, 5 or 6, depending on the accuracy of the sentence produced. A score of 6 was obtained if the keyword was used grammatically and semantically

1. *Apparent (adjective)*
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
VI. I associate this word with _____, _____, _____, _____.

Fig. 6.2 A sample VKS item from Set A

accurately, even if there were errors in other parts of the sentence. The score was reduced to 5 if the keyword was used in the wrong grammatical category (e.g., as an adjective instead of a noun), with the wrong derivation or conjugation (e.g., wrong plural or wrong past participle), or with the wrong preposition (e.g., "an insight on"). Option V received a score of 4 if the meaning of the keyword was demonstrated accurately, but the word was not used appropriately in the sentence, and a score of 2 if the meaning was inaccurate. Figure 6.3 illustrates the scoring procedure.

In option VI, the free association test, participants could produce up to four words they associate with the keyword. The associates may be connected syntagmatically (i.e., frequently occurring together or in collocation with another word, as in *cat* → *purr* or *car* → *drive*) and paradigmatically, including coordinates (*cat* → *dog*), synonyms (*cat* → *feline*), antonyms, superordinates (*animal* as a superordinate of *cat*) and subordinates, such as *Siamese* for *cat* (see Wolter, 2001 for a discussion on association types). Each correct associate received one point, totaling a possible four points per VKS item. For an associate to be given a point, participants needed to have scored at least 3 in the respective VKS item. The VKS and free association test (each yielding a separate result) were also scored by two independent raters. Pearson's r interrater reliability scores were very high ($p < 0.001$). For Set A, the

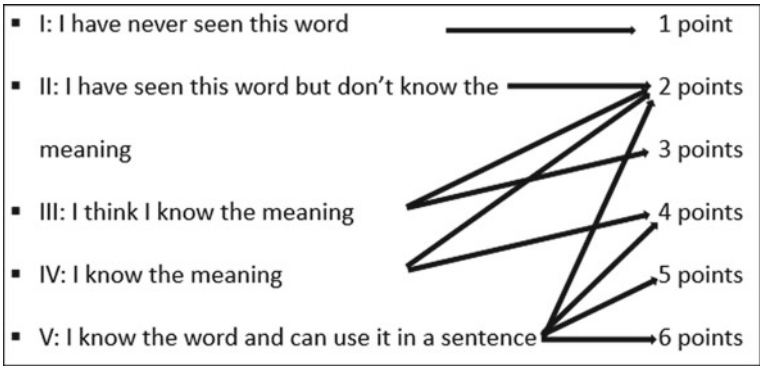

Fig. 6.3 VKS scoring procedure

VKS correlation was 0.941, and the association test correlation was 0.860; for Set B, the values were 0.937 and 0.875, respectively.

Although both the VKS and the association tests measure gains in depth, the VKS items may be scored differently to also account for gains in breadth (i.e., number of words known). This is because any word that was previously unknown (scores 1 or 2) and received a score of 3 or higher in the posttest can be said to represent a gain in breadth. By contrast, changes from receptive knowledge (scores 3 or 4) to productive knowledge (scores 5 or 6) illustrate a gain in depth.

6.2.6 Design

Figure 6.4 shows the overall research design in some detail. First, all participants wrote the control essay (without keywords). Five weeks later, the four intact groups sat the pretests and the receptive lexical test (LexTALE; Lemhöfer & Broersma, 2012). Participants were informed that the purpose of the pretest and the LexTALE was to measure their proficiency in English. The LexTALE also acted as a cognitively demanding task immediately following the pretests, so the keywords could be flushed from participants' memories, as recommended by Schmitt (2010). Then, at one-week intervals, the Timed CW and SW treatment groups performed their respective tasks (see below). For the treatments and posttests, vocabulary Sets A and B were counterbalanced between the 4 intact classes. Two weeks after each condition, the participants completed the posttests, which measured lexical retention. After the Timed CW and SW treatments and after posttest 1, learners continued with their regular classes; following posttest 2, participants completed the questionnaire.

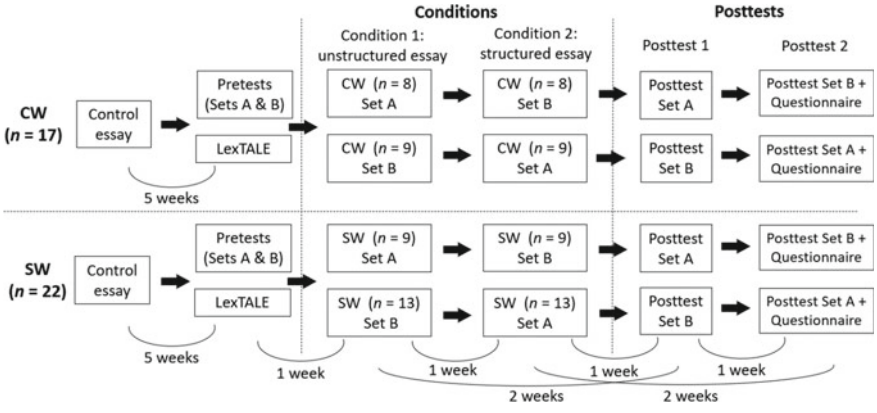

Fig. 6.4 Illustration of research design and procedure

6.2.7 Treatment Groups

The Sentence-writing (SW) Treatment Group. The sentences were written in two different 60-min sessions, ten sentences per session, with a different vocabulary set (A or B) counterbalanced (see Table 6.2). Participants were given glossaries and instructed to write a total of 20 grammatically correct sentences with at least 10 words each, one sentence per keyword. This was done to keep participants from producing overly simplistic sentences (Kim, 2008; Zou, 2017). The keywords needed to be used correctly, and the part of speech specified in the glossary could not be changed. For instance, the noun "insight" could not be used in its adjectival form "insightful". In addition, to avoid long-enough sentences that show no understanding of the keywords, such as "I really have no idea what the word 'constraint' means", I controlled the sentences by periodically supervising learners' performance.

The Composition-writing (Timed CW) Treatment Group. Two argumentative essays similar to the control essay in type, length (300–400 words), time limit (60 min), and scoring (two independent raters) were written in two different sessions. This time, however, participants had to include 10 keywords in each essay, using one vocabulary set per essay. The instructions for the first condition, the unstructured essay, prompted learners to focus on the keywords and disregard text quality (see Appendix G for the CW tasks). This was done to try to simulate occasions when learners decide to focus on keyword use (the secondary task) rather than on the writing proper (the primary task). I believe that this task would effectively replicate Zou's (2017) study (see Sect. 2.6) In this study, her lower-proficiency learners managed to incorporate the keywords, but text quality was compromised, which resulted in an increase in learning following Timed CW. The instructions for the second condition, structured essay, highlighted the importance of composing a well-structured text, which more closely replicates the expected quality of argumentative essays at university (see Appendix J for two sample essays produced by participants). However, learners disregarded the instructions and produced two structured essays instead (see Sect. 6.3 Analysis).

The topics selected were deemed familiar to learners in that age group and were chosen from a list of TOEFL topics for independent essays (ETS, 2020). Because the TOEFL is extensively piloted and internationally recognized and administered, the topics were considered reliable and valid. Each essay (control, unstructured, and structured) had a different topic; nevertheless, all topics required a similar structure: Learners needed to express agreement or disagreement with a given statement and to support their views with examples and clear argumentation.

When designing the CW task, the option of using the same topic for both essays (three, including the control essay) was considered. In fact, this is often done in the literature (e.g., Shaw & Liu, 1998; Storch & Tapper, 2009) to avoid the possibility of different topics affecting writing quality differently, thus distorting the results. Still, as acknowledged by Shaw and Liu (1998), rewriting on the same topic has a facilitating effect. Indeed, Skehan (2003, 2014), Foster & Skehan (1996) has shown that task repetition results in task familiarization, allowing learners to focus more on

complexity and accuracy. As explained by Skehan (2014, p. 217), "memory traces from the first performance are still having an effect and facilitate the subsequent processing". Since increase in cognitive load is central to the current investigation, it was decided to choose different topics of comparable structure. Unfortunately, the different topics made it impossible to score blindly, as the available raters knew the topics of the three essay conditions.

The unstructured and structured essays were rated similarly to the control essay: Based on the TOEFL independent-essay criteria (ETS, 2020), the essays were given holistic scores ranging from 1 (worst) to 5 (best) by two trained raters, and the final score was the average of both scores. A third rater was available to act as adjudicator in case the scores differed by more than one point, but this was not necessary. Pearson's r interrater reliability correlation was 0.832 ($p < 0.001$) for scores for the unstructured essay and 0.922 ($p < 0.001$) for the structured essay, representing very high interrater reliability.

6.2.8 Procedures

The glossaries, test of lexical learning, CW and SW tasks were piloted with a focus group of 17 graduate students. In the group interviews, these learners drew attention to inaccuracies and occasional lack of clarity. All these were subsequently corrected for the study proper. All tasks were conducted during regular 90-min classes at the beginning of participants' first academic year. The pretests, posttests, glossaries, and the questionnaire were provided to participants in paper form and collected after the treatment. The three essays (control, unstructured and structured), the sentence writing and the receptive lexical test (LexTALE) were conducted on laptops. Participants had no access to the internet, except for the LexTALE. Also, the proofreading tool in the word processor was turned off. The author of this book administered and monitored all tasks closely.

To hide the true purpose of the quasi-experiment, Timed CW participants were told that the aim of the study was to measure their ability to write argumentative essays under different conditions. SW learners were told that the purpose of the quasi-experiment was to measure their sentence-writing speed while being obliged to follow certain conditions. For them, the condition was to incorporate keywords in sentences, but they did not know whether other groups were writing sentences under different conditions.

6.3 Analysis

The VKS and association tests, used to measure and compare lexical learning following SW and Timed CW (i.e., to answer RQ1), yielded three outcome variables: VKS_6, VKS_3 and Association score. The first variable (VKS_6) is an ordinal variable with levels ranging from 1 (no knowledge) to 6 (full productive knowledge).

The second variable (VKS_3), also derived from the VKS and also ordinal, had three levels. Level 1 was the combination of VKS scores 1 and 2 (i.e., no knowledge or the ability to recognize word form), level 2 combined scores 3 and 4 (i.e., receptive knowledge of meaning with different levels of confidence), and level 3 joined scores 5 and 6, representing productive knowledge (see Sect. 6.2.5 for details on VKS scoring). Finally, the third variable (Association), was derived from the association test. This was the number of words that participants associated with each keyword, ranging from 0 to 4. Association scores followed a Poisson distribution and were analyzed as count variables (see Heck et al., 2012).

Generalized linear mixed models (GLMMs) were used to analyze VKS_6, VKS_3, and Association. For details on how random effects were entered in the models (i.e., following a minimal-to-maximal-that-improves-fit process) and on how to choose the model with best fit (i.e., using AIC), see Sect. 4.5.1. Two categorical variables were entered as random effects: Participants (the 39 participants) and Items (the 20 keywords). The categorical variables Set (vocabulary Sets A and B) and Class (the four intact classes from which the participants were selected) were not entered as random effects because they had fewer than five levels (Hajduk, 2019). Still, Participants and Items sufficed to account for the variation between Set and Class. The fixed effects were Group (SW and Timed CW), Time (pretest and posttest), Condition (unstructured and structured), and all possible interactions. Because Condition had only two levels, it was defined as a continuous variable (see Field, 2017) and thus entered as a covariate. The same learners wrote the unstructured and structured essays (see below), so the data are not independent. Therefore, Condition as a covariate solved the problem of non-independence of observations. Appendix K illustrates the full process of creation of the GLM models.

To detect signs of increased cognitive load in Timed CW, the within-subject variables Scores (measuring overall essay quality), Errors (measuring writing accuracy), and WPM (measuring writing fluency) were obtained from each of the three essays. As discussed in Chap. 3, it was assumed that a decrease in the scores of one or more of these measurements might indicate an increase in task-induced cognitive load. All three variables were normally distributed (z scores for skewness < 1.96). Importantly, data analysis showed that learners failed to follow task instructions, instead treating the unstructured and structured essays equally (i.e., as structured essays). Paired-sample t-tests found no difference between the unstructured and structured essays in Scores, $t(16) = -0.575$, $p = 0.573$, Errors, $t(16) = -1.507$, $p = 0.151$, and WPM, $t(16) = -0.101$, $p = 0.921$, which attest to the similarity in production between both essays. Consequently, the data were collapsed when calculating lexical learning. The raw data can be found online (see Silva et al., 2020).

6.4 Results

Before answering the research questions, it is important to note that all three variables obtained from the VKS and association tests—i.e., VKS_6 (all six levels), VKS_3

(levels 1–2, 3–4, 5–6 combined), and Association—registered lexical gains between the pretest and posttest ($p < 0.001$). This can be seen by referring to Time (i.e., pretest–posttest lexical gains) in Tables 6.4, 6.6, and 6.8 in the subsections below.

6.4.1 *Results for Tests Measuring Lexical Knowledge*

RQ1 asked whether Polish EFL learners acquire academic vocabulary to a similar degree after performing Timed CW and SW tasks. To obtain answers, I constructed three generalized linear mixed models (GLMM), one for each dependent variable: VKS_6, VKS_3, and Association (see Sect. 6.3). I discuss them separately below.

Results for VKS_6. The proportion of scores for VKS_6 in the Timed CW and SW groups is shown in Table 6.3 and illustrated in Fig. 6.5. It is noteworthy that

Table 6.3 Proportion of VKS_6 scores for Timed CW and SW in the pretest and posttest

	VKS score											
	1		2		3		4		5		6	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Timed CW ($n = 17$)	0.11	0.00	0.55	0.57	0.02	0.00	0.03	0.02	0.02	0.02	0.27	0.39
SW ($n = 22$)	0.10	0.02	0.53	0.42	0.04	0.03	0.04	0.03	0.02	0.03	0.27	0.47

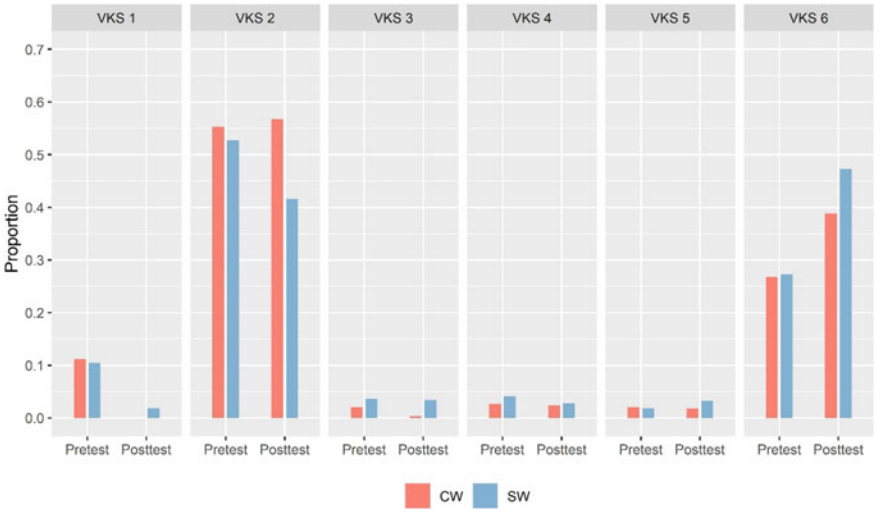

Fig. 6.5 Proportion of VKS_6 scores in the pretest and posttest for Timed CW and SW

SW learners showed a higher increase between pretest and posttests in level 6 scores (full productive knowledge) than Timed CW participants: 20% and 12%, respectively. This was mostly attributed to a decrease in level 2 scores (ability to recognize word form) among SW learners, from 53 to 42%.

Table 6.4 presents the results of the GLMM for VKS_6. The Group * Time interactions show a difference in learning between the SW and Timed CW groups ($p = 0.025$). The odds ratios of the interactions indicate that in the posttest, Timed CW is 46% more likely to score 1 in the VKS (versus all the other scores combined) than SW. Also, Timed CW participants are 32% less likely to score 6 in the posttest

Table 6.4 Fixed and random effect estimates for VKS_6

	Estimate (Std. error)	t	p	Odds ratio	95% CI Odds ratio
Fixed effects					
Intercept	–	–	–	–	–
Group	0.08 (0.26)	0.317	0.752	1.08	[0.66, 1.79]
Time	−1.01 (0.12)	−8.430	<0.001	0.36	[0.29, 0.46]
Condition	−0.03 (0.12)	−0.225	0.822	0.97	[0.78, 1.22]
Group * Time	0.38 (0.16)	2.423	0.025	1.46	[1.05, 2.03]
Group * Time[a]	−0.38 (0.16)	−2.423	0.025	0.68	[0.49, 0.95]
Group * Condition	0.04 (0.21)	0.176	0.860	1.04	[0.69, 1.56]
Time * Condition	0.26 (0.15)	1.765	0.078	1.30	[0.97, 1.74]
Group * Time * Condition	−0.21 (0.21)	−1.026	0.305	0.81	[0.54, 1.21]

	Variance (Std. error)	95% CI	Wald Z	p	Intraclass correlation (ICC)
Random effects					
Residual	3.29	–	–	–	–
Participants (Intercept)[b]	0.285 (0.073)	[0.172, 0.472]	3.897	<0.001	0.0613
TimelParticipants (intercept-slope correlation)	0.302 (0.314)	[−0.350, 0.756]	0.962	0.336	0.0650
Items (Intercept)[c]	0.769 (0.095)	[0.603, 0.980]	8.066	<0.001	0.1655

Note Number of data points = 1560; items = 40; participants = 39. Probability distribution: multinomial; link function: cumulative negative log–log (for better fit, see Heck et al., 2012, p. 320). Reference categories (descending) = SW, Pretest, Unstructured, and VKS score 1

[a]Reference category of target: VKS level 6. The link function changed to cumulative complementary log–log due to the change in reference (see IBM SPSS, 2020)

[b]Covariance structure = first-order autoregressive (AR1)

[c]Covariance structure: variance component

Table 6.5 Proportion VKS_3 scores for CW and SW in the pretest and posttest

	VKS score					
	1		2		3	
	Pre	Post	Pre	Post	Pre	Post
Timed CW ($n = 17$)	0.66	0.57	0.05	0.03	0.29	0.41
SW ($n = 22$)	0.63	0.43	0.08	0.06	0.29	0.50

(versus all the other scores combined) than SW learners.[2] Importantly, most of the knowledge gained was in breadth, not in depth, since most of the change occurred between levels 1 or 2 and 6 (i.e., from not knowing the meaning to productive word knowledge).

Of note, the intraclass correlation (ICC) shows the variance in the data explained by the random effects after the variance explained by the fixed effects (Carson & Beeson, 2013; Hajduk, 2019; Heck et al., 2012). This means, for example, that the intercept for Participants accounted for 6.13% (ICC = 0.0613) and the intercept for Items accounted for 16.55% (ICC = 0.1655) of the variance in VKS_6 scores that is not explained by the fixed effects. Put differently, VKS_6 scores varied considerably among participants and keywords, together explaining over 22% of the variation in scores that would have gone unexplained had random effects not been included. This extra variance explained means less statistical error unaccounted for, therefore increasing the statistical power and reliability of the analysis.

Results for VKS_3. A different way to answer RQ1 is by looking at how students performed on different types of word knowledge, represented by the variable VKS_3 (level 1: no knowledge or ability to recognize word-form; level 2: receptive knowledge of meaning to different degrees of certainty; level 3: productive knowledge). An increase from level 1 to levels 2 or 3 in the posttest represents a gain in breadth of lexical knowledge while an increase from level 2 to level 3 (i.e., receptive to productive knowledge) may indicate gains in depth. The proportion of scores for VKS_3 is depicted in Table 6.5 and Fig. 6.6.

Similarly to VKS_6, most of the knowledge gained was in breadth, not in depth, since scores mostly changed from levels 1 to 3. Also, the proportions for VKS_3 show more learning for SW than Timed CW. First, level 1 scores decreased by 20 points in SW and only 9 points in Timed CW. Second, level 3 increased 9 points more in SW than in Timed CW. The GLMM for VKS_3, presented in Table 6.6, confirms significantly higher lexical gains for SW than for Timed CW.

[2] An odds ratio of 1.46 indicates a proportion of 46%. Since the value in the column "estimate" is positive, it means "46% more likely". An odds ratio of 0.68 indicates a proportion of 32% (1 – 0.68 = 0.32). Since the value in the column "estimate" is negative, it means "32% less likely". Where the odds ratio is 46%, the reference categories are SW, pretest, and score 1. Therefore, the value given compares CW to SW in the posttest with reference to score 1, thus the reported results. Where the odds ratio is 32%, the reference category for the dependent variable VKS_6 changes to score 6, as indicated below the table.

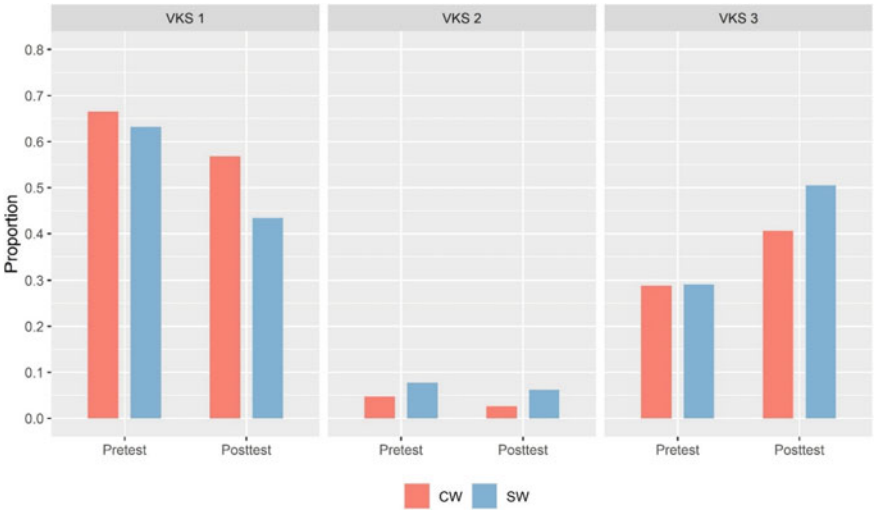

Fig. 6.6 Proportion of VKS_3 scores in the pretest and posttest for CW and SW

The odds ratios for the Group * Time interactions show that Timed CW learners are 65% more likely to score 1 (versus scores 2 and 3 combined) in the posttest than SW. Also, Timed CW learners have 39% lower chances of scoring a 3 (versus the combination of scores 1 and 2) than SW participants in the posttest. Importantly, the Time * Condition reached significance ($p = 0.045$). Given the odds ratio, this means that writing in the second condition—i.e., the structured argumentative essay (Timed CW) or the second set of 10 sentences (SW)—made participants 28% more likely to score 1 in the posttest than when writing in the first condition. This means that although participants treated both unstructured and structured essays as structured essays (see Sect. 6.3), there is some indication that unstructured essays yielded slightly more learning than structured essays. Regarding the ICC, the results show that the random effects combined explained 33.98% of the variance left unexplained by the fixed effects.

Results for Association. The descriptive statistics for Association are shown in Table 6.7. It appears that SW generated more learning than Timed CW since there was a higher increase in mean count for Association scores in SW. The GLMM, shown in Table 6.8, corroborates this.

The odds ratio for Time shows that participants had a 2.51 higher chance of achieving a higher score in Association in the posttest than in the pretest. This shows that both Timed CW and SW tasks yielded significant gains in depth of knowledge. The increase in scores was significantly higher for SW than for Timed CW, as shown by the Group * Time interaction ($p = 0.034$). Here, the odds ratio shows that Timed CW participants are 40% less likely than SW learners to achieve higher scores in Association in the posttest. As for the ICC, all random effects put together explain 40.42% of the variance left over from the fixed effects.

Table 6.6 Fixed and random effect estimates for VKS_3

	Estimate (Std. error)	t	p	Odds ratio	95% CI Odds ratio
Fixed effects					
Intercept	–	–	–	–	–
Group	0.05 (0.23)	0.220	0.826	1.05	[0.67, 1.64]
Time	−0.91 (0.11)	−8.267	<0.001	0.40	[0.32, 0.50]
Condition	−0.07 (0.11)	−0.704	0.482	0.93	[0.76, 1.14]
Group * Time	0.50 (0.15)	3.371	0.003	1.65	[1.21, 2.24]
Group * Time[a]	−0.50 (0.15)	−3.371	0.003	0.61	[0.45, 0.82]
Group * Condition	0.05 (0.19)	0.266	0.790	1.05	[0.72, 1.54]
Time * Condition	0.25 (0.12)	2.009	0.045	1.28	[1.01, 1.62]
Group * Time * Condition	−0.16 (0.18)	−0.882	0.378	0.85	[0.59, 1.22]

	Variance (Std. error)	95% CI	Wald Z	p	ICC
Random effects					
Residual	3.29	–	–	–	–
Participants (Intercept)[b]	0.207 (0.061)	[0.115, 0.370]	3.366	<0.001	0.0415
Time|Participants (intercept-slope correlation)	0.977 (0.295)	[−1.000, 1.000]	3.318	<0.001	0.1961
Items (Intercept)[c]	0.509 (0.077)	[0.378, 0.686]	6.573	<0.001	0.1021

Note Number of data points = 1560; items = 40; participants = 39. Probability distribution: multinomial; link function: cumulative negative log–log (for better fit, see Heck et al., 2012, p. 320). Reference categories (descending) = SW, Pretest, Unstructured, and VKS score 1
[a]Reference category of target: VKS level 6. The link function changed to cumulative complementary log–log due to the change in reference (see IBM SPSS, 2020)
[b]Covariance structure = first-order autoregressive (AR1)
[c]Covariance structure: variance component

Table 6.7 Descriptive statistics for Association for Timed CW and SW in the pretest and posttest

	Association scores (max. = 4)					
	Pretest			Posttest		
	Mean	*SD*	*Mdn*	*Mean*	*SD*	*Mdn*
Timed CW ($n = 17$)	0.54	1.03	0	0.78	1.09	0
SW ($n = 22$)	0.53	1.03	0	1.00	1.17	1

To answer RQ1, the three GLMMs (VKS_6, VKS_3, Association) showed significant more learning following SW than Timed CW. This was true in terms of breadth (as measured by VKS_6 and VKS_3) and depth of knowledge (as measured by Association). This finding will be discussed in more detail in Chap. 8.

Table 6.8 Fixed and random effect estimates for final model of Association

	Estimate (Std. error)	t	p	Odds ratio	95% CI Odds ratio
Fixed effects					
Intercept	−1.12 (0.24)	−4.740	<0.001	0.33	[0.20, 0.52]
Group	0.14 (0.36)	0.381	0.704	1.15	[0.56, 2.33]
Time	0.92 (0.16)	5.857	<0.001	2.51	[1.81, 3.48]
Condition	0.13 (0.19)	0.680	0.497	1.13	[0.79, 1.63]
Group * Time	−0.52 (0.22)	−2.320	0.034	0.60	[0.37, 0.96]
Group * Condition	−0.31 (0.26)	−1.174	0.241	0.74	[0.44, 1.23]
Time * Condition	−0.27 (0.14)	−1.853	0.065	0.77	[0.58, 1.02]
Group * Time * Condition	0.36 (0.19)	1.843	0.067	1.43	[0.98, 2.09]

	Variance (Std. error)	95% CI	Wald Z	p	ICC
Random effects					
Residual	1.00	–	–	–	
Participants (Intercept)	0.565 (0.160)	[0.324, 0.985]	3.528	<0.001	0.1023
Time|Participants (intercept-slope correlation)	−0.318 (0.248)	[−0.702, 0.209]	−1.282	0.200	0.0576
Items (Intercept)	0.614 (0.111)	[0.431, 0.874]	5.546	<0.001	0.1112
Time|Items (intercept-slope correlation)	−0.734 (0.081)	[−0.858, −0.532]	−9.038	<0.001	0.1330

Note Number of data points = 1560; items = 40; participants = 39. Reference category for predictors: SW, Pretest, Unstructured. Reference category for target: ascending. Probability distribution: Poisson; link function: log. Covariance structure = AR1

6.4.2 Results for Task-Performance Variables

RQ2 asked whether using keywords in essays affect their overall quality, accuracy, and fluency, which may be a sign of increased cognitive load. To answer this question, the within-subject textual variables Scores, Errors, and WPM were analyzed.

Table 6.9 Descriptive statistics for textual measures

Essay conditions ($n = 17$)	Essay scores		Normed errors		WPM	
	M	*SD*	*M*	*SD*	*M*	*SD*
Control	3.71	0.97	70.03	33.55	7.02	1.28
Unstructured	3.38	1.07	78.93	33.05	5.81	1.17
Structured	3.62	1.23	91.53	44.77	5.83	1.40

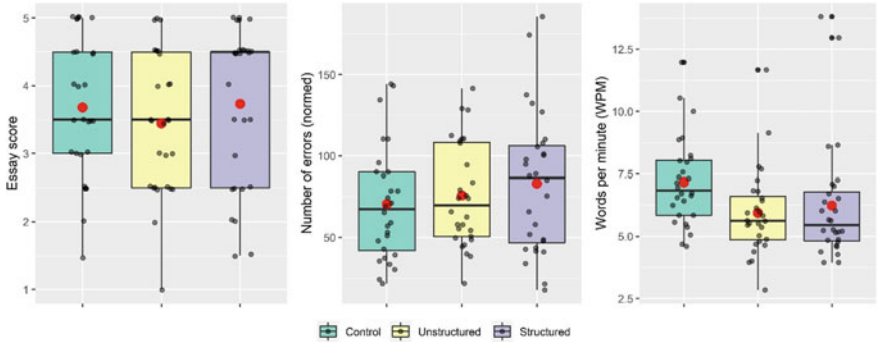

Fig. 6.7 Boxplots comparing essays for each textual measure. The red dots represent the means

The descriptive statistics are shown in Table 6.9. The paired samples t-tests found no difference in Scores between the control śand unstructured essays, $t(16) = 1.428$, $p = 0.173$, control and structured essays, $t(16) = 0.347$, $p = 0.733$, unstructured and structured essays, $t(16) = -1.054$, $p = 0.308$. This shows that using pre-specified keywords when writing did not reduce the overall quality of the essays and that unstructured and structured essay conditions had the same quality. Regarding Errors, the t-tests found a significant difference between the control and unstructured essays, $t(16) = -2.227$, $p = 0.041$, 95% CI $[-17.36, -0.42]$, $r = 0.49$, representing a medium effect size, and the control and structured essays, $t(16) = -2.811$, $p = 0.013$, 95% CI $[-37.70, -5.28]$, $r = 0.57$, with a large effect size (Plonsky & Oswald, 2014). There was no difference in the number of errors between the unstructured and structured essays: $t(16) = -1.507$, $p = 0.151$. Finally, differences were also found in WPM between the control and unstructured, $t(16) = 3.438$, $p = 0.003$, 95% CI [0.47, 1.96], $r = 0.65$, and the control and structured essays, $t(16) = 2.874$, $p = 0.011$, 95% CI [0.31, 2.06], $r = 0.58$, with large effect sizes. Again, there was no difference in WPM between the unstructured and structured essays: $t(16) = -0.101$, $p = 0.921$. To answer RQ2, even though the three essays were qualitatively similar (i.e., similar scores), learners wrote the unstructured and structured essays with keywords more slowly, but even with the extra time, participants produced more errors in these two essays than in the control. Figure 6.7 shows a side-by-side comparison of the three essays in each textual measure.

6.5 Discussion

The current study has attempted to further our understanding of how an explicit focus on language production may facilitate the incidental acquisition of academic words. To this end, this study drew on Laufer and Hulstijn's (2001) involvement load hypothesis (ILH) and compared incidental lexical learning following sentence

writing (SW) and composition writing (Timed CW). I aimed to find out whether Polish advanced-level learners of English acquire and retain academic words to a similar degree after writing sentences and 60-min timed argumentative essays.

When answering RQ1, the statistical analyses found strong evidence that using keywords when writing sentences generates more learning than when writing argumentative essays, both in breadth and depth of knowledge. This is rather remarkable considering that SW participants took on average 19.27 min ($SD = 3.97$) to complete the task whereas Timed CW learners needed an extra 40 min ($M = 59.97, SD = 9.74$). Gains in breadth were registered when comparing learning in VKS_6 (pertaining to six levels of vocabulary knowledge) and VKS_3 (pertaining to types of vocabulary knowledge). This is because, in both cases, gains focused mostly on learning new meanings of unknown or almost completely unknown words. Learning in depth of knowledge was mainly evidenced by comparing SW and Timed CW in association scores.

These results do not appear to corroborate the predictions of the ILH (i.e., that SW and Timed CW should generate similar lexical gains), neither in terms of breadth nor depth of knowledge. The findings also run counter the results of previous studies comparing these two tasks, three of which supported the ILH (i.e., Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017) while one (Zou, 2017) found that Timed CW yielded more learning than SW. These differences in research findings will be discussed in more detail in Chap. 8, which presents the joint discussion of the results from Study 2 and 3. This is because Study 2 (this one) and Study 3 (Chap. 7) are complementary and hence an in-depth discussion is more suitable after reporting both studies.

To answer RQ2, which asked whether the need to use pre-specified keywords in essays affects the quality, accuracy, and fluency of writing, I compared the control essay (i.e., without keywords) to two treatment essays. This was done to detect signs of increased cognitive load in Timed CW and to explore how this higher pressure on attentional resources may have affected lexical learning. In this study, I adopted three task-performance measurements to detect signs of increased cognitive load (Klepsch et al., 2017; Paas et al., 2003): Scores, Errors, and WPM. The results showed that Timed CW participants needed significantly more time to compose texts with keywords than without in order to maintain similar text quality. Still, even after needing more time to produce such essays, participants made more errors in these essays than in the control. This reinforces previous findings by Lee (2018), who also showed that higher task complexity resulted in an increase in time-on-task and in a decrease in accuracy. As a result, to answer RQ2, there is some evidence to suggest that the need to incorporate keywords in essays increased task cognitive load, which may help explain why Timed CW generated less learning than SW (despite warranting far more time from learners). Again, these findings will be discussed in more detail in Chap. 8 together with findings from Study 3.

One methodological issue with this study was my failure to persuade participants to write an unstructured essay. I believe this task would have simulated Zou's (2017) CW task (see Sect. 2.6). In her study, it appears that participants treated keyword use, not essay writing, as the primary task. This is evinced by the fact that her participants

did indeed incorporate the keywords but did not write good quality essays (judging by the sample essay provided by the author). My intention with the unstructured Timed CW task was to achieve exactly this, namely, persuade participants to disregard quality and treat keyword-use as the primary task. It is possible that when learners do this, the lexical learning potential of Timed CW increases, which would explain why Zou (2017) found higher learning following Timed CW than SW.

Still, this failure to persuade learners to disregard text quality was expected. Experienced educators know that learners' task performance is often different from what was intended by teachers and task designers. In fact, it may be impossible to make learners behave in pre-determined ways, even when task instructions are clear, thorough, and strict (Takavoli, 2014). A reason for this is that, as noted by Breen (1987/2009 as cited in Manchón, 2014, p. 41), learners' interpretation of a task is shaped by their perception of themselves, the task, and the task situation. In a similar argument, Macaro (2014, p. 61) has stated that learners' response to a task "will differ according to how they interpret the goal of the task or according to the goal they set themselves in relation to the task" (see also Manchón, 2014; Nicolás Conesa et al., 2014 for a similar opinion). Consequently, in this study, it seems that learners approached both essay conditions as structured essays because this is what they had been trained to do in their Writing Practice lessons. On a positive note, participants' choice to focus on essay quality rather than on the incorporation of the keywords, thus treating essay writing as the primary task, lends support to the incidental nature of my study.

Because essay-writing was considered the primary task, participants likely allocated more attentional resources to the production of a well-structured essay than to keyword use. This increase in cognitive load of structured-essay writing may be what reduces lexical learning, to the point of equating the learning of Timed CW to the learning following SW (as in Kim's, 2008 study), or even making Timed CW less conducive to vocabulary learning than SW (as in the current study). What follows is that if the cognitive load induced by structured essays is reduced, lexical learning should increase. One way to reduce this cognitive load is not to put L2 writers under time pressure. That is, to write untimed argumentative essays. To this aim, Study 3, reported in the next chapter, compares the learning of academic words following SW, Timed CW, and Untimed CW.

6.6 Conclusion

The study reported in this chapter has compared the academic vocabulary learning potential of two types of tasks: sentence writing (SW) and timed composition (essay) writing (CW). The results showed more learning following SW than CW under time pressure both for breadth and depth of lexical knowledge. Furthermore, measures of cognitive load (writing fluency, accuracy, and overall quality) indicated that writing essays with pre-specified keywords under time pressure taxed learners' cognitive resources. This is because CW essays took longer to write but kept similar overall

quality and were less accurate than the control essays (without the need to use pre-specified keywords). One possibility then is that CW yielded less lexical learning than SW because learners' cognitive resources were overloaded by the essay-writing process. If this is the case, writing essays without time pressure may alleviate the cognitive load and potentially increase lexical learning. This possibility is explored in Chap. 7.

Funding The research was supported by grant 2019/35/N/HS2/03550 from the National Science Centre Poland awarded to the author, Breno Barreto Silva.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118.

Biber, D., Gray, B., & Poonpon, K. (2013). Pay attention to the phrasal structures: Going beyond t-units—a response to WeiWei Yang. *TESOL Quarterly*, 192–201.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.

Byrnes, H., & Manchón, R. M. (2014). Task, task performance and writing development: Advancing the constructs and the research agenda. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 267–299). John Benjamins.

Cambridge University Press. (2020). Cambridge Dictionary. Retrieved from: https://dictionary.cambridge.org/. Accessed November 24, 2020.

Carson, R. J., & Beeson, C. M. L. (2013). Crossing language barriers: Using crossed random effects modelling in psycholinguistics research. *Tutorials in Quantitative Methods for Psychology, 9*(1), 25–41.

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics, 31*(5), 693–713.

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011a). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing, 28*(4), 561–580.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2011b). Predicting the proficiency level of language learners using lexical indices. *Language Testing, 29*(2), 243–263.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197–222.

Davies, M. (2012). *Corpus of contemporary American English* (1990–2012). Retrieved from: https://www.wordfrequency.info/. Accessed November 06, 2020.

De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning, 68*(4), 906–941.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84. https://doi.org/10.1017/S0272263104261034

ETS. (2020). TOEFL iBT [online]. Retrieved from: http://www.ets.org/toefl. Accessed November 06, 2020.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299–323.

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing, 26*, 80–95.

Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique feature analysis or involvement load hypothesis: Estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research, 47*, 859–869.

Hajduk, G. K. (2019). *Introduction to linear mixed models*. Retrieved from: https://ourcodingclub.github.io/tutorials/mixed-models/#crossed. Accessed November 27, 2020.

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2012). *Multilevel modelling of categorical outcomes using IBM SPSS*. Routledge.

IBM SPSS. (2020). IBM SPSS advanced statistics 26. Retrieved from: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/26.0/en/client/Manuals/IBM_SPSS_Advanced_Statistics.pdf. Accessed November 26, 2020.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84. https://doi.org/10.1191/0265532202lt220oa

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*, 377–403.

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands. *The American Journal of Psychology, 103*(3), 327–342.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1–18. https://doi.org/10.3389/fpsyg.2017.01997

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing, 21*, 1–17.

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university. *Journal of Second Language Writing, 28*, 39–52.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and applications. *TESOL Quarterly, 49*(4), 757–786.

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

Lee, J. (2018). The effects of task complexity and L2 proficiency on L2 written performance. *The Journal of Asia TEFL, 5*(4), 945–958. https://doi.org/10.18823/asiatefl.2018.15.4.4.945

Lee, J. (2019). Time-on-task as a measure of cognitive load in TBLT. *The Journal of Asia TEFL, 16*(3), 958–969. https://doi.org/10.18823/asiatefl.2019.16.3.12.958

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Leńko-Szymańska, A. (2020). *Defining and assessing lexical proficiency*. Routledge.

Macaro, E. (2014). Reframing task performance: The relationship between tasks, strategic behaviour, and linguistic knowledge in writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 53–77). John Benjamins.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Manchón, R. M. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 27–52). John Benjamins.

Manchón, R. M., & Williams, J. (2016). Introduction: SLA-L2 writing interfaces in historical perspective. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 567–586). De Gruyter.

McCarthy, P. M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–488.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

Meara, P., & Miralpeix, I. (2017). *Tools for researching vocabulary*. Multilingual Matters.

Meara, P., Miralpeix. (2018). *D_Tools*. Retrieved from: http://www.lognostics.co.uk/tools/D_Tools/D_Tools.htm. Accessed November 27, 2020.

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 1–22. https://doi.org/10.1016/j.jml.2020.104092

Miralpeix, I. (2006). Age and vocabulary acquisition in English as a foreign language (EFL). In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 89–106). Multilingual Matters.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nicolás-Conesa, F., Roca de Larios, J., & Coyle, Y. (2014). Development of EFL students' mental models of writing and their effects on performance. *Journal of Second Language Writing*, *24*, 1–19.

Ortega, L. (2012). Epilogue: Exploring L2 writing-SLA interfaces. *Journal of Second Language Writing, 21*, 404–415.

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

Pearson. (2020). *Longman dictionary of contemporary English.* Retrieved from: https://www.ldoceonline.com/. Accessed November 24, 2020.

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.

Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and sentence writing for second language vocabulary acquisition. *Applied Linguistics, 33*(1), 66–82.

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

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics, 24*, 146–161.

Rice, C. A., & Tokowicz, N. (2020). State of the scholarship: A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition, 42*, 439–470.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57.

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163–191). John Benjamins.

Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing, 28*, 1–19.

Schmidt, R. (1994). Deconstructing consciousness in search of useful definitions for applied linguistics. In J. Hulstijn & R. Schmidt (Eds.), *Consciousness in second language learning* (Vol. 11, pp. 11–26). AILA Review. http://www.aila.info/download/publications/review/AILA11.pdf#page=11

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows? *Language Learning, 64*(4), 913–951.

Shaw, P., & Liu, E.T.-K. (1998). What develops in the development of second-language writing? *Applied Linguistics, 19*(2), 225–254.

Silva, B., Kutylowska, K., & Otwinowska, A. (2020). *Data for paper entitled Incidental learning of academic words through writing sentences and compositions: Can an increase in cognitive load affect acquisition?* Figshare. https://figshare.com/s/b111f1e5ede514ddf904

Silva, B., Kutylowska, K., & Otwinowska, A. (2021). Learning academic words through writing sentences and compositions: Any signs of an increase in cognitive load? *Language Teaching Research*, 1–33.https://doi.org/10.1177/13621688211020421

Silva, B., & Otwinowska, A. (2019). VST as a reliable academic placement tool despite cognate inflation effects. *English for Specific Purposes, 54*, 35–49.

Skehan, P. (2003). *A cognitive approach to language learning* (2nd ed.). Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510–532.

Skehan, P. (2014). Limited attentional capacity, second language performance and task-based pedagogy. In S. Peter (Ed.), *Processing perspectives on task performance* (pp. 211–260). John Benjamins.

Storch, N., & Tapper, J. (2009). The impact of an EAP course on postgraduate writing. *Journal of English for Academic Purposes, 8*, 207–223.

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*, 1–10.https://doi.org/10.1177/2158244017730596

Takavoli, P. (2014). Storyline complexity and syntactic complexity in writing and speaking tasks. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 217–236). John Benjamins.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83.

Wesche, M., & Paribakht, T. M. (1996). Assessing vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review, 53*, 13–40.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition, 23*, 41–69.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236–259. https://doi.org/10.1093/applin/amp024

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency. *Studies in Second Language Acquisition, 27*, 567–592.

Zhang, D., & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: Issues, findings and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education, 2*(1), 1–30.

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.

Chapter 7
Study 3—Incidental Learning of Academic Words Through Writing: Can a Decrease in Cognitive Load Affect Acquisition?

7.1 Introduction

Chapter 6 explored the academic vocabulary learning potential of sentence writing (SW) and timed composition (essay) writing (CW). The results showed more learning following SW than CW, possibly because writing essays under time pressure overloaded learners' cognitive resources, reducing the attention allocated to the keywords. The study reported in the current chapter goes one step further and compares the learning of academic vocabulary following three different tasks: SW, 60-min Timed CW, and Untimed CW (the new condition). After reporting the study and briefly discussing its results, Chap. 8 will bring a general discussion on findings of Study 2 (Chap. 6) and 3 (current chapter).

7.2 Method

7.2.1 Aims and Research Questions

This quasi-experiment extends Study 2 to include a third group, namely Untimed CW. Similarly to Study 2, here I measure the acquisition of academic keywords, also provided to students in glossaries. Thanks to adding the third group, it is now possible to compare the acquisition of academic words following the writing of sentences (SW), 60-min (timed) argumentative essays (Timed CW), and untimed argumentative essays (Untimed CW). In this study, I complemented the data from Study 2 (for the SW and Timed CW groups) and collected data for the Untimed CW group. As a result, the data reported here were obtained at the beginning of two different academic years.

As for the Untimed CW, by eliminating the time pressure from argumentative essay writing, hence providing L2 writers with plentiful planning, formulating, and reviewing time, it is possible that attentional resources will be freed. This may result

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_7

in better quality essays (relative to Timed CW) and/or higher levels of lexical learning and retention. Considering the above, the two main research questions are as follows:

- RQ1. Do Polish EFL learners acquire and retain academic words to a similar degree after writing sentences, timed, and untimed argumentative essays?
- RQ2. Does writing untimed argumentative essays with pre-specified keywords reduce L2 writers' cognitive stress when compared to timed essays?

Regarding vocabulary acquisition, I hypothesize that writing untimed essays may generate more lexical learning than writing timed essays; or, writing untimed essays may generate as much learning as writing timed essays because participants will utilize the extra time available to focus on the content and structure of the essays, rather than on keyword use. As for SW, I expect a replication of the findings in Study 2: SW should lead to more lexical learning than Timed CW. Nevertheless, it is unclear whether SW will be less or more conducive to the learning of academic words than Untimed CW.

I explore these hypotheses in three ways. First, I compare the vocabulary learning induced by SW (effectively, the control group) to the learning following Timed CW and Untimed CW tasks. Then, I compare the control essay (without keywords) to the CW essays (each with 10 keywords) regarding lexical complexity, writing accuracy, and fluency in production. As in Study 2, these measures are used to assess quality in written production and to detect signs of increased cognitive load, which may affect lexical learning. Furthermore, and differently from Study 2, here I also assess cognitive load by having participants complete a self-rating scale wherein they report on the perceived difficult and on the level of mental effort of and frustration induced by the tasks (see Sect. 6.2.5; see also Appendix L).

7.2.2 Participants

In total, data were collected from 133 first-year students majoring in English at the Institute of English Studies, University of Warsaw. The data for 55 participants (the same learners as in Study 2) were collected in fall 2018 whereas the data for 78 participants were collected in fall 2019. However, as with Study 2, the number of participants was significantly reduced after they completed a questionnaire at the end of the study to ensure any learning was truly incidental (for more information on the questionnaire see Appendix E and Sect. 6.2.2). Eighteen participants reported having studied at least one of the keywords, mostly by consulting a dictionary after the treatments. These learners were eliminated from further analyses. Eighteen other learners reported suspecting of the true aim of the study and were therefore also excluded from the analyses. Finally, seven participants were excluded for quitting mid-experiment or otherwise failing to perform all tasks. At the end, 90 learners remained: 33 in the SW group (8 males, $M_{age} = 19.52$, $SD = 1.35$), 33 in the Timed CW group (7 males, $M_{age} = 19.16$, $SD = 0.68$), and 24 in the Untimed CW group (7 males, $M_{age} = 19.25$, $SD = 0.94$).

The learners were all selected from nine Writing Practice intact classes—four in 2018 and five in 2019—all following the same syllabus. Four classes were taught by me and five by another teacher from the same institution. As explained in Chap. 4, when participants come from separate groups (i.e., are nested within groups) the data is not truly independent. To solve this problem, Classes (i.e., the 9 intact classes) were entered in the GLMM as a random effect (see Sect. 7.3).

All learners were Polish native speakers ($n = 86$) or speakers of Slavic languages ($n = 4$). Of the latter, three were Ukrainian and one was Russian, and all self-reported their proficiency in Polish as B1 or B2 according to the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). Regarding English proficiency, only students at the B2 level or higher were accepted for enrolment (as assessed by the university's admission criteria). Additionally, I used five different proficiency measures—the same as in Study 2; see below—to confirm that participants in the SW, Timed CW, and Untimed CW groups had similar writing skills and comparable receptive and productive lexical proficiency in English.

7.2.3 *Measures of Participant Proficiency*

Proficiency measures included a receptive lexical knowledge test and four measures derived from a control essay written before the experiment. Independent samples *t*-tests were run for all comparisons and showed no significant differences between the three groups in any test. Table 7.1 shows descriptive statistics for all five measures used.

Table 7.1 Descriptive statistics for proficiency measures

			Textual measures							
Group	Receptive lexical knowledge		Essay score		Productive lexical knowledge				Productive accuracy	
	LexTALE				D		Frequency		Normed errors[a]	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
SW ($n =$ 33)	79.93	12.09	3.88	0.94	99.81	16.76	3.85	0.18	68.042	31.92
Timed CW ($n =$ 33)	73.97	12.77	3.71	0.98	97.03	15.08	3.91	0.21	76.46	34.80
Untimed CW ($n =$ 24)	80.57	8.06	3.58	1.01	95.31	15.41	3.85	0.19	85.60	37.10

[a](Number of errors/number of words in the text) * 1000 (see Biber et al., 2013; Friginal & Weigle, 2014; Jarvis et al., 2003 for a similar normalization of errors)

Receptive Lexical Knowledge. As in Study 2, the online version of Lemhöfer and Broersma's (2012) LexTALE was used (see www.lextale.com). For more details on the test see Sect. 6.2.3 and Appendix F. Participants' LexTALE scores (range = 53.75–98.75) confirmed that they were at level B2 or higher (above 70.7 indicates advanced proficiency). A One-way ANOVA was run on the LexTALE scores to compare the three groups on receptive lexical proficiency. The result showed a barely significant difference between the groups, $F(2, 87) = 3.115$, $p = 0.049$, $\eta_p^2 = 0.07$, representing a small effect size. To verify where the difference lay, I followed suggestions by Field (2017) and Howell (2010) and conducted two post hoc tests: Hochberg's GT2, due to the difference in sample sizes, and Games-Howell, as the data failed the assumption of homogeneity of variance (Levene test: $p = 0.024$). Hochberg's GT2 found no difference between the three groups ($p = 0.86$). Similarly, Games-Howell results showed no difference between SW and Timed CW ($p = 0.13$), SW and Untimed CW ($p = 0.97$), and Timed CW and Untimed CW ($p = 0.53$). As a result, participants in the three treatment groups were deemed comparable in terms of receptive lexical proficiency level.

Textual Measures

Control Essay Score . As in Study 2, participants in the three groups wrote a 300- to 400-word argumentative essay that served as a control essay and as a measure of writing proficiency. Participants were allowed 60 min to complete the task. Its design and scoring were parallel to the essay used in the Timed CW and Untimed CW groups, but without incorporating any pre-specified keywords (see Appendix G for the control and treatment essays. Please note that in Study 3 only the topic from the structured essay was used). The essays were scored holistically (from 1, worst, to 5, best) in the same manner and by the same two trained raters as in Study 2 (see Sect. 6.2.3 for details). The interrater reliability for control essay scores was Pearson's $r = 0.78$, $p < 0.001$. An ANOVA found no difference between SW, Timed CW, and Untimed CW in control essay scores: $F(2, 87) = 0.660$, $p = 0.52$.

Productive Lexical Knowledge. Before deriving the productive lexical knowledge measures from the control essays, participants' compositions were formatted in a similar manner as in Study 2 (see Sect. 6.2.3). The same two measures were derived from the control compositions: D (Malvern & Richards, 2002) as a measure of lexical variation, and word frequency, as a measure of lexical sophistication (see Sect. 6.2.3 for more details on these measures and how they were obtained). ANOVAs found no difference between the three groups for D scores, $F(2, 87) = 0.597$, $p = 0.55$, or frequency scores, $F(2, 87) = 1.094$, $p = 0.34$.

Normed Errors. The fourth textual measure assessed learners' productive accuracy in the control essay (and other essays written by the CW group; see below). Every instance of error increased a participant's error score by one point, and the sum of points was the participant's score. The errors were counted in a similar manner as in Study 2 and scored by the same independent raters (see Sect. 6.2.3). The interrater reliability for number of errors was very high (Pearson's $r = 0.91$, $p < 0.001$). An

ANOVA found no difference between the groups, $F(2, 87) = 1.739$, $p = 0.18$, again evincing no difference in proficiency between SW, Timed CW, and Untimed CW.

7.2.4 Task-Performance Measures of Cognitive Load

I adopted in this study the same task-performance measurements as those in Study 2: Scores, normed Errors, and words per minute (WPM). Here, they were used to compare the control, the timed, and the untimed essays to detect signs of increased cognitive load (i.e., to answer RQ2). Please refer to Sect. 6.2.4 for more details on how these variables were computed.

7.2.5 Another Measure of Cognitive Load: The Self-rating Scale

The self-rating scale was a questionnaire designed to assess participants' perceived task difficulty, level of frustration, and amount of effort induced by the task. The scale was adapted from Kruger et al. (2014) and followed suggestions by Klepsch et al.'s (2017) and Paas et al.'s (2003) work on cognitive load theory. There were two questionnaires, one for the SW task and the other for the CW tasks (see Appendix L for both questionnaires). The only difference between them was minimal, that is, only what was necessary to account for the difference in tasks (SW vs. CW). Figure 7.1 brings an example of the same item in the two different questionnaires (the italics highlight the differences).

Questionnaire for the SW task:

1) How difficult was it to write *the sentences* using the 10 words given? Please assess it on a scale of 1 to 6, where 1 means "very easy" and 6 means "very difficult". ( )

Questionnaire for the CW tasks:

1) How difficult was it to write *the essay* using the 10 words given? Please assess it on a scale of 1 to 6, where 1 means "very easy" and 6 means "very difficult". ( )

Fig. 7.1 Example SW and CW self-rating scale item

There were five items in the questionnaire, each with a 6-point Likert scale. An even number of points was preferred to stop participants from choosing a middle point, which is often the case when they are undecided (Nation, 2013; Schmitt, 2010). The first item, depicted in Fig. 7.1, assessed task difficulty; items 2 and 3 assessed mental effort; items 4 and 5 assessed level of frustration. Items 2 and 4 inquired about mental effort or frustration level, respectively, of the task performed (i.e., SW, Timed CW, or Untimed CW). Items 3 and 5 asked participants to compare the mental effort or frustration level, respectively, of the task performed to the mental effort or frustration level induced by the control essay. This was needed because participants did not fill a self-rating scale after the control essay, just after the three treatment tasks. Importantly, only participants in Study 3 performed this task, not the 39 learners from Study 2. This means that data from only 51 learners is available for analysis.

7.2.6 The Working Memory Task

I decided to measure participants' working memory (WM) because research has shown that WM correlates positively with language learning (e.g., Elgort et al., 2018), including "grammar and vocabulary learning", and language production (Biedroń & Pawlak, 2016, p. 407). To measure participants' WM, I used the Polish version of the Digit Span task, which is part of the Wechsler Adult Intelligence Scale (Brzeziński et al., 2004). The test consists of two parts, a forward digit span task and a backwards Digit Span task, both done orally in a one-to-one basis. Participants sit in a quiet room with the experimenter. They are provided with instructions in Polish (i.e., their L1) and two examples before starting each task. In both tasks, the examiner reads an increasingly long sequence of numbers in Polish (e.g., 3–4–6, 5–6–2–3). The numbers are read slowly and at a constant pace. After each sequence, the participants must repeat these numbers from memory. In the backwards task, the numbers must be held in memory to be repeated backwards (for example, the reply to sequence 3–4–6 must be 6–4–3). The number sequences start with three digits (the forward task) or two digits (the backwards task) and increase in difficulty to up to nine or eight digits, respectively. There are two sequences for each number of items (i.e., difficulty level), totaling 14 sequences.

As for the scoring, test-takers obtain one point for each correct sequence, and the test is stopped when the participants make two mistakes in a row. Two final scores are provided, one for the forward and another for the backwards digit span task. The final score for each participant is the average of both. The test was conducted outside classroom hours, and not all participants were present for the test, resulting in missing data (see Sect. 7.3). Also, because Polish is not the authors' L1, the test was conducted by other Polish colleagues (linguists) familiar with the test.

7.2.7 Instruments Needed to Measure Lexical Learning

Keywords. The 20 keywords were all selected from Coxhead's (2000) Academic Word List (AWL). They were the same 20 keywords as in Study 2—chosen based on Silva and Otwinowska's (2019) research on academic vocabulary (reported in Chap. 5)—as words of average difficulty for Polish learners of English at B2 level of proficiency or higher. Please refer to Sect. 6.2.5 in Chap. 6 for details on how the words were chosen. Here, the 20 keywords were also divided into the same two comparable sets of 10 keywords. However, because the comparability of both sets had already been ensured in Study 2, here participants wrote sentences or essays either with Set A or Set B, not both (as in Study 2). Put differently, each participant incorporated 10 words in their tasks, not 20. Sets were assigned randomly among participants. When combining the data from Study 2 and 3, only the data from the structured condition in Study 2 (the second condition) were taken. Therefore, only the data for one set (A or B) were taken per student (see Sect. 6.2.6). At the end, in the current study, Sets A and B were divided among the 90 participants as follows: 49 participants were given Set A (54.44%) and 41 participants performed their tasks with Set B (45.56%).

Glossary. To assist with task performance, all participants in SW, Timed CW, and Untimed CW were given a glossary right before performing their writing tasks (see Appendix H). These are the same glossaries as in Study 2, providing a definition and two examples for each keyword (see Sect. 6.2.5 for more details).

Test of Lexical Learning (Pretest and Posttest). The tests of lexical learning were the same and were scored in a similar manner as the tests in Study 2 (see Sect. 6.2.5). These were the VKS and the free association test. In the VKS, for each keyword, participants received a score from 1 to 6 (from no knowledge to full productive knowledge). In the association test, participants provided up to four words they associated with each keyword, resulting in a score of 0–4.

7.2.8 Design

Figure 7.2 shows the research design in some detail. First, all participants wrote the control essay. One week later, the groups sat the pretests and the receptive lexical test (LexTALE). Participants were informed that the purpose of the pretest and the LexTALE was to measure their proficiency in English. The LexTALE also acted as a cognitively demanding task immediately following the pretests, so the keywords could be flushed from participants' memories, as recommended by Schmitt (2010). Then, one week later, the SW, Timed CW, and Untimed CW treatments performed their respective tasks, followed by the self-rating scale of cognitive load. Two weeks after writing their tasks, participants completed the posttests and the incidental learning questionnaire. The Digit Span task was performed outside the classroom

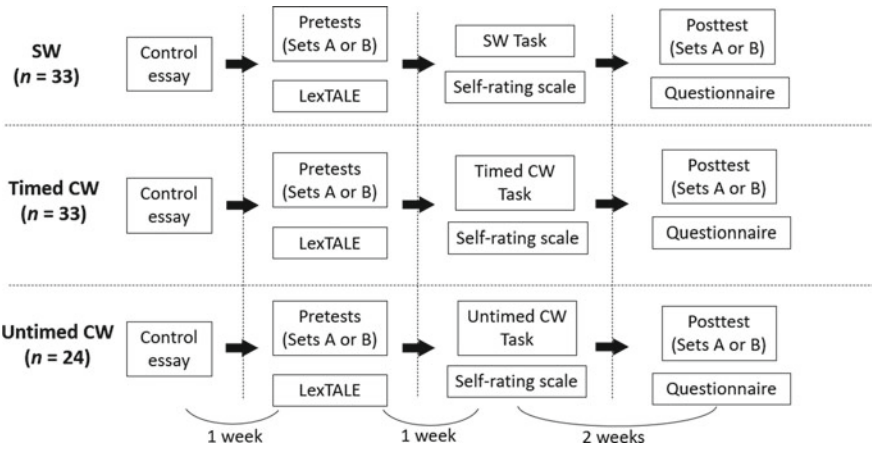

Fig. 7.2 Illustration of research design and procedure

at a place and time agreed with each participant individually, during or after the quasi-experiment.

7.2.9 Treatments

The Sentence-Writing (SW) Treatment Group. This is the same treatment group as in Study 2. Participants were given glossaries and instructed to write a total of 10 grammatically correct sentences with at least 10 keywords each, one sentence per keyword (see Sect. 6.2.7 for more details).

The Timed Composition-Writing (Timed CW) Treatment. This is the same treatment group as in Study 2. One argumentative essay similar to the control essay in type, length (300–400 words), time limit (60 min), and scoring (two independent raters) was written in one session. Participants had to include 10 keywords in the essay, either from Set A or B (see Appendix G for the structured CW task, the one used here; see Appendix J for two sample essays produced by participants). The instructions foregrounded the importance of composing a well-structured text to ensure participants treated essay writing, not keyword use, as the primary task. For more details on this task see Sect. 6.2.7. The topic was the following: "Do you agree or disagree with the following statement? *Parents are the best teachers.* Use specific reasons and examples to support your opinion."

The Untimed Composition-Writing (Untimed CW) Treatment Group. This task was the same task as the one for the Timed CW group—including the topic, length, and need to use 10 keywords (either Set A or B). The only difference between the Timed CW and the Untimed CW task was that here participants had no time limit.

The essays were written during 90 min-lessons, but participants could extend their writing into their 30-min break following the lesson. If any participant wanted to use even more time, they were taken to a separate, nearby quiet room to finish the task. Nevertheless, this was rarely needed.

7.2.10 Procedures

No piloting was needed for the glossaries, test of lexical learning, CW and SW tasks since they had been piloted and used in Study 2. The Digit Span task and the self-rating scale were not piloted because they are established measures in the field. All tasks were conducted during regular 90-min classes (except for the Digit Span task) at the beginning of participants' first academic year. Data from 39 participants were collected in fall 2018; however, only the data from the second condition (i.e., structured essay or the second time learners wrote sentences in SW; see Fig. 6.4) were included in the analyses. Data from the remaining 51 learners were collected one year later. The pretests, posttests, glossaries, the incidental learning questionnaire, and the self-rating scale were provided to participants in paper form and collected after the treatment. The Digit Span task was performed at a time that was suitable for each individual participant and the experimenter. The two essays, control and treatment essays (Timed CW and Untimed CW), the sentence writing, and the receptive lexical test (LexTALE) were conducted on laptops. Participants had no access to the internet, except for the LexTALE. Also, the proofreading tool in the word processor was turned off. All tasks were administered and monitored closely by the author.

To hide the true purpose of the quasi-experiment, CW participants were told that the aim of the study was to measure their ability to write argumentative essays under different conditions. SW learners were told that the purpose of the quasi-experiment was to measure their sentence-writing speed while being obliged to follow certain conditions. For them, the condition was to incorporate keywords in sentences, but they did not know whether other groups were writing sentences under different conditions.

7.3 Analysis

7.3.1 Choosing the Data

Study 3 reused data from Study 2 (SW and Timed CW). Of importance, however, because in Study 3 learners in Timed CW and Untimed CW wrote only one essay, and SW leaners wrote only one set of 10 sentences, data from only one of the timed essays and from one set of sentences in SW in Study 2 were utilized for analyses in Study 3. This was done to match the number of data points in Study 3. In Study

2, CW participants wrote two essays, first in the unstructured condition and then in the structured condition, and two sets of sentences (see Fig. 6.4). In Study 3 I only used data from the second condition and second set of sentences. I opted for the second condition because, as explained in Sect. 6.2.7, participants in Study 2 treated both essays as structured, and therefore choosing the second condition seemed more suitable.

As in Study 2, in Study 3 the VKS and association tests were used to measure and compare lexical learning following SW, Timed CW, and Untimed CW (i.e., to answer RQ1). The tests yielded the same three outcome variables as in Study 2: VKS_6, VKS_3, and Association (see Sect. 6.3 for more details). Additionally, scores in the Digit Span task yielded the continuous variable WM_Scores. Importantly, because the Digit Span task was conducted outside regular class hours, not all participants completed the task, resulting in missing data. In total, 58 participants (64.44%) completed this task. The procedure used to generate this missing data is described below.

7.3.2 Generating the Digit Span Missing Data

It is common to replace missing data with the mean or median of a given group (Field, 2017). Nonetheless, this is acceptable only when there are very few data points missing. This is because, as pointed out by Cheema (2014) and Honaker et al. (2011), by using means or medians, the researcher reduces de variance in the data and hence the standard error estimates. Lower standard error estimates increase the possibility of obtaining a significant p value (i.e., $p < 0.05$), thus making committing a Type I error more likely (i.e., finding a significant difference when one does not exist).

To overcome this problem, I generated the missing data via a multiple imputation procedure. In this procedure, the software utilizes regression analyses on all available predictors—here, Group (SW, Timed CW, and Untimed CW), Time (pretest and posttest), VKS_6, VKS_3, Association, and LexTALE scores—to fill the missing data with logical values. The minimum and maximum values, based on the existing range in WM_Scores (Digit Span), were set to 4 and 10.5. The multiple imputation performed five different analyses (i.e., imputations), each with 50 iterations (simulations). The final imputed values were the pooled values from the five analyses. Appendix M brings the descriptive statistics of the original data, of each of the five imputations, and of the final data. Note that the means and standard deviations of the original data ($M = 6.7328$, $SD = 1.3753$) and the final data ($M = 6.7339$, $SD = 1.3606$) differ only slightly.

7.3.3 Statistical Analyses

To answer RQ1, which asked whether writing sentences, timed and untimed essays generate similar lexical learning, generalized linear mixed models (GLMMs) were run to analyze the three outcome variables, that is, VKS_6, VKS_3, and Association. Three categorical variables were entered as random effects: Participants (the 90 participants), Items (the 20 keywords), and Class (the 9 intact classes). The fixed effects were Group (SW, Timed CW, Untimed CW), Time (pretest and posttest), the Group * Time interaction, and WM_Scores, entered as a covariate to control for the effect of participants' WM on lexical learning. Appendix N illustrates the full process of creation of the GLM models. The models reported here are the final models for each dependent variable.

To answer RQ2 (i.e., "Does writing untimed argumentative essays with pre-specified keywords reduce L2 writers' cognitive stress when compared to timed essays?"), I derived eight variables to detect signs of increased cognitive load in Timed and Untimed CW. From the control, timed, and untimed essays, I obtained the within-subject variables Scores, Errors, and words per minute (WPM), as in Study 2. All three variables were normally distributed (z scores for skewness <1.96) and variance was homogenous, thus parametric tests were used. When comparing the control essays to treatment essays (timed or untimed), paired-sample t-tests were used as the essays were written by the same participants. When comparing Timed CW to Untimed CW, independent samples t-tests were used. From the self-rating scale, I obtained five variables (see Appendix L for the questionnaire items that originated these variables). The variables Difficulty, Effort, and Frustration, based on items 1, 2, and 4, respectively, were used to compare the perceived difficulty, level of effort, and frustration between the three treatment tasks (i.e., SW, Timed CW, and Untimed CW). ANOVAs were used for these analyses. The variable Effort_to_Control, based on item 3, compared the level of effort between the treatment tasks and the control essay. Finally, the variable Frustration_to_Control, based on item 5, compared the level of frustration between the treatment tasks and the control essay. The variables Effort_to_Control and Frustration_to_Control were analyzed with one-sample t-tests.

7.4 Results

Before answering the research questions, it is important to note that all three variables obtained from the VKS and association tests—i.e., VKS_6 (all six levels), VKS_3 (levels 1–2, 3–4, 5–6 combined), and Association—registered lexical gains between the pretest and posttest ($p < 0.001$). This is can be seen by referring to Time (i.e., pretest–posttest lexical gains) in Tables 7.3, 7.5, and 7.7 in the subsections below. Also importantly, the covariate WM_Scores was only significant with the variable Association ($p = 0.021$; see Table 7.7). Even then, the effect is rather weak. The

odds ratio (0.957) shows that participants with higher WM_Scores were 4.3% less likely to have higher scores in Association in the posttest than participants with lower WM_Scores. This result was unexpected and may have been an artefact of the data. Also, this difference is minor and shows that working memory only barely helped predict Association scores.

7.4.1 Results for Tests Measuring Lexical Knowledge

RQ1 asked whether Polish EFL learners acquire academic vocabulary to a similar degree after performing SW, Timed CW, and Untimed CW. To obtain answers, I constructed three generalized linear mixed models (GLMMs), one for each dependent variable: VKS_6, VKS_3, and Association (see Sect. 7.3). The analyses and the logic behind them are parallel to those carried out in Study 2. I discuss the GLMMs separately below.

Results for VKS_6. The proportion of scores for VKS_6 is illustrated in Fig. 7.3 and shown in Table 7.2. Table 7.3 shows the results of the GLMM. Generally, most of the learning occurred from scores 1 or 2 (no knowledge or ability to recognize) in the pretest to score 6 (full productive knowledge) in the posttest. This means that most of the knowledge gained was in breadth, not in depth. Still, Untimed CW learners showed only a slightly higher increase between pretest and posttests in level 6 scores (full productive knowledge) than SW and Timed CW participants: 23% for Untimed CW relative to 19% for SW and 20% for Timed CW.

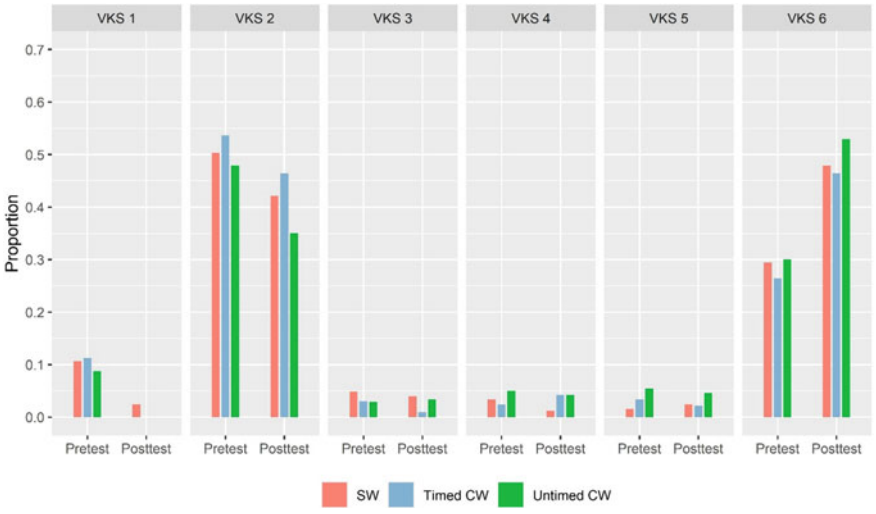

Fig. 7.3 Proportion of VKS_6 scores in the pretest and posttest for SW, timed CW, and untimed CW

Table 7.2 Proportion of VKS_6 scores for SW, timed CW, and untimed CW

	VKS Score											
	1		2		3		4		5		6	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Untimed CW ($n = 24$)	0.09	0.00	0.48	0.35	0.03	0.03	0.05	0.04	0.05	0.05	0.30	0.53
Timed CW ($n = 33$)	0.11	0.00	0.54	0.46	0.03	0.01	0.02	0.04	0.03	0.02	0.26	0.46
SW ($n = 33$)	0.11	0.02	0.50	0.42	0.05	0.04	0.03	0.01	0.02	0.02	0.29	0.48

Table 7.3 Fixed and random effect estimates for VKS_6

Fixed effects					
	Estimate (std. error)	t	*P*	Odds ratio	95% CI odds ratio
Intercept	–	–	–	–	–
Group untimed (GU)	0.132 (0.21)	0.630	0.529	1.11	[0.76, 1.72]
Group timed (GT)	−0.094 (0.20)	−0.462	0.644	0.93	[0.61, 1.36]
Time	0.832 (0.10)	8.511	<0.001	1.98	[1.89, 2.80]
WM_Scores	0.004 (0.03)	0.143	0.887	1.00	[0.94, 1.06]
GU * Time	0.12 (0.18)	0.697	0.487	1.10	[0.63, 1.25]
GT * Time	0.054 (0.15)	0.363	0.717	1.05	[0.78, 1.42]
GT[a] * Time	−0.065 (0.18)	−0.361	0.719	0.95	[0.66, 1.34]
Random effects					
	Variance (std. error)	95% CI	Wald Z	*p*	ICC
Residual	3.29	–	–	–	–
Participants (intercept)[b]	0.241 (0.059)	[0.149, 0.388]	4.090	<0.001	0.0473
Time | Participants (intercept-slope correlation)	0.683 (0.314)	[-0.308, 0.963]	2.177	0.029	0.1344
Items (intercept)[c]	0.868 (0.097)	[0.697, 1.082]	8.934	<0.001	0.1708

Note Number of data points = 1800; items = 20; participants = 90. Probability distribution: multinomial; link function: cumulative complementary log–log. Reference categories (predictors) = SW, pretest. Reference category target: VKS score 6

[a]Reference category: untimed CW

[b]Covariance structure = first-order autoregressive (AR1)

[c]Covariance structure: variance component

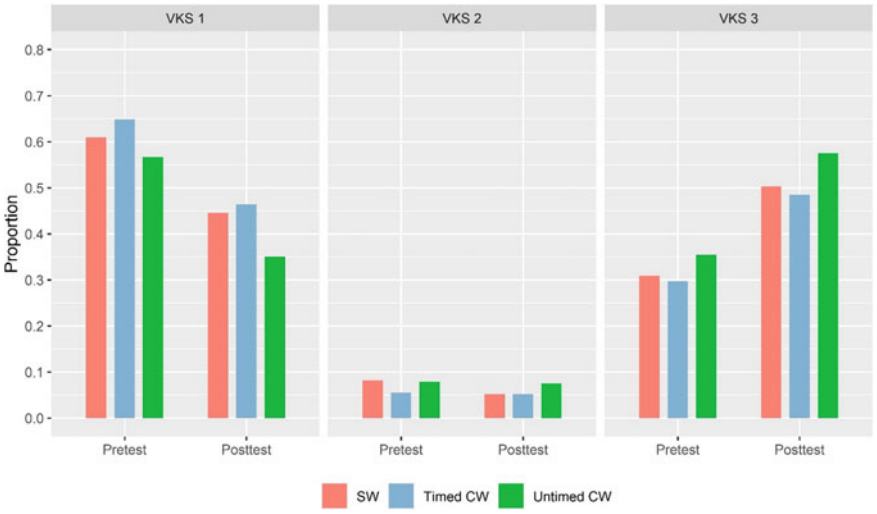

Fig. 7.4 Proportion of VKS_3 scores in the pretest and posttest for SW, timed CW, and untimed CW

Table 7.4 Proportion VKS_3 scores for SW, timed CW, and untimed CW in the pretest and posttest

	VKS Score					
	1		2		3	
	Pre	Post	Pre	Post	Pre	Post
Untimed CW ($n = 24$)	0.57	0.35	0.08	0.08	0.35	0.58
Timed CW ($n = 33$)	0.65	0.46	0.05	0.05	0.30	0.48
SW ($n = 33$)	0.61	0.45	0.08	0.05	0.31	0.50

However, this increase was not significant, as shown by the Group * Time interactions in Table 7.3. The intraclass correlation (ICC) shows that the random effects combined explain 21.81% of the variance left unexplained by the fixed effects. This means that VKS_6 scores varied considerably among participants and lexical items.

Results for VKS_3. Another way to answer RQ1 is by looking at how students performed on different types of word knowledge, represented by the variable VKS_3. It represents three levels of vocabulary knowledge: level 1—no knowledge or ability to recognize word-form; level 2—receptive knowledge of meaning to different degrees of certainty; level 3—productive knowledge. An increase from level 1 to levels 2 or 3 in the posttest represents a gain in breadth of lexical knowledge. An increase from level 2 to level 3 (i.e., receptive to productive knowledge) may indicate gains in depth. The proportion of scores for VKS_3 is depicted in Table 7.4 and Fig. 7.4. Similarly to VKS_6, most of the knowledge gained was in breadth, not in depth, since scores mostly changed from levels 1 to 3. Again, the proportions for

Table 7.5 Fixed and random effect estimates for VKS_3

Fixed effects					
	Estimate (std. error)	t	*p*	Odds ratio	95% CI odds ratio
Intercept	–	–	–	–	–
Group untimed (GU)	0.125 (0.11)	1.102	0.609	1.13	[0.78, 1.69]
Group timed (GT)	−0.141 (0.13)	−1.071	0.497	0.87	[0.08, 2.09]
Time	0.659 (0.04)	17.105	<0.001	1.93	[1.79, 2.09]
WM_Scores	0.023 (0.02)	1.267	0.221	1.02	[0.98, 1.06]
GU * Time	0.117 (0.06)	2.066	0.039	1.12	[1.01, 1.26]
GT * Time	−0.007 (0.14)	0.051	0.960	1.01	[0.77, 1.32]
GT[a] * Time	−0.110 (0.14)	−0.763	0.445	0.90	[0.68, 1.19]
Random effects					
	Variance (std. error)	95% CI	Wald Z	*p*	ICC
Residual	3.29	–	–	–	–
Participants (intercept)[b]	0.415 (0.096)	[0.264, 0.652]	4.339	<0.001	0.0961
Items (intercept)[b]	0.600 (0.080)	[0.462, 0.779]	7.481	<0.001	0.1389
Class (intercept)[b]	0.014 (0.043)	[0.00, 5.487]	0.328	0.743	0.0032

Note. Number of data points = 1800; items = 20; participants = 90. Probability distribution: multinomial; link function: cumulative complementary log–log. Reference categories (predictors) = SW, pretest. Reference category target: VKS score 3
[a]Reference category: untimed CW
[b]Covariance structure = variance component

Table 7.6 Descriptive statistics for association for SW, timed CW, and untimed CW in the pretest and posttest

	Association scores (max. = 4)					
	Pretest			Posttest		
	Mean	SD	Mdn	Mean	SD	Mdn
Untimed CW ($n = 24$)	1.00	1.40	0	1.66	1.55	2
Timed CW ($n = 33$)	0.56	1.05	0	0.93	1.21	0
SW ($n = 33$)	0.68	1.15	0	1.07	1.28	0

Table 7.7 Fixed and random effect estimates for association

Fixed effects					
	Estimate (std. error)	t	*p*	Odds ratio	95% CI odds ratio
Intercept	−0.394 (0.18)	−2.163	<0.05	0.67	[0.39, 1.17]
Group untimed (GU)	0.441 (0.23)	1.921	0.190	1.55	[0.60, 4.03]
Group timed (GT)	−213 (0.26)	−0.807	0.447	0.81	[0.08, 2.09]
Time	0.470 (0.03)	16.775	<0.001	1.74	[1.36, 2.24]
WM_Scores	−0.044 (0.02)	−2.451	0.021	0.957	[0.92, 0.99]
GU * Time	0.086 (0.13)	0.678	0.505	1.09	[0.84, 1.42]
GT * Time	0.014 (0.06)	0.225	0.845	1.01	[0.75, 1.37]
GT[a] * Time	−0.072 (0.13)	−0.545	0.591	0.93	[0.71, 1.22]
Random effects					
	Variance (std. error)	95% CI	Wald Z	*p*	ICC
Residual	3.29	–	–	–	–
Participants (intercept)[b]	0.252 (0.064)	[0.153, 0.416]	3.922	<0.001	0.0517
Time | Participants (correlation)	−0.182 (0.254)	[−0.605, 0.320]	−0.717	0.474	0.0373
Items (intercept)[b]	0.513 (0.085)	[0.371, 0.710]	6.037	<0.001	0.1053
Time | Items (correlation)	−0.533 (0.108)	[−0.711, −0.290]	−4.939	<0.001	0.1094
Class (intercept)[b]	0.103 (0.083)	[0.022, 0.497]	1.250	0.211	0.0212

Note Number of data points = 1800; items = 20; participants = 90. Probability distribution: poisson; link function: log. Reference categories (predictors) = SW, pretest. Reference category target: ascending

[a]Reference category: untimed CW

[b]Covariance structure = AR1

[c]Covariance structure: variance components

VKS_3 show more learning for Untimed CW than for Timed CW and SW. First, level 1 scores decreased by 22 points in Untimed CW, 19 points in Timed CW, and 16 points in SW. Second, level 3 increased by 23% in Untimed CW, 18% in Timed CW, and 19% in SW.

The Group * Time interaction for VKS_3, presented in Table 7.5, confirms significantly higher lexical gains for Untimed CW than for SW ($p = 0.039$), but there is no difference between SW and Timed CW ($p = 0.960$), or Timed CW and Untimed CW ($p = 0.445$). Still, the difference between Untimed CW and SW appears minor:

The odds ratio shows that Untimed CW learners were 12% more likely to score 3 (compared to scores 1 and 2 combined) in the posttest than SW participants, which is clearly not a substantial advantage, mainly considering the amount of time spent writing untimed argumentative essays. The ICC shows that the random effects together explained 23.82% of the variance left over from the fixed effects.

Results for Association. The descriptive statistics for Association are shown in Table 7.6 and the GLMM in Table 7.7. Untimed CW participants had higher average scores in the pretest than the other participants. In terms of pretest–posttest gains, the Untimed CW group increased mean scores by 0.66, whereas Timed CW increased by 0.37 and SW registered mean gains of 0.39. The median increase was also higher for Untimed CW: 2 points, versus 0 in Timed CW and SW.

The odds ratio for Time in Table 7.7 shows that participants in the three treatment groups had a 74% higher chance of achieving a higher score in Association in the posttest than in the pretest. This shows that all three groups yielded significant gains in depth of knowledge. However, the increase in scores was not significantly different between any of the groups, as shown by the Group * Time interactions. As for the ICC, all random effects put together explain 17.82% of the variance left over from the fixed effects.

In answering RQ1, the GLMMs show very little evidence of any difference in the amount of lexical learning between SW, Timed CW, and Untimed CW. Only the model for VKS_3 found a statistically significant advantage for Untimed CW over SW, but this difference was unsubstantial. Interestingly, the descriptive statistics of all three models (VKS_6, VKS_3, Association) showed more learning for Untimed CW than for the other two groups, although this difference did not reach statistical significance. These findings will be discussed in more detail in Chap. 8.

7.4.2 Results for Measures of Cognitive Load

RQ2 asked whether writing untimed argumentative essays with pre-specified keywords reduces the level of cognitive load when compared to timed essay writing. We answered this question in two ways. First, I analyzed the within-subject textual variables Scores, Errors, and WPM. Then, I compared the results for the five variables derived from the self-rating scale. These are reported separately below.

Results for Task-Performance Variables. The descriptive statistics are shown in Table 7.8. Regarding overall essay scores, t-tests found no difference between the control and timed essays, $t(32) = 1.856$, $p = 0.073$, control and untimed essays, $t(23) = -1.541$, $p = 0.137$, but found a statistically significant difference between timed and untimed essays, $t(54.841) = -2.141$, $p = 0.027$, 95% CI [−1.04, −0.06], $r = 0.28$, representing a small effect size (Plonsky & Oswald, 2014). This shows that using pre-specified keywords when writing did not reduce the overall quality of essays, but the untimed essays had slightly better quality than the timed essays, as illustrated in Fig. 7.5.

Table 7.8 Descriptive statistics for textual measures

	Essay scores		Normed errors		WPM	
Essay conditions	M	SD	M	SD	M	SD
Control (timed essay) ($n = 33$)	3.71	0.98	76.46	34.80	7.47	1.49
Timed CW	3.36	1.09	94.96	41.50	6.23	1.31
Control (untimed essay) ($n = 24$)	3.47	0.98	85.60	37.10	8.08	1.58
Untimed CW	3.92	0.75	75.16	22.26	5.35	1.08

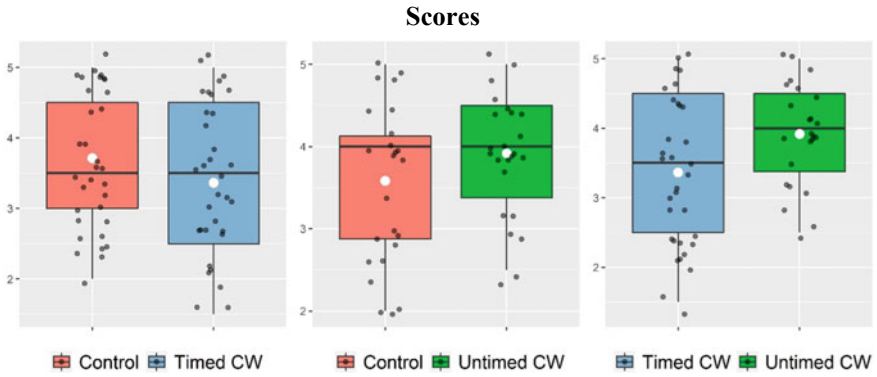

Fig. 7.5 Boxplots comparing essay scores. The white dots represent the means. The control essays were written by different students in the timed CW and untimed CW groups

Concerning Errors, the t-tests found a significant difference between the control and timed essays, $t(32) = -3.524$, $p < 0.001$, 95% CI [−29.20, −7.81], $r = 0.53$, representing a medium effect size, and the timed and untimed essays, $t(51.184) = 2.321$, $p = 0.024$, 95% CI [2.67, 36.94], $r = 0.31$, with a small effect size. There was no difference in the number of errors between the control and untimed essays: $t(23) = 1.547$, $p = 0.136$. It appears that incorporating keywords in essays increased the number of errors for timed essays only. In fact, untimed essays had fewer errors than the control (although the difference was not significant) and timed essays, as illustrated in Fig. 7.6.

Finally, differences were also found in WPM between the control and timed essays, $t(32) = 3.997$, $p < 0.001$, 95% CI [0.61, 1.88], $r = 0.58$, the control and untimed essays, $t(21) = 8.684$, $p < 0.001$, 95% CI [2.08, 3.39], $r = 0.88$, both with large effect sizes, and the timed and untimed essays, $t(53) = 2.618$, $p = 0.012$, 95% CI [0.88, 0.34], $r = 0.34$, with a small effect size, as illustrated in Fig. 7.7.

To sum up, using keywords in essays was not detrimental to quality. Effectively, when no time limit was given, essay quality improved, even though learners performed a secondary task (i.e., keyword use). Multitasking reduced accuracy in the timed essay, but not in the untimed essay, although participants took significantly longer writing the timed essay than the control. Thus, similarly to Study 2, these

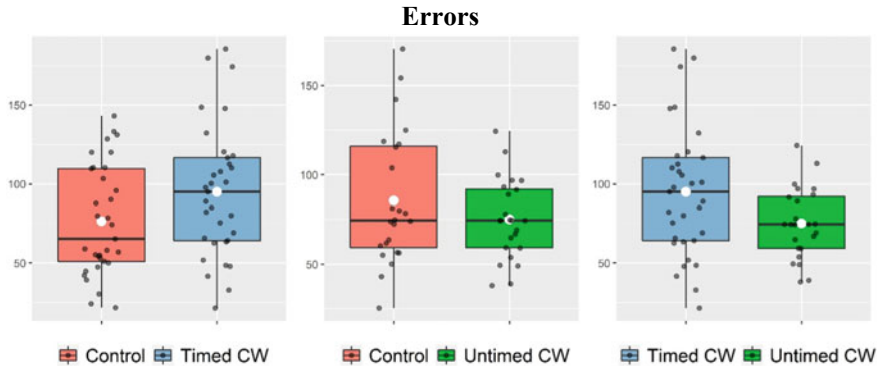

Fig. 7.6 Boxplots comparing essay errors. The red dots represent the means. The control essays were written by different students in the timed CW and untimed CW groups

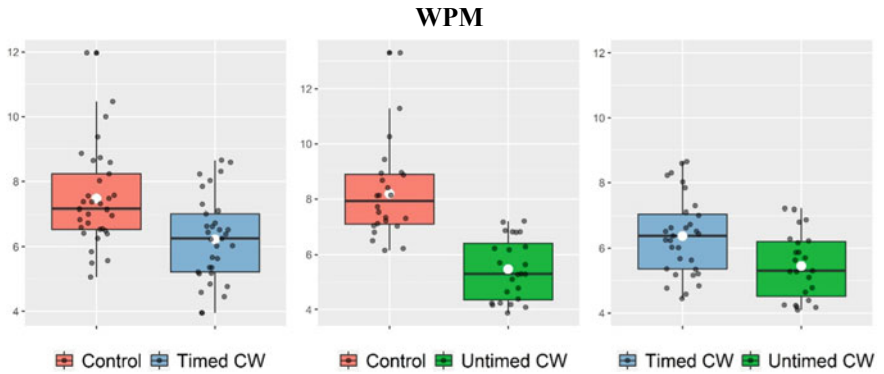

Fig. 7.7 Boxplots comparing essay WPM. The red dots represent the means. The control essays were written by different students in the timed CW and untimed CW groups

findings suggest that when writing timed well-structured essays with keywords, L2 writers are less fluent and less accurate, which is a sign of increased cognitive load. However, cognitive load does not appear to increase when there is no time pressure.

Results for Self-rating Scale. One-way ANOVAs were used to compare Difficulty, Effort, and Frustration between SW, Timed CW, and Untimed CW. The variables Effort_to_Control and Frustration_to_Control, comparing the level of effort and frustration of each treatment to the control essay, were analyzed with one-sample t-tests. The reference level was set to 3.5 (between options 3 and 4 in the scale), which represent learners' choice had they been able to select a middle "I don't know" option. This means that the one-sample t-tests will compare the difference between the data and this middle point. The descriptive statistics can be found in Table 7.9.

Table 7.9 Descriptive statistics for self-rating scale measures

Groups	Difficulty		Effort		Frustration		Effort_to_Control		Frustration_to_Control	
	M	SD	M	SD	M	SD	M	SD	M	SD
Untimed CW	3.62	0.23	4.58	0.25	2.88	1.62	4.64	1.43	2.96	1.49
Timed CW	3.62	0.28	4.56	0.31	3.19	1.47	5.06	1.63	3.56	1.75
SW	3.00	0.33	3.64	0.37	3.36	1.43	4.79	1.47	4.00	1.55

The descriptive statistics show that Timed CW and Untimed CW believed their tasks were more difficult and demanded more effort than participants in SW. Nevertheless, SW learners seem to have been more frustrated by their tasks than learners' in the other two groups. Still, the ANOVA found no difference between the groups for Difficulty, $F(2, 48) = 1.362$, $p = 0.266$, Effort, $F(2, 48) = 2.524$, $p = 0.091$, and Frustration, $F(2, 48) = 0.439$, $p = 0.647$. This shows that participants in the three groups perceived their tasks to be of similar difficulty and to induce a similar level of effort and frustration.

Regarding Effort_to_Control, one-sample t-tests found significant differences for all groups: Untimed CW, $t(23) = 4.294$, $p < 0.001$, 95% CI [0.67, 1.91], $r = 0.67$, Timed CW, $t(15) = 5.882$, $p < 0.001$, 95% CI [1.00, 2.13], $r = 0.84$, and SW, $t(10) = 2.269$, $p = 0.025$, 95% CI [0.17, 2.10], $r = 0.64$, all representing large effect sizes. Still, the strongest effect was found for Timed CW, which accordingly, had the highest mean of the three groups. Interestingly, SW learners also found that writing 10 sentences with keywords demanded more effort than writing a full essay without keywords (i.e., the control essay) thus showing that keyword use increased cognitive load even when writing sentences.

Results for Frustration_to_Control showed no differences for any of the groups: Untimed CW, $t(23) = -1.783$, $p = 0.088$, Timed CW, $t(15) = 0.143$, $p = 0.888$, and SW, $t(10) = 1.070$, $p = 0.310$. Of note, untimed essays were rated as less frustrating than control essays, showing that the higher cognitive load induced by keyword use was counteracted by the lack of time pressure. Also of importance, SW learners were the most frustrated and Timed CW participants were almost exactly as frustrated as when writing the control, despite the need to incorporate keywords. Both findings were unexpected. In fact, the variables Frustration and Frustration_to_Control yielded counterintuitive results as participants in SW seemed to be the most frustrated of all learners. This finding will be discussed in more detail in the next section.

To answer RQ2, it appears that freeing L2 writers of time restrictions when writing argumentative essays reduces cognitive load. First, the textual measures showed untimed essays had better quality and were more accurate than timed essays, while being as accurate as the control. Second, untimed essays demanded less effort than timed essays, as measured by Effort_to_Control, and were less frustrating than the control and the other two tasks.

7.5 Discussion

Similarly to Study 2 (Chap. 6), the current study set out to increase our understanding on how a more explicit focus on written production may facilitate the incidental acquisition of academic words. In this study, I also drew on Laufer and Hulstijn's (2001) involvement load hypothesis to compare the lexical learning potential yielded by writing tasks. Here, however, I added a third group—i.e., the untimed argumentative essay writing group (Untimed CW)—to the sentence writing (SW) group and the 60-min timed argumentative essay writing group (Timed CW) already explored in

Study 2. Furthermore, in the current study, I controlled for participants' working memory, known to predict lexical learning (Elgort et al., 2018), and included a self-rating scale questionnaire as an additional measure of cognitive load. The results were as follows.

When answering RQ1, the statistical analyses found scant evidence that untimed essays generate more lexical learning than SW and Timed CW. Moreover, the results indicate that SW yields as much learning as Timed CW, which serves as counterevidence to Study 2, wherein SW yielded more learning than Timed CW. These are interesting and somewhat unexpected results that will be discussed in detail in Chap. 8, the general discussion. Still, it is worth noting that the results of the current study corroborate the predictions of the ILH (that is, that writing sentences and longer texts generates similar levels of learning) and of previous empirical findings (i.e., Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017). Nevertheless, neither Study 2 nor 3 found support for Zou's (2017) study, which found that CW was more conducive to learning than SW.

To answer RQ2, which asked whether writing untimed essays with keywords was less cognitively demanding than writing timed essays with keywords, I adopted the same task-performance measures as in Study 2 (i.e., Scores, Errors, and WPM) and the self-rating scale questionnaire. The results showed that the overall quality of untimed essays was higher than the quality of timed essays while being similar to the control. Also, untimed essays were as accurate as the control essays, despite keyword use. Still untimed essays were more accurate than timed essays, which expectedly, were less accurate than the control ones, as in Study 2. Regarding fluency, learners in the Untimed CW group were the slowest writers, but this is an inherent characteristic of the task. What is noteworthy is that timed essays were written more slowly than control essays, again replicating findings from Study 2. All this appears to indicate that the lack of time pressure freed Untimed CW learners' attentional resources. This is because untimed essays were written as well and as accurately as control essays (i.e., without keywords), while having better quality and being more accurate than timed essays (where keyword-use was also required).

Results from the self-rating scale appear to support these findings. The results demonstrate that although untimed essays demanded more effort than the control, they also required less effort than timed essays. Interestingly, Untimed CW was the only group that rated writing untimed essays with keywords as *less* frustrating than writing (timed) control essays without pre-specified keywords. Timed CW participants rated their task as equally frustrating, while interestingly, SW learners found writing sentences with keywords more frustrating than writing control essays. In fact, the SW group also had the highest score for levels of frustration of the three groups.

These results appear counterintuitive in that writing argumentative essays with keywords, especially timed essays, should be more frustrating than writing sentences with keywords. Before discussing reasons for these results, it is important to revisit the relevant items in the self-rating scale. Items 4 and 5 of the scale, devised following similar wording to that of Kruger et al.'s (2014), read as follows (items taken from scale for the SW task in Appendix L; italics not present in the original task):

Item 4: "How *stressed, annoyed or frustrated* were you while writing these sentences? Please assess it on a scale of 1 to 6, where 1 means "very low" and 6 means "very high"."

Item 5: "Writing these sentences with 10 words made me more *stressed, annoyed or frustrated* than writing the first essay (without the 10 words). Please assess it on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree"."

One plausible reason why SW participants reported being more frustrated than learners in the other groups might be that they simply did not like the SW task, and hence may have felt "frustrated" and/or "annoyed" with having to perform this task. This being the case, items 4 and 5 may have measured learners' appreciation for the task, not levels of cognitive load. Timed CW and Untimed CW participants' level of frustration, on the other hand, may either also represent their appreciation for the task, or the level of cognitive load (e.g., "stress") induced by the tasks (since Timed CW learners reported higher frustration levels than Untimed CW participants). Put differently, it appears that items 4 and 5 of the self-rating scale lack validity, and therefore, constitute a limitation of the current study.

7.6 Conclusion

This chapter has reported on the findings of Study 3, which complemented Study 2, reported in Chap. 6. Study 3 compared the learning of academic vocabulary following sentence writing (SW), timed composition (essay) writing (Timed CW), and Untimed CW. The results showed only minor evidence that writing untimed essays with keywords generated more learning than writing sentences or timed essays. There was no difference in learning following SW and Timed CW, which contradicts the results from the Study reported in Chap. 6. This is an interesting result not least because Untimed CW appeared to be less cognitively demanding than Timed CW (i.e., untimed essays were more accurate and had higher overall quality than timed essays), and hence should, as hypothesized, have generated significantly more learning than the Timed CW task, and possibly than SW. Chapter 8 will discuss in more depth the combined results of Studies 2 and 3.

Funding The research was supported by grant 2019/35/N/HS2/03550 from the National Science Centre Poland awarded to the author, Breno Barreto Silva.

References

Biber, D., Gray, B., & Poonpon, K. (2013). Pay attention to the phrasal structures: Going beyond t-units—A response to WeiWei Yang. *TESOL Quarterly*, The Forum, 192–201.

Biedroń, A., & Pawlak, M. (2016). The interface between research on individual difference variables and teaching practice: The case of cognitive factors and personality. *Studies in Second Language Learning and Teaching, 6*(3), 395–422. https://doi.org/10.14746/ssllt.2016.6.3.3

Brzeziński, J., Gaul, M., Hornowska, E., Jaworowska, A., Machowski, A., & Zakrzewska, M. (2004). *Skala Inteligencji D. Wechsleradladorosłych. Wersjazrewidowana–Renormalizacja,WAIS-R(PL).* Pracownia Testów Psychologicznych PTP.

Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods, 13*(2), 53–75. https://doi.org/10.22237/jmasm/1414814520

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

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

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Friginal, E., & Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing, 26*, 80–95.

Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique feature analysis or involvement load hypothesis: Estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research, 47*, 859–869.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software, 45*(7), 1–47.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Cengage Learning.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*, 377–403.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1–18. https://doi.org/10.3389/fpsyg.2017.01997

Kruger, J., Hefer, E., & Matthew, G. (2014). Attention distribution and cognitive load in a subtitled academic lecture: L1 vs. L2. *Journal of Eye Movement Research, 7*(5), 1–15.

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

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods, 44*, 325–343.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

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

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Silva, B. (2019). *Learning academic words through writing: Can cognitive load affect task involvement* (Unpublished MA thesis). School of Education, University of Nottingham.

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*, 1–10. https://doi.org/10.1177/2158244017730596

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.

Chapter 8
General Discussion for Study 2 (Chapter 6) and Study 3 (Chapter 7)

8.1 Introduction

The first study reported in this book (Chap. 5) measured the receptive knowledge of academic vocabulary of first- and second-year BA-level students. It was carried out in order to find a threshold in the Vocabulary Size Test (VST; Nation & Beglar, 2007) that could reliably identify Polish higher-education learners in need of extra practice with academic words. The threshold that was identified, a VST score of 9900, revealed that 45.28% of first- and second-year Polish English majors did not possess sufficient knowledge of academic words. This is an issue that needs remediating, given the importance of academic vocabulary for success at university. Unfortunately, these often abstract and morphologically complex academic words (Corson, 1997) are typically not learnt incidentally through input alone (Knoch et al., 2015). Therefore, a more explicit and systematic approach to the teaching and learning of these words is needed, one that would promote lexical learning alongside other tasks carried out daily in the academic context. Higher-education students produce a substantial amount of written output at university, and hence these tasks were chosen for their potential to be applied in practice. To this aim, Studies 2 and 3 investigated the effectiveness of embedding academic words in sentence-writing (SW) and composition-writing (CW) tasks. Their design, results and limitations will be jointly discussed in the sections below.

8.2 Quick Review of Research Design

Studies 2 and 3 investigated how writing sentences (SW) and compositions (CW) with keywords provided in a glossary may facilitate the incidental lexical acquisition and retention of academic words. Both studies drew on Laufer and Hulstijn's (2001) involvement load hypothesis (ILH), which predicts a similar level of task-induced involvement load—and of incidental lexical learning—for SW and CW (see Chap. 2).

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_8

Incidental learning is defined in this book as learning that takes place when students perform a primary task involving the processing of some information and are not aware that they will be tested afterwards on their recall of that information. Such a definition is not without controversies; however, it is based on widely accepted definitions in the fields of SLA (e.g., Hu & Nassaji, 2016; Laufer, 2003; Pichette et al., 2012) and psycholinguistics (e.g., Craik & Lockhart, 1972; De Vos et al., 2019; Hulstijn, 2003), as discussed in Sect. 2.2.

In Studies 2 and 3 on incidental learning, participants' primary task was writing, which involved the processing of novel academic words. Participants were not aware of the true purpose of the quasi-experiment and were not told in advance that they would be tested afterwards on their recall of the words. Additionally, to ensure that any learning in Studies 2 and 3 was truly incidental, learners were given a questionnaire (see Appendix E) at the end of the quasi-experiments. Based on the results of this questionnaire, the data produced by any participant who suspected of the true purpose of the experiment and/or studied any of the keywords during or after the treatment were excluded from the analyses. Using the questionnaire, as suggested by De Vos et al. (2018), appears to be a valuable addition of my studies in comparison with previous studies in incidental vocabulary acquisition. This is because incidental-learning research has rarely, if at all, controlled for participants' perceptions of the study and for whether these participants had any extra exposure to the keywords outside experimental settings (De Vos et al., 2018; Rice & Tokowicz, 2020).

In both studies, I aimed to find out whether Polish advanced-level learners of English acquire and retain academic words to a similar degree after writing sentences and argumentative essays. First, in both studies, all participants wrote a 60-min control essay without the need to use any keyword. Then, in Study 2, each participant incorporated 20 keywords, divided into Sets A and B, each with 10 keywords. The keywords were used either in 20 sentences (SW), one keyword per sentence, or in two 60-min timed argumentative essays (Timed CW), 10 keywords per essay. In Study 3, each participant used only 10 keywords in sentences or essays, either utilizing Sets A or B, distributed randomly among participants. In addition to SW and Timed CW from Study 2, Study 3 also investigated the incidental lexical learning yielded by the writing of argumentative essays when no time limit was required (Untimed CW).

I hypothesized that the need to incorporate keywords in essay writing would increase the task cognitive load, as compared to the control essay, which may reduce lexical learning, especially for Timed CW. There were two other competing hypotheses in Study 3. First, L2 writers in the Untimed CW group may acquire more vocabulary than Timed CW learners because the lack of time pressure in Untimed CW may allow learners to allocate more attentional resources to the keywords than participants in Timed CW. Second, Untimed CW participants may acquire as much vocabulary as Timed CW L2 writers if learners in Untimed CW use the extra time available mostly to increase the quality of text production, not to focus on keyword use. In this case, learners in Untimed CW would devote as much attention to the incorporation of the keywords as learners in Timed CW—and thus the amount of

lexical learning between these groups would be similar—but would produce better quality texts.

To measure task-induced cognitive load, I utilized the following measures. Study 2 used three textual measures derived from the control and treatment essays (Timed CW): (1) holistic scores (Scores), measuring the overall quality of the essays; (2) normed errors (Errors), the number of errors per essay controlled for text size; and (3) words per minute (WPM), a measure of fluency in production. The control and treatment essays were compared in the three measures to assess changes in quality, accuracy, and fluency caused by the need to incorporate pre-specified keywords in the treatment essays. Study 3 adopted the same three measures. Additionally, in Study 3 participants answered a self-rating scale in order to assess the perceived level of difficulty, effort, and frustration between the control, Timed and Untimed CW, and SW.

8.3 Study 2 (Chap. 6): Lexical Gains and Cognitive Load

8.3.1 Lexical Gains in SW and Timed CW

In Study 2, RQ1 sought to compare the lexical learning induced by SW and Timed CW. To this end, three measures were used: VKS_6, VKS_3 and Association score. VKS_6 (pertaining to six levels of vocabulary knowledge) and VKS_3 (pertaining to types of vocabulary knowledge) focused mostly on learning new meanings of unknown or almost completely unknown words, hence measuring learning in breadth of knowledge. The Association variable, on the other hand, derived from the free association test, measured gains in depth of knowledge and also registered more learning for SW than for Timed CW. The results clearly showed more learning following the SW task than following the Timed CW task in the three dependent variables, thus in both breadth and depth of knowledge, that is VKS_6, VKS_3 (breath of knowledge) and the Association variable (depth of knowledge). Consequently, the results do not corroborate the predictions of the ILH (i.e., that SW and CW should generate similar lexical gains), neither in terms of breadth nor depth of knowledge. The findings also run counter the results of previous studies comparing these two tasks, three of which supported the ILH (i.e., Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017) while one (Zou, 2017) found that CW yielded more learning than SW. Zou's (2017) and Kim's (2008) studies have incorporated a more controlled design than the other two and will therefore be discussed below (see also Sect. 2.6 for a brief criticism of the research design in Gohar et al.'s, 2018 and Tahmasbi & Farvardin's, 2017 studies).

Zou (2017) claimed that her CW task generated more lexical processing—and therefore, learning—than SW because CW necessitates that information be chunked and organized hierarchically. That is, learners must connect the keywords and their

individual contexts semantically, creating coherent and cohesive sentences and paragraphs, and organizing these contexts into a well-argued, coherent output (i.e., the essay). As Zou (2017) claims, all this chunking and effort to compose a coherent whole likely increased processing of the keywords and consequently enhanced lexical learning. This argument, although intuitively satisfying, fails to consider learner-related and task-related aspects that may have influenced results. These aspects will be discussed in the next sections.

If increased processing is the only aspect underlying Zou's (2017) results, writing coherent argumentative essays—as reported in this book—should have been equally or more conducive to vocabulary learning than SW. This is because the higher reasoning skills necessary to produce argumentative essays (Ruiz-Funes, 2014) may have generated high levels of planning, formulating, and reviewing (Hayes & Flower, 1980), and increased the need for chunking and hierarchical organization. This would have enhanced exposure to and processing of the keywords, thus increasing learning; however, in Study 2, contrary to Zou's (2017) findings, writing argumentative essays yielded lower gains than SW.

The Role of Proficiency in Lexical Learning Through Writing. One reason for the discrepancy between my findings and Zou's (2017) findings may be participants' English proficiency level—intermediate in Zou's (2017) study and advanced in mine—and writing skills, as discussed in Sect. 3.3.1. One way that proficiency may affect lexical learning may be related to how different learners approach the composition of complex texts. As explained in Sects. 3.2 and 3.3.1, often only more experienced writers, usually more proficient writers, are able to generate well-organized ideas during the planning stage of writing (Flower & Hayes, 1981). Lower-proficiency (or less experienced) writers, by contrast, tend to devise messy, disconnected ideas in the planning stage, and these ideas will need a significant amount of organization and restructuring during the formulating and reviewing stages of complex writing. Put differently, higher-proficiency learners may process keywords considerably more when mentally planning their writing, but less during the writing proper. Conversely, lower-proficiency writers may process keywords mostly during writing by dint of the need to restructure their ideas. All this necessary writing and rewriting leads to more keyword *use*, and output production is known to promote lexical learning (e.g., Elgort et al., 2018; Joe, 1998; Pichette et al., 2012; Swain, 1985; see also Sect. 2.4). This may be one reason why the Timed CW task in Study 2 yielded less lexical learning than the CW task in Zou's (2017) study.

Another way proficiency may affect lexical learning through written output production may be directly connected to the heavy cognitive demands of the writing process (Gánem-Gutierrez & Gilmore, 2018; Manchón, 2014; Manchón & Roca de Larios, 2007; Ortega, 2012). According to Stevenson et al.'s (2006) inhibition hypothesis, such high cognitive demands may overload L2 writers' cognitive resources, particularly among inexperienced, lower-proficiency writers (see Sect. 3.3.1). Research seems to corroborate this hypothesis. For instance, Ruiz-Funes (2015) has shown that only more proficient learners with higher writing expertise are able to simultaneously tackle the linguistic and conceptual (i.e., argumentation and

coherence) demands of a complex writing task. Lower-level learners, by contrast, are overloaded by such tasks and as a result devote less attention to processes such as planning and monitoring, tending to focus instead on formal aspects of the text, including syntax and lexical use (Kellogg et al., 2013; Manchón et al., 2009; Ortega, 2012; Roca de Larios et al., 2016; Schoonen et al., 2009; Stevenson et al., 2006).

All this may mean that higher-proficiency learners are able to produce coherent compositions while allocating sufficient, but not exaggerated, attention to form (e.g., lexical items). What follows is that the advanced L2 writers in Study 2 may have split their attention between text production and keyword use. Nonetheless, as essay writing was the primary task, learners focused on the production of good quality essays while still setting aside enough cognitive resources to be able to incorporate the keywords satisfactorily. This argument is supported by the fact that the overall essay quality (as measured by holistic scores; see Sect. 8.3.2 below) of Timed CW was similar to that of the control essays, where keyword use was not required. The attention allocated to the keywords was sufficient to generate lexical learning, but not as much learning as if L2 writers had been able to focus more on keyword use, as in the SW task, where learners did not need to worry about the complexities of formal writing. By contrast, Zou's (2017) intermediate L2 writers, unable to cope both with the quality of text production and keyword use, focused mostly on the latter, as predicted by previous research (see above). This enhanced attention allocated to the keywords, together with the cyclical, recursive nature of formal writing—which increased lexical processing even further—may help explain why CW generated more learning than SW. If it is true that Zou's (2017) learners' primary focus was on keyword use, then essay quality may have suffered.

This was indeed the case, as explained in Sect. 2.6. A sample essay from Zou's (2017, p. 67) intermediate participants clearly indicates a failure to produce an accurate, coherently complex text, and shows what is seemingly a much stronger focus on the incorporation of the keywords. For instance, the essay is too short and riddled with errors, to the point of rendering several passages incomprehensible (see Fig. 2.1). To make matters worse, the keywords in Zou's (2017) example are clustered, often as lists in the same sentence, making it impossible for readers to know whether the meanings were known, and the words were used accurately by the L2 writers.

The argument that differences in participants' L2 proficiency and experience with writing may explain differences in lexical learning is further supported by a close comparison between Zou's (2017) and Kim's (2008) studies. As pointed out above, Zou's (2017) Chinese learners—non-English majors studying in Mandarin, thus likely never required to write complex essays in English—may have been cognitively overwhelmed by the writing task and hence allocated most of their attention to the keywords, enhancing learning. In other words, learners may well have written the essays mostly to incorporate words (i.e., the primary task), which may call into question the incidental nature of the study. Comparatively, Kim (2008), who found similar lexical gains following CW and SW, investigated upper-intermediate and advanced learners. These learners attended an Intensive English Program prior to starting their BA studies or were undergraduate students at the same university in the US. These participants' higher proficiency and likely broader experience in L2

writing—that is, characteristics similar to those of the participants in this book—may have enabled them to split their cognitive resources between conceptual and formal aspects of the essay (Ruiz-Funes, 2015). This reduced the attention paid to the keywords, therefore equating the learning yielded by CW and SW. Nevertheless, the argumentative essays written by my students (300–400-word long and timed) may have been more cognitively demanding than Kim's (2008) shorter, descriptive compositions (see Ruiz-Funes, 2014 for a similar argument; see Table 2.2). Thus, the argumentative essays forced my learners to allocate even more attentional resources to text production, therefore reducing the lexical learning yielded by Timed CW in Study 2.

The Role of Multitasking in Lexical Learning Through Writing. This argument tallies with the cognitive load theory (although see also Kellogg's, 1990 overload hypothesis in Sect. 3.3 for a similar argument for L1 writing). The theory states that working memory (WM) is limited and that freer attentional resources often lead to better task performance and possibly learning (Klepsch et al., 2017; Lee, 2019; Paas et al., 2003). In both Kim's (2008) study and the ones reported here, the shifting between tasks needed when multitasking (i.e., writing and use of keywords provided in a glossary) warranted the allocation of attentional resources (Kellogg, 1990; Kellogg et al., 2013; Olive, 2004, 2011; see also Robinson, 2001; Skehan, 2009, 2014). This is especially true because both tasks demanded similar WM resources (Wickens, 1981, 2008; see also Sect. 3.3.2), relying mostly upon verbal and visual WM (Olive et al., 2008). Therefore, these two tasks competed for cognitive resources, resulting in decreased task performance (see below) and less lexical learning. Such increased cognitive load might have affected my high-proficiency learners more than it affected Kim's (2008) participants, as suggested above, possibly because my argumentative essays were more complex and longer and necessitated the allocation of more attentional resources than Kim's CW task.

8.3.2 Signs of Increased Cognitive Load in Timed CW

RQ2 asked whether the need to use pre-specified keywords in essays affects the quality, accuracy, and fluency of writing. The results showed that Timed CW participants needed significantly more time (as measured by WPM) to compose texts with keywords than without in order to maintain similar text quality (as measured by Scores). Still, even after needing more time to produce such essays, participants made more errors in these essays than in the control (as measured by Errors). Consequently, in answering RQ2, evidence suggests that the need to incorporate keywords in timed argumentative essays increased the cognitive load of the task. This may help explain why the Timed CW task generated less learning than the SW task, as discussed in Sect. 8.3.

Such decrease in accuracy may also be explained by Skehan's (2003, 2009, 2014) trade-off or limited capacity hypothesis and, at least partly, by Robinson's (2001,

2005, 2007, 2011) cognition hypothesis. The trade-off hypothesis posits that attentional resources are limited, and that greater complexity in some aspects of a task may affect complexity in other aspects, as well as accuracy (see Sect. 3.4.2 for more details). For example, making more complex lexical choices (e.g., using the keywords) results in less complex and accurate syntactic constructions (see Skehan, 2009), which in Study 2 translated into higher scores in Errors. Also, time pressure (as in Timed CW) leaves little room for learners to focus on form, demonstrated here by a decrease in lexical learning and in accuracy. In the words of Ruiz-Funes (2014, p. 183), participants' "effort to meet the expectations of such register may have loaded their working memory capacity, preventing them from simultaneously attending to linguistic accuracy demands". This is in line with what has been explained about multitasking above and in Sect. 3.3.2.

In a similar vein, Robinson's (2001) cognition hypothesis postulates that manipulating tasks characteristics along the resource-dispersing dimension diverts learners' attention from language production, which decreases complexity, accuracy, and fluency in production (see Sect. 3.4.3 for more details). Here, the Timed CW task involved multitasking (i.e., essay writing and keyword use). This may be understood as tapping into Robinson's (2011) "± single task", which is a resource-dispersing feature. Accordingly, the Timed CW essays were less fluent and less accurate than the control essay (no keyword use). Again, multitasking decreased performance, which aligns well with Skehan's (2003) trade-off hypothesis and findings from dual-task research in L1 (see Sect. 3.3.2 and above). Still, there is no evidence to suggest that the multitasking induced by keyword use in Timed CW decreased the complexity of the essays. On the contrary, the holistic scores in the control and Timed CW essays did not differ significantly, suggesting similar overall quality. Nevertheless, overall quality only indirectly suggests complexity but does not equate with it. It may be that the increase in complexity of Timed CW due to the need to incorporate keywords may have been detrimental to syntactic complexity, as evidenced, for example, in a study conducted by Frear and Bitchener (2015). However, I did not adopt any measure of syntactic complexity in Study 2, and therefore, this may be considered a limitation of this study. Despite this, there is convincing evidence that keyword use increased the cognitive load of the Timed CW task, as explained above, which may have reduced lexical learning to such an extent that SW yielded more learning than Timed CW.

Considering the discussion above, essay writing may be more conducive to lexical learning than sentence writing, as in Zou's (2017) study, in at least two possible scenarios. First, writing unstructured texts where the use of keywords, not overall text quality, is treated as the primary task would direct attentional resources to the keywords, enhancing their learning. This was unsuccessfully attempted in Study 2 in the form of the unstructured Timed CW condition (see Fig. 6.4 and Sect. 6.3) and is, therefore, another limitation of this study. Second, composing without a time limit should free up writers' cognitive resources, which may enable them to produce essays that have better quality and are more accurate than essays written under time pressure. Also, the writing of untimed essays may allow learners to attend more closely to the keywords, hence enhancing learning (Kormos & Trebits, 2012; Roca

de Larios et al., 2016; Schoonen et al., 2009). To test this hypothesis, Untimed CW group was added to the research design in Study 3.

8.4 Study 3 (Chap. 7): Cognitive Load and Lexical Gains

Study 3 asked similar research questions to those in Study 2. RQ1 sought to find out whether there was a difference in the amount of learning following SW, Timed CW, and Untimed CW. RQ2 asked whether the lack of time pressure in Untimed CW reduced cognitive load when compared to Timed CW. Below, I will start by answering RQ2. Then, I will discuss RQ1 and compare the results to those in Study 2.

8.4.1 Signs of Increased Cognitive Load in Timed CW, but not in Untimed CW

The results for the Timed CW group replicate the findings from Study 2. That is, timed essays had the same overall quality as control essays but were written more slowly and less accurately than the control (see Table 7.8). Regarding Untimed CW, the textual measures showed that the overall quality of essays was similar to the control essays and higher than timed essays. Additionally, the untimed essays were more accurate than the timed essays (as measured by number of errors). In fact, despite keyword use, untimed essays ($M = 75.16$) were also more accurate than the control ($M = 85.60$), but this difference did not reach statistical significance. Finally, as expected, learners took longer writing the untimed essays than the control essays and timed essays; however, this should not be interpreted as increase in cognitive load but rather as an inherent characteristic of the Untimed CW task.

Concerning results from the self-rating scale used to assess cognitive load (see Table 7.9), participants reported higher levels of difficulty and effort for Timed CW and Untimed CW than for SW, but these differences were not statistically significant. When levels of effort were compared to the control essay, all three groups reported that writing with keywords demanded more effort than writing control essays. Nonetheless, Untimed CW participants reported less effort than the other groups, even SW. Finally, Untimed CW learners were the only to report their tasks as less frustrating than the control essays.

Overall, in answering RQ2, the results show an increase of task-induced cognitive load for the Timed CW task, as in Study 2, but not for Untimed CW task. This is because untimed essays were as accurate as and rated similarly to the control, while being more accurate and rated better than timed essays. Furthermore, untimed essays demanded less effort than timed essays and sentence writing and were less frustrating than the control and the other two tasks. To conclude, as hypothesized,

the increase in cognitive load expected when multitasking did not seem occur when writing untimed essays (see Sect. 8.3.1), or if it occurred, it was counteracted by the lack of time pressure.

As explained in Sect. 8.3.2 above, time pressure may overload learners' WM capacity, making it difficult for them to focus on form (e.g., Ruiz-Funes, 2014). Therefore, when there is no time pressure, such as when writing untimed essays, L2 writers have spare attentional capacity to focus on form, thus maintaining textual accuracy. Also, lack of time pressure allows for more pre-task and in-task planning, which has been shown to positively affect task performance. For instance, Skehan (2003), see also Foster and Skehan (1996), drawing on his trade-off hypothesis, demonstrated that the provision of planning time increased accuracy in production. Similarly, Kellogg (1990) showed that the availability of pre-task planning time increases the quality of written production. Regarding Robinson's (2001) cognition hypothesis, providing planning time, a resource-dispersing feature of task, should increase task accuracy, and this is indeed what recent meta-analytic studies (e.g., Johnson, 2017) and reviews of literature (Johnson, 2020) have shown (see Sect. 3.4.3 for more details). For example, a study conducted by Ellis and Yuan (2004) found that giving learners unpressured in-task planning time improved accuracy in production, just like the untimed essays in Study 3.

The provision of unlimited writing time allowed participants in the Untimed CW group to maintain the quality and accuracy of their essays, even though they were required to incorporate pre-specified keywords. This is a clear sign that participants attentional resources were not overloaded when writing untimed essays. This being true, as discussed in Sect. 8.3.1 above, it is possible that the advanced participants in Study 3 managed to allocate extra cognitive resources to the writing proper (i.e., the primary task)—thus maintaining overall quality and accuracy—and extra resources to keyword use, hence increasing lexical learning. Unfortunately, this does not appear to be the case.

8.4.2 Lexical Gains in SW, Timed CW, and Untimed CW: Unexpected Findings

To answer RQ1, the statistical analyses found no difference in lexical learning between SW, Timed CW, and Untimed CW, neither in breadth nor in depth of knowledge. Effectively, only the results for the variable VKS_3 (pertaining to types of vocabulary knowledge) showed a minor advantage of Untimed CW over SW, but there was no difference in learning between Untimed CW and Timed CW or between Timed CW and SW. In other words, it appears that writing argumentative essays without time pressure may be more conducive to learning than SW only, and only slightly and in breadth of knowledge, not in depth. These results mostly support the predictions of the ILH (Laufer & Hulstijn, 2001), since generally the learning following SW was similar to the learning yielded by CW, and findings from previous

research (Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017). Nevertheless, once again the results do not support Zou's (2017) findings (i.e., that CW yields more learning than SW). The findings also fail to replicate the results from Study 2, where SW was more conducive to vocabulary learning than Timed CW. These findings are discussed in detail below, starting with the untimed essays.

It was unclear whether untimed essays would yield more learning than or the same amount of lexical learning as Timed CW. In Sects. 3.4.3 and 4.3.3, I hypothesized that (1) untimed essays might be more conducive to learning than untimed essays, as learners freer attentional resources would allow them focus more on the keywords; or, that (2) untimed and timed essays might generate similar levels of lexical learning as long as participants decide to use the extra time available to allocate extra attention to text development (the primary task), not to keyword use (the secondary task). The results support the second hypothesis. Indeed, the untimed essays received higher scores and were more accurate than the timed essays (see above), which indicates that the participants did utilize the extra time available to focus on text quality. In fact, this preference for the primary task is a sign that learners did not suspect the true purpose of the quasi-experiment, otherwise they may have used the extra time to learn the keywords. This provides supporting evidence for the incidental nature of the current study.

Still, since the lack of time pressure allowed learners to improve text quality, it stands to reason that they spent more time planning, formulating, and reviewing the text than writers in Timed CW. This undoubtedly increased contact with the keywords owing to the higher amount of planning and re-planning, writing and rewriting, and constant reviewing. On the one hand, this forced learners to re-read the keywords repeatedly, therefore enhancing exposure to input. On the other, this likely made learners rewrite some of these keywords a few times, thus increasing output production. That is, even having decided to use the extra time to improve text quality, not to focus on keyword use, Untimed CW learners had more exposure to and practice of the keywords (relative to Timed CW learners). So why did this not increase learning—making Untimed CW more conducive to lexical acquisition than Timed CW—when exposure to input and output production are known to enhance learning?

Why Enhanced Input Failed to Improve Learning in Untimed CW. Obviously, during the recursive writing process, learners read the words in the glossary, in the examples in the glossary, and in the text, more so when composing untimed essays since they spent longer writing (thus reading) than when composing the timed essays. One could argue that this is comparable to encountering novel words repeatedly when reading a text, which is known to facilitate learning. There is a large body of research showing that the number of occurrences of a novel word in a text correlates positively with lexical learning, meaning that the more a word is repeated in a text, the higher learning tends to be (e.g., Brown et al., 2008; Chen & Truscott, 2010; Horst et al., 1998; Hu, 2013; Kweon & Kim, 2008; Malone, 2018; Pigada & Schmitt, 2006; Rott, 1999; Saragi et al., 1978; Teng, 2019, 2020; Vidal, 2011; Waring & Takaki, 2003; Webb, 2007, 2008; see also Webb, 2019, for a recent literature review on incidental

learning through input). In fact, according to Uchihara et al.'s (2019) recent meta-analysis, the correlation between the number of occurrences of novel words and their learning is on average moderate ($r = 0.41$). However, a higher number of repetitions increases learning because readers need to infer the meaning from diverse contexts and retrieve the meanings of already correctly inferred words. None of these processes take place when writing essays with keywords provided in a glossary, as in the studies presented in this book.

First, when inferring a word from a written text, readers need to process this word within its context, that is, they need to take advantage of the contextual information provided in order to infer the meaning of the word (e.g., Elgort & Warren, 2014; Frishkoff et al., 2010; Mulder et al., 2019 see also Sect. 2.3). When this happens, and assuming the word is inferred correctly, the novel word is processed semantically (i.e., deeply), which, as postulated by Craik and Lockhart's (1972) depth of processing hypothesis, facilitates learning (see Sect. 2.4). Obviously, when writing with keywords provided in a glossary, contextual inferencing does not occur. Second, when a novel word appears several times in a text, they occur in semantically different (diverse) contexts, not in that one context composed by the essay writers. Words that appear in diverse contexts are better acquired than words that appear in non-diverse contexts (e.g., Elgort et al., 2015; Joseph & Nation, 2018). Last, once a novel word is inferred correctly, its meaning will possibly be retrieved each time the word appears in a text. Such retrieval has been demonstrated to enhance lexical learning and retention (e.g., Elgort et al., 2015; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006; Soderstrom et al., 2016; Van den Broek et al., 2018; see also a meta-analysis by Rowland, 2014). For instance, Van den Broek et al. (2018) conducted three studies comparing the lexical learning yielded by inferencing and retrieval. The results showed that retrieval was more conducive to lexical retention than inferencing in all three studies. That is, participants retained words better when they were inferred correctly and then retrieved successfully in a subsequent context than when they were inferred correctly in multiple contexts.

When writing with keywords provided in a glossary, whether the essay is timed or untimed, learners consult the glossary to incorporate the keywords in their writing. The essay—and therefore all the sentences containing the keywords—is read several times as learners strive to produce high quality texts (as in Study 2 and 3). All this reading increases exposure to the keywords, but does not promote inferencing; it also does not provide multiple diverse contexts with the same keyword, as each keyword was used only once in the essay; finally, it does not induce retrieval, since even if learners forget the meaning of the keyword they have just used, they will likely consult the glossary (i.e., read), not attempt to retrieve the meaning of the word, thus processing the word more deeply (Craik & Lockhart, 1972). Consequently, at least when learners treat essay writing as the primary task, it seems irrelevant whether they spend less time (Timed CW) or more time (Untimed CW) reading the essays: The extra exposure in untimed essays may not increase processing, and therefore learning, of the keywords. Still, the extra time available in Untimed CW probably resulted in more rewriting, including the rewriting of passages containing the keywords. Yet, this extra output production did not enhance learning. This finding is discussed below.

Why Enhanced Output Production Failed to Improve Learning in Untimed CW. It is possible that when writing, only few, if any, of the keywords were re-written a few times in multiple original contexts until learners decided on a final sentence to keep in the essay. This makes sense since the writers appeared to be more concerned with the text itself, the primary task, not with keyword use. In this case, it is likely that once the writers had been able to create one sentence with a keyword, this sentence was maintained throughout the writing process. If the sentence was not preserved in its exact same form, it was likely kept in a semantically similar (i.e., non-diverse; see above) context. That is, the sentences containing keywords may have been written in original contexts mostly once, in a similar fashion to what may have happened in the Timed CW task.

Nonetheless, it is precisely the generation of original contexts that promotes learning through output, as discussed in Sects. 2.4 and 2.5. For instance, Joe (1998) drew on the construct of generation (Slamecka & Graf, 1978; Wittrock, 1974) to conclude that a more creative incorporation of lexical items into original contexts resulted in higher lexical acquisition. Possibly, in timed and untimed essays, such creative use occurred only once with most keywords, if not all of them. Similarly, the ILH posits that *evaluation* is strong only when words are incorporated in original contexts owing to the high levels of semantic (i.e., deep) processing involved. However, such deep processing may occur only when the keyword is used in the first, perhaps only, original context (i.e., not when sentences are merely re-written to improve form, as may have been the case in Timed CW and Untimed CW). In a related vein, Swain's (1985, 1995, 2000), Swain and Lapkin (1995) output hypothesis has long postulated that oral or written production enhances learning when learners experience communication breakdowns or realize that they have linguistic problems in their output. This, the hypothesis states, pushes learners to modify their output, deepening language processing (Craik & Lockhart, 1972) and thus enhancing learning. Again, in the case of my studies, it is possible that while pushed output pervaded the writing process, most of the keywords were not responsible for communication breakdowns, at least not more than once (i.e., when they were first incorporated into the essay).

Summing up, Untimed CW did not yield more learning than Timed CW. This was true even though participants spent significantly more time writing untimed essays, and thereby were likely more exposed to and used the keywords more times than Timed CW participants. Higher exposure did not increase learning in that no inferencing was needed, contexts were non-diverse, and retrieval was not induced. Production failed to improve learning in Untimed CW likely because the keywords were used in original contexts mostly once, and communication breakdowns induced by the keywords, if they occurred, they occurred in a similar number to that in Timed CW. Obviously, these assumptions need further research to be confirmed (see Chap. 9). And importantly, this rationale considers that L2 writers treated essay writing as the primary task, as in Studies 2 and 3 reported here. Had learners focused mostly on keyword use, as appears to have been the case in Zou's (2017) study, Untimed CW would likely have enhanced lexical learning.

Why Statistical Analyses May Be to Blame. It is also possible that Untimed CW did not yield more learning than Timed CW (and SW) in Study 3 because of problems with the statistical analyses. Indeed, in all three variables analyzed, gains for Untimed CW were higher than for Timed CW and SW, particularly in VKS_3 and Association (see Tables 7.2, 7.4, and 7.6 for the descriptive statistics of VKS_6, VKS_3, and Association, respectively). The variable VKS_3 combined scores 1 and 2 (no knowledge or ability to recognize the word), 3 and 4 (ability to recall meaning with less or more confidence), and 5 and 6 (full productive knowledge with or without mistakes). Therefore, VKS_3 measured different types of lexical knowledge (i.e., form recognition, meaning recall, and productive knowledge). As shown in Table 7.4, most learning in VKS_3 between the pretest and posttest for all groups occurred between level 1 (no knowledge or recognition) to level 3 (productive knowledge). The proportion of level 3 scores increased by 23% for Untimed CW, 18% for Timed CW, and 19% for SW, and yet the statistical model showed only a minor advantage of Untimed CW over SW. Table 7.6 also shows more pretest–posttest gains in Association for Untimed CW over the two other groups. In a variable whose maximum score was 4, Untimed CW mean scores increased by 0.66, while Timed CW showed gains of 0.37, and SW of 0.39. Additionally, between the pretest and posttest, the median score for Untimed CW jumped from 0 to 2, while it remained the same for Timed CW and SW (i.e., 0). Comparatively, in Study 2, Association scores for SW and Timed CW increased 0.47 and 0.24, respectively, while the median increased from 0 to 1 for SW while remaining unchanged for Timed CW. In other words, the difference in SW and Timed CW gains in Study 2 was narrower than the difference between Untimed CW and the other groups in Study 3. Yet, the gains were only statistically significant in Study 2.

As a result, it is possible that in Study 3, the statistical analyses failed to find a significant difference between the groups when in fact one exists (i.e., a type 2 error; see Field, 2017; Perry, 2011). Put differently, the statistical models may have lacked sufficient power to identify a significant difference between the groups (i.e., Hajduk, 2019; Howell, 2010; Salkind, 2011). In fact, when discussing the reliability of linear mixed models (LMMs), Meteyard and Davies (2020) drew attention to the pervasive, and yet unfortunate, lack of statistical power in psychological research. To overcome this issue, the researchers recommended a minimum of "30–50 participants, and 30–50 items or trials for each of those participants completing each condition" (Meteyard & Davies, 2020, p. 17). Brysbaert and Stevens (2018) have made a similar recommendation, that is, a minimum of 40 participants and 40 items per participant. Study 2, which likely had sufficient power, had 39 participants included in the analyses, each of whom incorporated 20 keywords in their essays or sentences. Thus, the participants sat pretests and posttest with 20 items each. Therefore, in Study 2, there were 40 items (data points) per participant, as per recommendation (see Sect. 4.4.2 for more details on how the data are organized in LMMs). By contrast, the 90 participants in Study 3 used only 10 keywords in their task, thus totaling only 20 items per participant. In addition to the lack of data, there is also the fact that generalized LMMs (GLMMs), used in Studies 2 and 3, are often more unstable than LMMs (Clark, 2019).

Indeed, the process of computing the three GLMMs in Study 3 showed some signs of lack of stability in the models, most likely due to the insufficient data (Meteyard & Davies, 2020) described above. For example, when building the model for VKS_6, the tenth model (see Appendix N for the model-building process) was unable to compute a reliable confidence interval for the working memory (WM) covariate and for the Time * Untimed_CW interaction. The 95% CI found for the estimates were, respectively, (−1823.61, 1823.60) and (−15,693.36, 15,693.12). These enormous, thus unreliable, CIs attest to the lack of precision (hence, the lack of power) of the analysis (Field, 2017). I solved this problem by eliminating the intercept for Class from the model, thus achieving more realistic CIs for the estimates: (−0.065, 0.056) for WM and (−0.459, 0.220) for the Time * Untimed_CW interaction. Still, the fact remains that the data were not as stable as in Study 2.

If the above is true, then the lack of statistical power is a limitation of Study 3. Unfortunately, as explained in Sect. 4.4.1, LMMs are still a novelty, and when designing the study, I was unaware of the recommendations put forward by Meteyard and Davies (2020) and Brysbaert and Stevens (2018). Future research should address this issue. It is also worth noting that if the lack of power is the reason why the analyses did not report significantly higher gains for Untimed CW over Timed CW and SW, then the rationale discussed above (on the role of input and output on learning) may be flawed. This means that Untimed CW participants may have indeed re-written sentences with keywords in semantically diverse contexts, hence increasing processing and learning. The only way to find out whether this is the case is by addressing this issue in future research. This possibility will be discussed in more detail in Chap. 9.

Results for SW and Timed CW. It remains unclear why SW and Timed CW yielded similar lexical gains in Study 3, whereas SW generated more learning than Timed CW in Study 2. One possibility is the lack of power in Study 3, as discussed above. However, this appears unlikely. The descriptive statistics shown in Tables 7.2, 7.4, and 7.6 in Study 3 (see also the section above) reveal almost identical gains for SW and Timed CW in VKS_6, VKS_3, and Association. Another possibility relates to the fact that in Study 2 each participant wrote sentences and essays with 20 keywords, while in Study 3 they used only 10 (although they utilized both Sets A and B, randomly distributed among participants). The results for the random effect (lexical) Items were highly significant in both studies for all variables, showing that different keywords generated different levels of learning (see Tables 6.4, 6.6, and 6.8 for results from Study 2; Tables 7.3, 7.5, and 7.7 for results from Study 3). Thus, it could be that the results in Study 2 and 3 differ exactly owing to the difference in the keywords. Nevertheless, this again is unlikely for at least two reasons. First, as discussed in Sect. 6.2.5, Sets A and B were statistically similar in every respect (i.e., frequency, concreteness, length, and part of speech), and thus should not produce different results. Second, the inclusion of Items as a random effect in all models ensured that the difference in learning among keywords was controlled for.

A final reason why in Study 3 Timed CW and SW generated similar vocabulary learning may relate to participants' lack of appreciation for the SW task. The two

variables measuring frustration derived from the self-rating scale showed that writing sentences was perceived as more frustrating than writing timed or untimed essays (see Table 7.9). This was somewhat surprising, but as explained in Sect. 7.5, these findings may reflect learners' dislike for, not the level of cognitive load induced by, the SW task. Consequently, it is possible that in Study 3 SW did not yield more learning than Timed CW because learners did not take the SW task as seriously as they should have done.

Unfortunately, the self-rating scale was not implemented in Study 2, so the results cannot be directly compared to those in Study 3. Yet, part of the data analyzed in Study 3 was taken from Study 2 (see Sect. 7.2.2), and this is crucial for the following reason. In Study 2, participants wrote sets of 10 sentences twice, in two different days (see Fig. 6.4), just like they wrote two essays (i.e., the unstructured and structured essays). Logic dictates that if learners disliked the SW task in Study 2, as they did in Study 3, they disliked it more the second time (i.e., the second day) they were required to write the sentences. As it happens, the data from Study 2 that was analyzed in Study 3 was exactly the data from the second day, when levels of dissatisfaction with the SW task were probably higher (Sect. 7.2.2 explains why only the data from the second day were included in Study 3). Consequently, it stands to reason that the subset of data from Study 2 that was analyzed in Study 3 was the subset that registered lower gains for SW. This decreased overall SW gains in Study 3 and equated the learning to that of Timed CW. However, there is no statistical evidence to substantiate this explanation, and therefore, further research is needed. Still, if this explanation is true, this is another limitation of Study 3.

8.5 The Use of the Keywords in the Essays: A Qualitative Analysis

In addition to answering the research questions from Studies 2 and 3, I analyzed a random subset of essays from these studies in order to better understand how the keywords were used. This qualitative analyzes is reported in this section. First, I will give examples demonstrating how the vast majority of the keywords were properly utilized in essay writing. The next sub-section will then shift attention to some keywords that were erroneously incorporated in the timed and untimed essays. These analyses are important findings that may reveal problems with the research design here and in previous studies, and therefore, may contribute to the betterment of future research.

8.5.1 The Proper Use of Keywords

A qualitative analysis of timed and untimed essays in Studies 2 and 3 revealed that the overwhelming majority of keywords were incorporated accurately and spread out evenly across the text. Below, there are few unedited examples taken from timed and untimed essays from both studies. The topic chosen is the one used in Study 3 and in the structured condition in Study 2: "Do you agree or disagree with the following statement? '*Parents are the best teachers*'. Use specific reasons and examples to support your opinion". The keywords and necessary prepositions are highlighted in bold. The vocabulary set used is indicated in brackets. The authors of the extracts are kept anonymous.

Timed Essays:

Example 1: *Furthermore, the parents can sometimes be too strict with their children, which can lead children to have many* ***constraints*** *, which might limit their development and actually be harmful. Thus, it is important that the parents* ***differentiate*** *their role as their children's teacher* ***from*** *being an actual oppressor.* (Set A)

Example 2: *Moreover, children usually* ***incorporate*** *the moral code they had been brought up in* ***into*** *their future family life. This creates a continuity in preserving specific values, which one consders important. At school, they are exposed to different types of characters, methods and traditions, therefore they have* ***insight into*** *the way other families work.* (Set A)

Untimed Essays:

Example 1: *Due to this fact, their role model will impact on us in one way or another because its* ***acquisition*** *will be taught in us* ***implicitly*** *as we grow.* (Set A)

Example 2: *Children can* ***derive*** *a great amount of things* ***from*** *their parents. When parents take time to teach their kids anything it* ***reinforces*** *family's connection, but also has a huge impact on the child's future.* (Set B)

As can be seen in the examples above, the keywords were not chunked together in few sentences (see Appendix J for two examples of whole essays). This is important because it shows the reader that writers understood the meaning of the keywords and were able to use them mostly accurately. At the same time, it demonstrates the writers' effort to compose well-structured essays, which was expected since the essay was the primary task. Compare this to an extract of the example provided by Zou (2017, p. 67) from one of her participants (see Fig. 2.1 for the entire example):

> People who has **assiduous trait** will get **lassitude** easily. Because they usually worry about some events are **indispensable, pernicious** or **apprehensive**. Sometimes, it is unnecessary to care too much about them, because these events are **ostensible**. (Keywords in bold)

Zou's (2017, p. 57) instructions for her CW task were the following: "Write a composition that coherently connects the 10 target words, and correct use of all words was required for task completion". The extract contains 36 words, seven of which (19.44%) are keywords. As discussed in Sect. 2.6, it is impossible for the reader (and the researcher) to know whether the meanings had been understood and whether the keywords were used accurately. There simply is not enough context surrounding the keywords to enable the reader to draw these conclusions. Therefore,

Zou's (2017) claim that the keywords were connected coherently and used accurately is untenable.

There are at least two reasons why her participants may have clustered the keywords so closely together. First, the keywords were rather advanced for Zou's (2017) intermediate learners, even more advanced than the keywords used in Studies 2 and 3 reported in this book, which investigated advanced learners. The average frequencies in COCA for the words in Sets A and B were, respectively, 8.57 (SD = 0.47) and 8.66 (SD = 0.69). The seven keywords in the extract above (out of a total of 10 equally difficult keywords) had in the same corpus a mean frequency of 5.99 (SD = 1.38). Only one of the keywords (i.e., "trait") is of relatively high frequency (therefore less advanced). It is possible that Zou's (2017) keywords were too difficult for the learners, and thus they needed to cluster the words together in order to be able to incorporate them in the text. The conclusion is that future research exploring learning through writing should pay more attention to the keywords chosen to avoid overly advanced words which L2 writers may be unable to cope with. A second reason for the clustering of the keywords in Zou's (2017) sample essay maybe the rather short text, which totaled 74 words. On the one hand, Zou did not stipulate a time limit for the writing (see Table 2.2). On the other, the CW task instructions did not require a minimum or a maximum text length. Clearly, her participants opted for short texts, despite the lack of time limit, texts that were too short to incorporate the 10 keywords adequately. Future research should bear this in mind and require a minimum text length, which could go a long way in avoiding the chunking of keywords.

8.5.2 The Improper Use of Keywords

The qualitative analysis of the essays in Studies 2 and 3 found that few keywords in Sets A and B were, at times, used improperly. These words were the following: Set A: "paradigm", "constitute", and "incorporate"; Set B: "derive", "affective", and "variability". Of course, these keywords were usually used accurately, as shown in the extracts above. Still, they were used erroneously enough times as to stand out from the others. Some unedited examples can be found below (keywords in bold; vocabulary set in brackets):

Timed Essays:

Example 1: *Parents can be **paradigms** to their children as role-models.* (Set A)

Example 2: *Our law **constitutes** that you cannot drink and drive.* (Set A)

Example 3: *Most of them are just **incorporate** in some offices or any other institute where they are working, for example.*

Example 4: *The drug is **derived** from Colombia; or I finally **derived** a good mark*. (Set B)

Example 5: *It is very **affective** when parents teach children.* (Set B)

Example 6: *On the other hand, well qualified teachers might give a wide* ***variability*** *of scientific knowledge.* (Set B)

Untimed Essays:

Example 1: *The* ***paradigm*** *of education children learned from their parents at home.* (Set A)

Example 2: *Parents are* ***constituted*** *to be the biggest influence on people.* (Set A).

Example 3: *In our generation it is easier for children to* ***derive*** *information through other ways such as the internet*. (Set B)

Example 4: *Parents may not always have an* ***affective*** *way of teaching due to bias.* (Set B)

Example 5: *The* ***variability*** *of positive and negative experiences that people had with their parents.* (Set B)

It was expected that some of the keywords would be used incorrectly, even though the glossary brought the definition and two examples for each keyword (see Appendix H for the glossaries). One reason for this is that the keywords in Studies 2 and 3 were academic words, which, as explained in Sect. 1.5, are typically abstract and morphologically complex (Corson, 1997; Lubliner & Hiebert, 2011; Vidal, 2011), which makes them difficult to be understood and used. For instance, the words "paradigm" (Set A) and "derive" (Set B) may have been used incorrectly because the definitions provided in the glossary were insufficient to generate a clear understanding of the words' meanings (even though the definitions were taken from reputable advanced learners' dictionaries; see Sect. 6.2.5). One of the two definitions of "paradigm" in the glossary was "a model of something, a very clear and typical example of something". Considering the examples above, it appears that some learners understood "paradigm" as a synonym of "example" or "type". Similarly, the definition for "derive", that is, "to get something from something else", may have led some participants to believe this word to be a synonym of "get", "obtain", or "come from". Again, these mistakes occurred even though two examples were provided for each keyword. This points to the importance of giving clear definitions and examples in order to ensure accurate use of the keywords.

Yet, previous research has provided only short definitions lacking in specificity and no examples. For instance, Kim's (2008, p. 325) glossary defined the keyword "vexed" as an adjective meaning "worried; distressed". Zou's (2017, p. 75) glossary defined "lassitude" as a noun meaning "a state of tiredness". "Vexed" and "lassitude" are not academic words but are highly advanced and abstract words. As a result, it is likely that their participants faced similar difficulties to mine. Zou's (2017) extract shown in Sect. 8.6.1 above illustrates possible difficulties with the keywords. Unfortunately, Kim (2008) provided no sample essays. All this shows that future research should try to address this issue by more carefully devising, and piloting, glossaries.

Other keywords from Sets A and B may have been misused for a different reason. Some examples are the words "constitute" and "incorporate" in Set A, and "affective" and "variability" in Set B. In these cases, it is possible that learners misunderstood the words not only because of the aforementioned insufficiency of the glossary, but

also because the words resemble other existing words in English. These types of words are called *synforms*, a term coined by Laufer (1988 as cited in Peters, 2019, p. 127). In Studies 2 and 3, "constitute" may have reminded participants of "constitution", while "incorporate" was likely confused with "corporate", or "corporation"; "affective" may have conjured up the word "effective", and "variability" may have been misconstrued as meaning "variety". The examples provided above show these misunderstandings clearly. The takeaway is fairly straightforward: When selecting keywords, future research on lexical learning through writing should avoid any word that resembles another (or other) existing word(s) in the target language. Even though Laufer (1988, as cited in Peters, 2019) has long drawn attention to the problems associated with using synforms in research, many studies do not appear to control for this confounding variable (for examples of such studies see Chang et al., 2020; Chen & Truscott, 2010; González-Fernández & Schmitt, 2020; Hui, 2020; Kweon & Kim, 2008; Pellicer-Sánchez & Schmitt, 2010; Vidal, 2011; Webb, 2007). Not controlling for synforms was an unfortunate omission of Studies 2 and 3 that I intend to redress in future research. Still, this omission is unlikely to have distorted scores in that each vocabulary set contained the same number of such words.

8.6 Issues in Proceduralizing Incidental Lexical Learning

Before ending this discussion, this chapter will analyze in more depth some issues underlying the definition of incidental learning. In Sect. 2.2, this book drew attention to some of the controversies surrounding this definition. There, I made it clear that some researchers believe that incidental learning may occur via input only (e.g., Krashen, 1989; Webb, 2019) while others maintain that output production may also generate incidental lexical learning (e.g., Hu & Nassaji, 2016; Hulstijn & Laufer, 2001; Laufer, 2003; Ortega, 2009). I also underscored that incidental learning is often contrasted with intentional learning (i.e., any learning that results from a deliberate attempt to memorize words during a task). However, in line with other researchers, I argued that the definition of incidental learning as lacking intentionality is problematic in that it is impossible, in any input or output task, to fully rule out learners' intention to commit words to memory (e.g., De Vos et al., 2018; Elgort et al., 2017; Webb, 2019). Then, after following common definitions in the fields of SLA and psycholinguistics, I stated that incidental learning may occur through input or output as long as participants perform the primary task while processing information (here, keywords) without being aware of the true purpose of the experiment and without being told in advance that they will be tested afterwards on their recall of that information. Studies 2 and 3 in this book met these criteria. Additionally, as suggested by De Vos et al. (2018), the post-experiment questionnaire ensured that learners who suspected of the true purpose of the studies and/or studied the keywords prior to the posttest were eliminated from the analyses.

Still, incidental learning through writing is more explicit than incidental learning through input (e.g., reading). This is because, in an experiment, in order to make

learners use novel words accurately in writing, researchers must provide glossaries, just as learners would consult an outside source (e.g., dictionaries) when writing texts in real life (see Sect. 3.3.2). However, more explicit learning cannot be deemed intentional learning, since in the latter learners would need to be directly encouraged to make an effort to remember the keywords. Put differently, learning through writing, as proceduralized in Studies 2 and 3, is more explicit than learning through input, but not intentional. As a result, one could argue that the learning yielded by the SW and CW tasks was *semi-incidental*, not *fully incidental*, considering that the learning processes underlying writing are certainly different, at a minimum more explicit, than those of reading. "Semi-incidental" is not an existent construct in the literature, but it is not difficult to imagine the usefulness of differentiating fully incidental learning from semi-incidental learning in future research.

A fully incidental reading task could be proceduralised as simply a reading task with a posttest, not a pretest, measuring learning. Still, to achieve this, researchers would need to embed in the text either rare keywords, or plausible nonwords, to ensure participants have no prior knowledge of the lexical items being tested. Nonetheless, the rarity of the keywords and nonwords—the latter may be perceived similarly to rare words since they would be unknown to all participants—would make these keywords more salient in the text, thus enhancing learning. In fact, research has shown that nonwords may benefit from repetition in reading more than real words do because of their salience (e.g., Uchihara et al., 2019). Researchers may also opt for employing less advanced keywords, but in this case a pretest would need to be used to measure existing knowledge. Here, it may be argued that the pretest could draw attention to the purpose of the experiment, hence making the experiment not fully incidental. A solution is to pretest participants weeks, or even months, before the experiment, even before leaners become aware of the existence of an experiment. But in this case, it would be impossible to account for any learning occurring prior to the commencement of the experiment proper, not to mention the ethical issues involved in pretesting learners without their consent.

Ensuring fully incidental learning through reading becomes yet more problematic when researchers include other variables in the analyses. For instance, one may wish to explore the influence of learners' knowledge of lexical learning strategies on their ability to learn incidentally through input. Also, researchers may be interested in whether the amount of vocabulary knowledge in an L2 correlates with incidental lexical learning in an L3. In both cases, extra research instruments must be added to the design, and these instruments may inadvertently draw attention to the true purpose of the experiment.

In other words, it might be impossible to design fully incidental studies, even when measuring learning through input, if a criterion for a study to be considered fully incidental is that it simulates real-life contextual word learning (e.g., without rare or nonwords, tests, and controlling variables). The more variables are explored and need to be controlled for, therefore increasing the reliability of the results, the more likely the study is to draw attention to the purpose of the experiment, and to increase the explicitness of the learning process. This being the case, incidental learning may be better understood as a continuum, with study designs falling anywhere between fully

incidental and semi-incidental. This continuum may also be applicable to research exploring the incidental lexical learning through writing, although such studies would almost certainly be always closer to the semi-incidental end of the scale.

Acknowledging the existence of such a continuum may improve the interpretability of future research findings. However, it would be necessary to develop a valid and reliable tool that can objectively classify study designs within this continuum. Should such tool be developed, it may be used as a measure of level of explicitness, with study designs that are closer to being fully incidental being less explicit. Level of explicitness could then be included in statistical models as a covariate, thereby controlling for the influence of this variable on incidental learning and increasing the reliability of the findings.

8.7 Conclusion

This chapter has discussed in detail the findings from Studies 2 (Chap. 6) and 3 (Chap. 7) and outlined extra findings that may help improve future research. Study 2 compared the academic vocabulary learning yielded by writing sentences (SW) and writing essays under time pressure (Timed CW); Study 3 replicated this design and added one more condition: Untimed CW.

Study 2 found that SW generated more lexical learning than Timed CW. Also, the study showed that writing essays with keywords took longer and was less accurate than writing essays without the need to incorporate pre-specified keywords (i.e., the control essay). This was taken to suggest that writing essays while having to incorporate keywords (i.e., multitasking) may have increased cognitive load, which may help explain why the Timed CW condition yielded less academic vocabulary learning than SW. Put differently, writers may have been overwhelmed by the writing process and the multitasking needed, and thus had to allocate their limited attentional resources to the writing proper, not to keyword use, hence reducing lexical learning. This being the case, it was hypothesized that writing essays without time pressure (the Untimed CW condition added to Study 3) may free attentional resources and as a result enhance vocabulary acquisition.

Study 3 followed a similar design but failed to replicate the findings from the previous study. First, the results suggest that removing the time pressure from the essay-writing task may have indeed reduced the cognitive load of the task since untimed essays had overall higher quality and were more accurate than timed essays. Nevertheless, learning was only marginally higher for Untimed CW than for SW and Timed CW. Moreover, SW and Timed CW registered similar learning, which goes counter to the results from Study 2. The reason for these conflicting results remains unknown, but some possibilities include lack of power in the statistical analysis from Study 3 and lack of enhanced processing of the keywords in the Untimed CW condition. The next chapter concludes this book, underscores practical implications, outlines several research limitations, and makes suggestions for future research.

References

Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition through reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language, 20*(2), 136–163.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1).

Chang, R., Yang, X., & Yang, Y. (2020). Prediction differs at sentence and discourse level: An event-related potential study. *Applied Psycholinguistics*, 1–19. https://doi.org/10.1017/S0142716420000235

Chen, C., & Truscott, J. (2010). The effects of repetition and L1 lexicalization on incidental vocabulary acquisition. *Applied Linguistics, 31*(5), 693–713.

Clark, M. (2019). *Mixed models for big data: Explorations of a fast penalized regression approach with mgcv.* Retrieved from https://mclark.github.io/posts/2019-10-20-big-mixed-models/. Accessed November 27, 2020.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718.

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

De Vos, J. F, Schriefers, H., & Lemhöfer, K. (2019). Noticing vocabulary holes aids incidental language word learning: An experimental study. *Bilingualism: Language and Cognition, 22*(3), 500–515.

De Vos, J. F., Schriefers, H., Nivard, M. G., & Lemhöfer, K. (2018). A meta-analysis and meta-regression of incidental second language word learning from spoken input. *Language Learning, 68*(4), 906–941.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2017). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition, 40*, 341–366.

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

Elgort, I., Perfetti, C., Rickles, B., & Stafura, J. (2015). Contextual learning of L2 word meanings: Second language proficiency modulates behavioral and ERP indicators or learning. *Language, Cognition and Neuroscience, 30*(5), 506–528.

Elgort, I., & Warren, P. (2014). L2 Vocabulary learning from reading: Explicit and tacit lexical knowledge and the role of learner and item variables. *Language Learning, 64*(2), 365–414.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26*, 59–84. https://doi.org/10.1017/S0272263104261034

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365–387.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299–323.

Frear, M. W., & Bitchener, J. (2015). The effects of cognitive task complexity on writing complexity. *Journal of Second Language Writing, 30*, 45–57.

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2010). Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology, 35*(4), 376–403.

Gánem-Gutierrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning, 68*(2), 469–506.

Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique feature analysis or involvement load hypothesis: Estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research, 47*, 859–869.

González-Fernández, B., & Schmitt, N. (2020). Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics, 41*(4), 481–505.

Hajduk, G. K. (2019). *Introduction to linear mixed models.* Retrieved from: https://ourcodingclub.github.io/tutorials/mixed-models/#crossed. Accessed November 27, 2020.

Hayes, J. R., & Flower, L. S. (1980). *Identifying the organization of writing processes.* Retrieved from: https://www.researchgate.net/publication/200772468_Identifying_the_organization_of_writing_processes. Accessed September 6, 2020.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207–223.

Howell, D. C. (2010). *Statistical methods for psychology* (7th ed.). Cengage Learning.

Hu, M. H.-C. (2013). The effects of word frequency and contextual types on vocabulary acquisition from extensive reading: A case study. *Journal of Language Teaching and Research, 4*(3), 487–495.

Hu, M.H.-C., & Nassaji, H. (2016). Effective vocabulary learning tasks: Involvement load hypothesis versus technique feature analysis. *System, 56*, 28–39.

Hui, B. (2020). Processing variability in intentional and incidental word learning: An extension of Solovyeva and DeKeyser (2018). *Studies in Second Language Acquisition, 42*, 327–357. https://doi.org/10.1017/S0272263119000603

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell Publishing.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539–558.

Joe, A. (1998). What effects to task-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics, 19*(3), 357–377.

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: A research synthesis and meta-analysis. *Journal of Second Language Writing, 37*, 13–38. https://doi.org/10.1016/j.jslw.2017.06.001

Johnson, M. D. (2020). Planning in L1 and L2 writing: Working memory, process, and product. *Language Teaching, 53*, 433–445.

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology, 166*, 190–211.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands. *The American Journal of Psychology, 103*(3), 327–342.

Kellogg, R. T., Whiteford, A. P., Turner, C. E., Cahil, M., & Mertens, A. (2013). Working memory in written composition: An evaluation of the 1996 model. *Journal of Writing Research, 5*(2), 159–190.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, 1–18. https://doi.org/10.3389/fpsyg.2017.01997

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university. *Journal of Second Language Writing, 28*, 39–52.

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality, and aptitude in narrative task performance. *Language Learning, 62*(2), 439–472.

Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence by the input hypothesis. *The Modern Language Journal, 73*(4), 440–464.

Kweon, S., & Kim, R. (2008). Beyond raw frequency: Incidental vocabulary acquisition in extensive reading. *Reading in a Foreign Language, 20*(2), 191–215.
Laufer, B. (2003). Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence. *Canadian Modern Language Review, 59*(4), 567–588.
Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1–26.
Lee, J. (2019). Time-on-task as a measure of cognitive load in TBLT. *The Journal of Asia TEFL, 16*(3), 958–969. https://doi.org/10.18823/asiatefl.2019.16.3.12.958
Lubliner, S., & Hiebert, E. H. (2011). An analysis of English-Spanish cognates as a source of general academic language. *Bilingual Research Journal, 34*(1), 76–93.
Malone, J. (2018). Incidental vocabulary learning in SLA: Effects of frequency, aural enhancement, and working memory. *Studies in Second Language Acquisition, 40*, 651–675.
Manchón, R. M. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 27–52). John Benjamins.
Manchón, R. M., & Roca de Larios, J. (2007). On the temporal nature of planning in L1 and L2 composing. *Language Learning, 57*(4), 549–593.
Manchón, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal and problem-solving nature of foreign language composing processes: Implications for theory. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 102–129). Multilingual Matters.
Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language, 112*, 1–22. https://doi.org/10.1016/j.jml.2020.104092
Mulder, E., Van de Ven, M., Segers, E., & Verhoeven, L. (2019). Context, word, and student predictors in second language vocabulary learning. *Applied Psycholinguistics, 40*, 137–166.
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.
Olive, T. (2004). Working memory in writing: Empirical evidence from the dual-ask technique. *European Psychologist, 9*(1), 32–42.
Olive, T. (2011). Working memory in writing. In V. W. Berninger (Ed.), *Past, present, and future contributions of cognitive writing research to cognitive psychology* (pp. 485–504). Psychology Press.
Olive, T., Kellogg, R. T., & Piolat, A. (2008). Verbal, visual, and spatial working memory demands during text composition. *Applied Psycholinguistics, 29*, 669–687.
Ortega, L. (2009). Sequences and processes in language learning. In M. H. Long & C. J. Doughty (Eds.), *Handbook of second and foreign language teaching* (pp. 81–105). Wiley-Blackwell.
Ortega, L. (2012). Epilogue: Exploring L2 writing-SLA interfaces. *Journal of Second Language Writing, 21*, 404–415.
Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*(1), 63–71.
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.
Perry, F. L. (2011). *Research in applied linguistics: Becoming a discerning consumer* (3rd ed.). Routledge.
Peters, E. (2019). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). Routledge.
Pichette, F., de Serres, L., & Lafontaine, M. (2012). Sentence reading and sentence writing for second language vocabulary acquisition. *Applied Linguistics, 33*(1), 66–82.
Pigada, M., & Schmitt, N. (2006). Vocabulary acquisition through extensive reading: A case study. *Reading in a Foreign Language, 18*(1), 1–28.
Rice, C. A., & Tokowicz, N. (2020). State of the scholarship: A review of laboratory studies of adult second language vocabulary training. *Studies in Second Language Acquisition, 42*, 439–470.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*, 1–32.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics, 45*, 193–213.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 3–37). John Benjamins. https://doi.org/10.1075/tblt.2

Roca de Larios, J., Nicolás-Conesa, F., & Coyle, Y. (2016). Focus on writers: Processes and strategies. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second language writing* (pp. 267–286). De Gruyter.

Roediger, H. L., III., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition, 21*(4), 589–619.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463.

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. In H. Byrnes & R. M. Manchón (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163–191). John Benjamins.

Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing, 28*, 1–19.

Salkind, N. J. (2011). *Statistics for people who (think they) hate statistics* (4th ed.). Sage Publications.

Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System, 6*(2), 72–78.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77–101). Multilingual Matters.

Skehan, P. (2003). *A cognitive approach to language learning* (2nd ed.). Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510–532.

Skehan, P. (2014). Limited attentional capacity, second language performance and task-based pedagogy. In S. Peter (Ed.), *Processing perspectives on task performance* (pp. 211–260). John Benjamins.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604.

Soderstrom, N. C., Kerr, T. K., & Bjork, R. A. (2016). The critical importance of retrieval and spacing for learning. *Psychological Science, 27*(2), 223–230.

Stevenson, M., Schoonen, R., & de Glopper, K. (2006). Revising in two languages: A multidimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing, 15*, 201–233.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. M. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House Publishers.

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 Henry Widdowson* (pp. 125–144). Oxford University Press.

Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford University Press.

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.

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*, 1–10. https://doi.org/10.1177/2158244017730596

Teng, M. F. (2019). The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading. *The Language Learning Journal, 47*(2), 145–158.

Teng, M. F. (2020). Retention of new words learned incidentally through reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research, 24*(6), 785–812. https://doi.org/10.1177/1362168819829026

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, 1–41. https://doi.org/10.1111/lang.12343

Van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual richness and word learning: Context enhances comprehension but retrieval enhances retention. *Language Learning, 68*(2), 546–585. https://doi.org/10.1111/lang.12285

Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning, 61*(1), 219–258.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130–163. Retrieved from http://nflrc.hawaii.edu/rfl/October2003/waring/waring.pdf

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics, 28*(1), 46–65.

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

Webb, S. (2019). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). Routledge.

Wickens, C. D. (1981). *Processing resources in attention, dual task performance, and workload assessment*. Technical report. Engineering-psychology Research Laboratory. University of Illinois.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(3), 449–455.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*(2), 87–95.

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.

Chapter 9
Conclusions, Practical Implications, Limitations, and Suggestions for Future Research

9.1 Introduction

This chapter presents overall conclusions from the three studies reported in this book. It begins with Study 1 (reported in Chap. 5). The main purpose of this study was to find an assessment tool that may be useful for university placement purposes without much distortion in scores due the presence of cognates in tests. I will briefly remind the reader of the aims and design of this study, summarize the findings, and lay out some implications for pedagogy. Then, I highlight some research limitations in Study 1 and suggest possibilities for further research. The chapter then focuses on Studies 2 and 3, reported in Chaps. 6 and 7, following a similar structure to that of the first study. Studies 2 and 3 compared the academic vocabulary learning following sentence writing (SW), timed essay writing (Timed CW) and Untimed CW (Study 3 only).

9.2 Study 1: Assessing Academic Vocabulary Knowledge for Placement Purposes

9.2.1 A Summary of the Research Design and Findings

Study 1 had two main aims. First, it used two measurements of receptive lexical knowledge to assess whether first- and second-year BA students at the Institute of English Studies possess sufficient knowledge of academic vocabulary. One measurement was Nation and Beglar's (2007) Vocabulary Size Test (VST), assessing general vocabulary knowledge. The second measurement, a Yes/No test, was a tailor-made academic-vocabulary meaning-recall test (AVT; Sect. 5.2.3). The second aim of the study was to verify the practicability and reliability of using the VST for academic placement purposes with English majors. To achieve this, I combined the scores

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7_9

obtained in the VST and the AVT and used statistical analyses to find a threshold in the VST at or above which learners may be considered sufficiently proficient in academic vocabulary. By combining the scores of both tests, I found a manner of assessment that is less sensitive to cognate inflation effects (i.e., when scores are artificially inflated due to cognate guessing, thus making students appear more proficient than they really are; see Sect. 1.6).

The analyses found signs of cognate inflation both in the AVT and the VST since cognates had generally higher scores than noncognates. In the VST, the results showed higher cognate inflation among lower-frequency bands. Furthermore, the results demonstrated that 45.28% of Polish English majors would benefit from such extra practice. Statistical analyses found that a score of 9900 in the VST is a suitable a threshold at or below which first- and second-year Polish learners majoring in English may be considered in need of extra practice of academic words. This threshold is important for two reasons. First, had participants taken only the VST, their scores would have been inflated, especially in the lower-frequency bands (i.e., more advanced vocabulary), thus showing that learners may be perceived as more proficient than they really are. Second, had learners sat only the AVT test, as Yes/No test, examinee variability (i.e., differences in self-confidence and/or linguistic background) would also have distorted the scores. This is because some participants lacked self-confidence, therefore scoring low in the AVT while scoring above the threshold in the VST. These participants were likely unsure of their answers and hence decided not to tick several correct items in the AVT. Other learners, especially multilingual learners, took more risks in the AVT, thus having higher scores here but scoring below the threshold in the VST.

As a result, in line with previous research, I have demonstrated that cognate inflation effects (e.g., Elgort, 2013; Petrescu et al., 2017) and examinee variability (Mochida & Harrington, 2006; Schmitt, 2010) may misrepresent learners' true lexical knowledge. A novel contribution of this study lies in the investigation of the combination of the tailor-made AVT and the VST, which appears to have ameliorated the effects of such score distortions.

9.2.2 Implications for Pedagogy

The VST is freely available online, can be administered in paper or electronic form and is easily scored. In practice, establishing a VST threshold below which students are in need of academic vocabulary instruction facilitates academic placement while increasing its reliability. Once a threshold has been found, the VST may be used with little concern regarding distortions in score. Nevertheless, if teachers or researchers decide not to identify and implement a threshold, I recommend the following.

If possible and practicable, following suggestions by Allen (2018), Elgort (2013), and Laufer and McLean (2016), one may ascertain that the number of cognate items in vocabulary tests is proportional to the number of cognates found in learners' L1. However, this is impractical, if not impossible, when learners originate from different

linguistic backgrounds. In contexts were most or all learners have the same L1, as in the Polish context, tests could be created with the correct proportion of cognates. Still, as explained previously, it may be rather difficult to verify such proportion. If it is not possible to find this proportion, cognates should be kept in the test, as without them "it would be impossible to produce valid vocabulary size estimates" (Elgort, 2013, p. 269). This being the case, one must remember that cognates may be over- or under-represented in the test and that the resulting distortion in scores should be considered when interpreting the results. I would go further and argue that the misuse of exotics (e.g., "puma", "yoga"; see Table 5.8), that is, words that rarely change orthographically between languages sharing the same script, indicates a monolingual bias in test construction that should also be avoided. Exotics were excessively employed in band 11K of the VST, resulting in an exceptionally large cognate inflation effect (see Sect. 5.5 and Fig. 5.1).

Another implication, this time more specific to the Polish context, concerns the number of Polish first- and second-year English majors who scored below the VST threshold. As the cluster analyses used in Study 1 have showed, 48 out of 106 learners (45.28%) scored below the 9900-threshold, averaging a score of 8575 in the VST. This means that almost half of the first- and second-year students at the Institute of English Studies would benefit from extra practice with academic vocabulary. This may be also true for third, fourth, or even fifth year students for at least three reasons. First, the institute does not offer any course geared towards the explicit instruction and practice of academic vocabulary. Second, my (rather limited) teaching experience at the institute and anecdotal evidence from colleagues and students indicate that lecturers of content (i.e., most lecturers) do not focus on language (e.g., academic vocabulary), whereas teachers of language do not emphasize academic words. Third, and perhaps most importantly, research has shown that academic words are not learnt incidentally (e.g., in lectures or during reading assignments) even after three years attending classes at English-medium universities in English-speaking countries (e.g., Knoch et al., 2014, 2015). As a result, if the first point is true and the second point is to be believed, there is a high likelihood that at least one third of students at the Institute of English Studies obtain their bachelors' degree without sufficient knowledge of academic vocabulary. This is just reasonable speculation at this point, so more research is needed. Still, it is an argument that is sensible enough to warrant further investigation.

9.2.3 Limitations and Suggestions for Future Research

One obvious limitation, already mentioned above, concerns the fact that only first- and second-year Polish English majors from one institution were investigated in Study 1. In the future, it would be interesting to expand the sample of participants to comprise learners from further years, and possibly English majors from other institutes and universities in Poland. For one thing, this would provide a better picture regarding the current level of academic vocabulary knowledge of such students in the country.

For another, such cross-sectional studies would shed light on knowledge at different levels, hence providing some information on the progress of English majors. In this case, it is perhaps more useful to have studies measuring knowledge of academic words longitudinally to ensure that current pedagogical practice in higher education facilitates lexical learning. Generally, assessing Polish English majors' knowledge of academic vocabulary, cross-sectionally and longitudinally, should provide a much clearer picture of the current situation and should help higher-education institutions around the country make well-informed choices.

Study 1 assesses only the receptive knowledge of academic vocabulary, which is yet another limitation. These results cannot be interpreted to fully predict proficiency in listening or any productive skill. Still, there is a large body of research showing that receptive vocabulary knowledge correlates highly and positively with the four main skills, including academic reading and writing (Laufer & Ravenhorst-Kalovski, 2010; Milton et al., 2010; Paribakht & Webb, 2016), so the current findings should not be underestimated. In the future, it may be worth investigating learners' productive knowledge of academic vocabulary and comprehension skills to make sure they can cope with the production and interpretation of academic texts. If this is done in several higher-education institutions in Poland, across the five years, and also longitudinally, as suggested above, the results will be rather informative and might affect substantial changes.

There are at least two more limitations to Study 1. The first one is that learners were tested on single-word lexical items only. There is a considerable body of research that indicates that multiword items are an important part of academic discourse (e.g., Byrd & Coxhead, 2010; Hyland, 2008, 2012; Simpson-Vlach & Ellis, 2010) and that learners' use of these formulaic sequences is an effective predictor of lexical proficiency (e.g., Crossley et al., 2015). Consequently, future research should try to measure knowledge of multiword items. Second, I did not control for the varying degrees of formal and semantic overlap of cognates in the analysis. As discussed in Sect. 5.5, some bands in the VST had rather unpredictable scores for cognates and noncognates, which may be explained by this difference in formal and semantic overlap. This is because more similar cognates are usually recognized faster (e.g., Comesaña et al., 2015; De Groot, 2011; Dijkstra et al., 2010; Duyck et al., 2007; Mulder et al., 2015) and possibly acquired better than less similar cognates (Otwinowska & Szewczyk, 2019; Otwinowska et al., 2020). If the semantic and formal similarities are controlled for, these variables may be entered in the statistical model as covariates to reduce error, thus increasing power and generating more reliable results.

9.3 Studies 2 and 3: Incidental Lexical Learning Through SW and CW

Given that almost 50% of the first- and second-year learners failed to demonstrate adequate receptive lexical knowledge of academic vocabulary, it is paramount to explore the effectiveness of different tasks in the teaching of academic words since, to my knowledge, evidence in this area is still scarce. To this aim, the second and third studies reported in this book (Chaps. 6 and 7, respectively) compared the academic vocabulary learning potential of sentence writing (SW) and composition writing (CW) tasks. Writing argumentative essays is a rather quotidian task at university and was therefore the type of essay explored here.

9.3.1 A Summary of the Research Design and Findings

Studies 2 and 3 assessed the potential of L2 writing (SW and CW tasks) in the Polish academic context to yield incidental learning of academic words provided in a glossary. Study 2 compared SW to 60-min Timed CW tasks whereas Study 3 adopted SW, Timed CW, and Untimed CW tasks. In both studies, all participants wrote a control essay (i.e., without keywords) prior to the quasi-experiment proper and answered a questionnaire at the end of the quasi-experiment to ensure learning was truly incidental. To measure increase in cognitive load, I compared the holistic scores, the number of errors, and the number of words composed per minute in the control essay and in the timed and untimed essays in both studies. Study 3 also employed a self-rating scale assessing participants' perceived level of task difficulty, effort, and frustration to assess increase in cognitive load.

The results are inconclusive, as discussed in Chap. 8 (see also Sect. 9.3.3 below). In Study 2, SW generated significantly more learning than Timed CW in all three variables tested. In Study 3, SW, Timed CW, and Untimed CW yielded similar lexical learning. The results showed signs that Untimed CW may have generated more learning than the other groups, but this learning reached statistical significance only in one of the variables, and only barely. In terms of cognitive load, the results in both studies confirmed my predictions that Timed CW is more cognitively demanding than writing control essays (without keywords). This is because timed essays took longer to write and were less accurate than control essays. Also as predicted, Untimed CW put less pressure on L2 writers' cognitive resources than the control and timed essays.

In these studies, I aimed to verify experimentally whether writing sentences or argumentative essays with the novel keywords would lead learners to acquire similar amounts of vocabulary, thus supporting the involvement load hypothesis (ILH; Laufer & Hulstijn, 2001). Study 2, which found more lexical learning following SW than Timed CW, did not lend support to the ILH and to previous studies (i.e., Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017). Study

3, wherein SW, Timed CW, and Untimed CW generated similar lexical learning, corroborated the predictions of the ILH and the results of previous research. The exception is Zou's (2017) findings, which showed more gains for CW than for SW.

These studies add to the existing body of evidence regarding vocabulary learning in several ways. One is the use of academic words, so far unexplored in quasi-experimental studies of this type. Another novel contribution concerned the use of textual measures to control for participants' writing proficiency and, together with the self-rating scale, to investigate increases in cognitive load due to the incorporation of keywords. Also, through a qualitative analysis of the keywords used in the essays (see Sect. 8.6), these studies are the first L2 writing studies to draw attention to the importance of well-designed and well-piloted glossaries, otherwise learners may be unable to use the keywords accurately. Moreover, the use of a post-task questionnaire provided considerably more control over extra-treatment exposure, which has been overlooked in studies of similar type. Using the questionnaire ensured that any learning reported here was truly incidental, which reinforces my findings. Finally, Studies 2 and 3 utilized generalized linear mixed models (GLMMs) during the statistical analyses. This advanced statistical technique is more powerful than more traditional statistical methods, and my studies seem to be the first to employ GLMMs when assessing lexical learning via essay writing.

9.3.2 Implications for Pedagogy

Several implications may be drawn from these findings. The most obvious is that writing sentences and essays generate considerable lexical learning and retention when learners are obliged to incorporate novel words with the help of external sources. As a result, these are tasks that may be effectively utilized at university to promote the learning of academic words. However, if learners are under heavy cognitive load, as is the case with Timed CW, sentence writing may be more or equally conducive to lexical learning than essay writing. This being the case, it seems unreasonable to assign essay-writing tasks solely to further vocabulary acquisition, at least in the classroom, where they must be timed. This is because essay writing is far more time-consuming and equally effective as or less effective than writing sentences. Based on these considerations, I would suggest limiting in-classroom essay writing to the practice of writing skills or for assessment purposes, not for lexical learning. If vocabulary acquisition is also a goal, there must be no time limit assigned to the writing, and it is therefore preferably done outside the classroom. In the classroom, SW may be more efficient.

Another practical implication of Studies 2 and 3 concerns the use of glossaries, or similarly, of dictionaries. In Sect. 8.6, I demonstrated that some of the academic keywords were incorporated inaccurately despite the provision of glossaries with definitions and examples. Also, research has shown that learners often fail to consult dictionaries properly because they misunderstand the meaning or use of the words or focus on the wrong definition of polysemous lexical items (Nesi & Haill, 2002).

This problem may be exacerbated with academic words. This is because these words are typically abstract (Corson, 1997) and in some cases may be polysemous and have meanings that differ in general and academic use. For instance, the AWL item "significant" may mean "important or noticeable", as in "there has been a significant increase in the number of women students in recent years" (Cambridge University Press, 2020), or may be connected to statistical analysis (basically meaning $p < 0.05$). Both meanings are used in academic writing, but only the former is utilized in general English. Considering the above, it is advisable that university teachers underscore the importance of using high-quality monolingual dictionaries, and that they practice dictionary use in the classroom. Well-designed glossaries may also be provided (and their use practiced), but this seems unnecessary not least because learners have access to monolingual dictionaries in their mobile devices. University teachers, especially language teachers and teachers of writing skills, should also, when necessary, draw attention to words that may confuse learners.

9.3.3 Limitations and Suggestions for Future Research

Studies 2 and 3 have at least two limitations that are similar to those in Study 1. First, only single-word lexical items were employed, but multiword items pervade academic discourse (e.g., Simpson-Vlach & Ellis, 2010). Future research should address this issue. Also, all participants investigated were first-year students. It may be that learners from other years may be more accustomed to essay writing, which may affect the lexical learning yielded by this task. Consequently, it is advisable that future studies recruit participants from different academic years, both BA and MA students.

Some limitations already mentioned elsewhere are the following. In Study 2, the task instructions failed to persuade learners to write unstructured essays, that is, to allocate more attentional resources to keyword use in lieu of essay writing (the primary task). Doing so could have increased the lexical learning yielded by the Timed CW task, and thus simulate what I believe happened in Zou's (2017) study, where CW generated more learning than SW (see Sect. 6.5). Future research could explore different ways to address this issue. Furthermore, in Study 3, items 4 and 5 in the self-rating scale (see Sect. 7.5 and Appendix L) did not reliably address task-induced cognitive load. The problem likely stemmed from the task script, where the words "annoyed" and "frustrated" may have been misconstrued by participants as assessing satisfaction, not cognitive load. In the future, one solution is that the task script eschews using these words, opting for "stressed" instead.

Another potential problem with Study 3 is that it only used the data from the second condition in Study 2 (i.e., the second day learners wrote sentences or essays). If the SW task frustrated learners in Study 2, as it did in Study 3, in the sense that they did not like the task, then participants were even more frustrated on the second day, which may have been detrimental to lexical learning. Using only these data in Study 3 possibly reduced the average learning induced by SW, thus equating its learning to

the learning generated by Timed CW. In the future, researchers requiring learners to write sentences with keywords should not ask participants to perform the task twice or more. If there are too many keywords, and therefore the SW task must be split into different sessions, researchers should explore different ways to motivate learners.

There are at least two other possible limitations in Studies 2 and 3. One is that all participants were adults, B2 or higher, and Polish or, in a few cases, speakers of other Slavic languages. It is possible that the findings would have been different with adolescents or with learners of a different proficiency level and/or L1. Future research may thereby incorporate learners of different proficiency level and, if desired, varied linguistic backgrounds. The second possible limitation concerns the use of glossaries. Some researchers may frown upon the use of glossaries in incidental learning tasks. There may two reasons for this. First, the provision of glosses increases processing of the keywords, therefore increasing lexical learning (e.g., Hulstijn et al., 1996; Nation, 2013; Rott, 2005; Schmitt, 2008, 2010; Teng, 2020; Watanabe, 1997). Still, as explained in Sect. 3.3.2, the glossaries in Studies 2 and 3 simulated dictionary use, which is unavoidable if writers are to accurately incorporate novel lexical items in their essays. This explains why all studies comparing SW and CW tasks (i.e., Gohar et al., 2018; Kim, 2008; Tahmasbi & Farvardin, 2017; Zou, 2017; see Table 2.2) utilized glossaries in their research design. Moreover, glossaries were provided in all tasks (SW, Timed CW, and Untimed CW), and hence any increase in lexical learning likely affected all tasks equally. Second, it could be argued that the provision of glossaries drew attention to the keywords, making lexical learning more explicit, and therefore, not incidental. Nonetheless, as discussed a few times in this book, including in the previous chapter (Sect. 8.5), more explicit learning is not the same as intentional learning.

Future Research Paths comparing Timed CW and Untimed CW . In Sect. 8.4.2, I discussed two reasons why, in Study 3, Untimed CW and Timed CW registered statistically similar gains in lexical knowledge, which was unexpected. One reason is that Study 3 may have lacked statistical power, which is a limitation of the study. Consequently, I suggested that it is possible that Untimed CW yielded more learning than Timed CW, but the analyses did not show this. Further research with more statistical power may shed light on this issue. Another reason for the similar amount of learning following Timed CW and Untimed CW is the following. Despite the enhanced exposure and likely increase in output production in Untimed CW relative to Timed CW (because learners likely had the time to rewrite sentences with keywords more in Untimed CW than in Timed CW), these may not have promoted deeper processing of the keywords. Therefore, Untimed CW did not generate higher levels of vocabulary learning than Timed CW. To pursue this hypothesis, possible future research designs are outlined below.

It would be interesting if a future study investigated the amount of revision and rewriting in the Untimed CW condition. One technique that could be employed here is the keystroke logging software, which records the whole writing process. This has been used extensively in writing research (e.g., Chukharev-Hudilainen et al., 2019; Latif, 2008; Leijten & Van Waes, 2013; Sabbaghan, 2013; Van Waes et al., 2009).

The preferred software for such task appears to be Inputlog (Leijten & Van Waes, 2020). Inputlog synchronizes with text processors such as Microsoft Word to record the whole writing process, which can then be re-played, paused, rewound, and fast forwarded; it records each keystroke and each pause in milliseconds, allowing the researcher to see when writers are thinking, or perhaps reading the text. Inputlog even records the position of the cursor, so researchers know, for example, when writers go back in the text and what they rewrite. This could be used with the Untimed CW tasks, for instance, to check whether the sentences with keywords were rewritten or not, and whether the rewritten sentences generated more lexical learning than those that were not rewritten. Keystroke logging could even be combined with stimulated recall techniques (see Sabbaghan, 2013 for an example of such research design). In this case, the researcher may re-play the writing process while interviewing the writer and pause the process (or rewind/fast forward) in order to ask questions related to specific events. For example, the researcher may note a long pause in the middle of the writing process, may show this pause to the participant, and may inquire about the reason for the pause.

Another valuable technique that could be used in isolation or combined with keystroke logging is eye-tracking technology. Essentially, eye-trackers track eye movements over a visual scene, recording quick eye pauses (i.e., fixations) and faster eye movements (i.e., saccades) in milliseconds. The assumption is that the longer it takes processing a certain region of interest (ROI), that is, looking at this ROI, the longer the cognitive engagement with the material is, and the higher the probability of learning (Conklin & Pellicer-Sánchez, 2016). Researchers have used eye-tracking to measure lexical processing mostly in reading (e.g., Altmann & Kamide, 2007; Chaffin et al., 2001; Elgort et al., 2017; Godfroid & Schmidtke, 2013; Godfroid et al., 2018; Lai et al., 2013; Libben & Titone, 2009; Williams & Morris, 2004), but also in writing (Alamargot et al., 2010; Torrance et al., 2016); eye-tracking has also been used, for instance, in research with video subtitles (e.g., Szarkowska & Gerber-Morón, 2018). Anson and Schwegler (2012), Conklin and Pellicer-Sánchez (2016), and Roberts and Siyanova-Chanturia (2013) have all provided valuable introductory information on the research possibilities afforded by this technology. For example, eye-trackers may record the sum duration of all fixations in a given ROI, such as a keyword (this is called Total Reading Time). Also, the researcher may be interested in how long participants spend re-reading a certain ROI (say, a sentence containing a keyword). This is called Second Pass Reading Time (see Conklin & Pellicer-Sánchez, 2016, p. 456 for an interesting illustration of several types of measurements). It would even be possible to synchronize Inputlog and eye-tracking software, thus allowing the researcher to have a much deeper insight into L2 writers' output production (via Inputlog) and exposure to input (via eye-tracking).

9.4 Final Conclusions

This book has empirically demonstrated a few valid points. It has showed that Polish students majoring in English lack knowledge of academic vocabulary, which may significantly hinder their ability to comprehend and produce academic texts. Then, this book discussed convincing evidence that requiring learners to incorporate keywords in their writing, inside or outside the class, facilitates lexical learning. Therefore, it showed that in the academic context, quotidian tasks such as writing argumentative essays or sentence with keywords may help address the insufficiency in knowledge of academic words. I hope that the findings reported here will be useful for researchers and practitioners alike. They have been for me. This project has helped me as a researcher to understand the nature of academic vocabulary learning and will help me as a teacher working in the academic environment.

References

Alamargot, D., Plane, S., Lambert, E., & Chesnet, D. (2010). Using eye and pen movements to trace the development of writing expertise: Case studies of a 7th, 9th and 12th grader, graduate student, and professional writer. *Reading and Writing, 23*, 853–888.

Allen, D. (2018). Cognate frequency and assessment of second language lexical knowledge. *International Journal of Bilingualism*, 1–16. https://doi.org/10.1177/1367006918781063

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language. *Journal of Memory and Language, 57*, 502–518. https://doi.org/10.1016/j.jml.2006.12.004

Anson, C. M., & Schwegler, R. A. (2012). Tracking the mind's eye: A new technology for researching twenty-first-century writing and reading processes. *College Composition and Communication, 64*(1), 151–171.

Byrd, P., & Coxhead, A. (2010). *On the other hand: Lexical bundles in academic writing and in the teaching of EAP* (Vol. 5, pp. 31–64). University of Sydney Papers in TESOL.

Cambridge University Press. (2020). *Cambridge dictionary* [online]. Retrieved from: https://dictionary.cambridge.org/. Accessed November 06, 2020

Chaffin, R., Morris, R. K., & Seely, R. E. (2001). Learning new word meanings from context: A study of eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 225–235.

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H.-H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition, 41*, 583–604. https://doi.org/10.1017/S027226311900007X

Comesaña, M., Soares, A. P., Ferré, P., Romero, J., Guasch, M., & García-Chico, T. (2015). Facilitative effect of cognate words vanishes when reducing the orthographic overlap: The role of stimuli list composition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 614–635.

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32*(3), 453–467.

Corson, D. (1997). The learning and use of academic English words. *Language Learning, 47*(4), 671–718.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics, 36*(5), 570–590.

De Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction.* Psychology Press.

Dijkstra, T., Miwa, K., Brummelhuis, B., Sappelli, M., & Baayen, H. (2010). How cross-linguistic similarity and task demands affect cognate recognition. *Journal of Memory and Language, 62*, 284–301.

Duyck, W., Van Assche, E., Drieghe, D., & Hartsuiker, R. J. (2007). Visual word recognition by bilinguals in a sentence context: Evidence for nonselective lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 663–679.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the vocabulary size test. *Language Testing, 30*(2), 253–272.

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2017). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition, 40*, 341–366.

Godfroid, A., Ahn, J., Choi, I., Ballard, L., Cui, Y., Johnston, S., Lee, S., Sarkar, A., & Yoon, H.-J. (2018). Incidental vocabulary learning in a natural reading context: An eye-tracking study. *Bilingualism: Language and Cognition, 21*(3), 563–584.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports, and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183–205). University of Hawai'i, National Foreign Language Resource Center.

Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique feature analysis or involvement load hypothesis: Estimating their predictive power in vocabulary learning. *Journal of Psycholinguistic Research, 47*, 859–869.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80*(3), 327–339.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150–169.

Kim, Y. (2008). The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning, 58*(2), 285–325.

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university. *Journal of Second Language Writing, 28*, 39–52.

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing, 21*, 1–17.

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S.W.-Y., Lee, M.-H., Chiou, G.-L., Liang, J.-C., & Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review, 10*, 90–115.

Latif, M. A. (2008). A state-of-the-art review of the real-time computer-aided study of the writing process. *International Journal of English Studies, 8*(1), 29–50.

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

Laufer, B., & Mclean, S. (2016). Loanwords and vocabulary size test scores: A case of different estimates for different L1 learners. *Language Assessment Quarterly, 13*(3), 202–217.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30. Retrieved in March 2018 from http://files.eric.ed.gov/fulltext/EJ887873.pdf

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392.

Leijten, M., & Van Waes, L. (2020). *Inputlog*. Retrieved from: https://www.inputlog.net/. Accessed November 27, 2020.

Libben, M. R., & Titone, D. A. (2009). Bilingual lexical access in context: Evidence from eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(2), 381–390. https://doi.org/10.1037/a0014875

Milton, J., Wade, J., & Hopkins, N. (2010). Aural word recognition and oral competence in a foreign language. In R. Chacón-Beltrán, C. Abello-Contesse, & M. D. M. Torreblanca-López (Eds.), *Insights into non-native vocabulary teaching and learning* (pp. 83–97). Multilingual Matters.

Mochida, A., & Harrington, M. (2006). The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing, 23*(1), 73–98.

Mulder, K., Dijkstra, T., & Baayen, R. H. (2015). Cross-language activation of morphological relatives in cognates: The role of orthographic overlap and task-related processing. *Frontiers in Human Neuroscience, 9*(16), 1–18.

Nation, I. S. P. (2013) *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nesi, H., & Haill, R. (2002). A study of dictionary use by international students at a British university. *International Journal of Lexicography, 15*(4), 277–305.

Otwinowska, A., Foryś-Nogala, M., Kobosko, W., & Szewczyk, J. (2020). Learning orthographic cognates and non-cognates in the classroom: Does awareness of cross linguistic similarity matter? *Language Learning*, 1–47. https://doi.org/10.1111/lang.12390

Otwinowska, A., & Szewczyk, J. M. (2019). The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism, 22*(8), 974–991.

Paribakht, T. S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes, 21*, 121–132.

Petrescu, M. C., Helms-Park, R., & Dronjic, V. (2017). The impact of frequency and register on cognate facilitation: Comparing Romanian and Vietnamese speakers on the vocabulary levels test. *English for Specific Purposes, 47*, 15–25.

Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition, 35*, 213–235. https://doi.org/10.1017/S0272263112000861

Rott, S. (2005). Processing glosses: A qualitative exploration of how form-meaning connections are established and strengthened. *Reading in a Foreign Language, 17*(2), 95–124.

Sabbaghan, S. (2013). How noticing is affected by replay of writing process during stimulated recall. *Procedia: Social and Behavioural Sciences, 83*, 629–633.

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

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

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

Szarkowska, A., & Gerber-Morón, O. (2018). Viewers can keep up with fast subtitles: Evidence from eye movements. *PLoS ONE, 13*(6), 1–30. https://doi.org/10.1371/journal.pone.0199331

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*, 1–10. https://doi.org/10.1177/2158244017730596

Teng, M. F. (2020). Retention of new words learned incidentally through reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research, 24*(6), 785–812. https://doi.org/10.1177/1362168819829026

Torrance, M., Johansson, R., Johansson, V., & Wengelin, A. (2016). Reading during the composition of multi-sentence texts: An eye-movement study. *Psychological Research, 80*, 729–743. https://doi.org/10.1007/s00426-015-0683-8

Van Waes, L., Leijten, M., & Van Weijen, D. (2009). Keystroke logging in writing research: Observing writing processes with Inputlog. *German as a Foreign Language Journal, 2*(3), 41–64.

Watanabe, Y. (1997). Input, intake and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in Second Language Acquisition, 19*, 287–307.
Williams, R. S., & Morris, R. K. (2004). Eye movements, word familiarity, and vocabulary acquisition. *European Journal of Cognitive Psychology, 16*(2), 312–339.
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.

Appendix A

A.1 389 Polish–English Cognate Types in the AWL

English	Polish	English	Polish
conformist	konformistyczny	initiation	inicjacja
cooperate	kooperować	initiative	inicjatywa
cooperation	kooperacja	innovation	innowacja
coordinate	koordynować	innovative	innowacyjny
coordinated	skoordynowany	innovator	innowator
coordination	koordynacja	inspector	inspektor
coordinator	koordynator	instability	niestabilność
specific	specyficzny	institute	instytut
traditional	tradycyjny,	institution	instytucja
abstract	abstrakcyjny	instruct	instruować
academic	akademicki	instruction	instrukcja
academy	akademia	instructor	instructor
accumulate	kumulować	integrate	integrować
accumulation	skumulowanie	integration	integracja
adapt	adaptować	intelligence	inteligencja
adaptation	adaptacja	intelligent	inteligentny
adequacy	adekwatność	intensify	intensyfikować
adequate	adekwatny	intensity	intensywność
administration	administracja	intensive	intensywny
administrative	administracyjny	interaction	interakcja
alternative	alternatywny	interactive	interaktywny
analogous	analogiczny	interpret	interpretować

(continued)

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

(continued)

English	Polish	English	Polish
analogy	analogia	interpretation	interpretacja
analysis	analiza	irrational	irracjonalny
analyst	analityk	isolate	izolować
analytic	analityczny	isolated	odizolowany
analyze	analizować	isolation	izolacja
arbitrary	arbitralny	isolationism	izolacjonizm
aspect	apekt	legal	legalny
assist	asystować	liberal	liberalny
assistant	asystent	liberalism	liberalizm
author	autor	liberalize	liberalizować
authority	autorytet	licence	licencja
automatic	automatyczny	locate	lokować
category	kategoria	location	lokacja
chemical	chemiczny	logic	logika
civil	cywilny	logical	logiczny
classic	klasyczny	major	major
classical	klasyczny	manipulate	manipulować
classics	klasyka	manipulation	manipulacja
code	kodować	maximize	maksymalizować
coded	zakodowany	maximum	maksymalny
colleague	kolega	mechanism	mechanizm
comment	komentować	media	media
commentary	komentarz	mediation	mediacja
commentator	komentator	medical	medyczny
commission	komisja	method	metoda
communicate	komunikować	methodical	metodyczny
communication	komunikacja	methodological	metodologiczny
communicative	komunikatywny	migrate	migrować
compatibility	kompatybilność	migration	migracja
compatible	kompatybilny	military	militarny
compilation	kompilacja	minimal	minimalny
compile	kompilować	minimize	minimalizować
complex	kompleks	minimum	minimum
component	komponent	ministerial	ministerialny
computer	komputer	ministry	ministerstwo
concentrate	koncentrować	modification	modyfikacja
concentrated	skoncentrowany	modified	modyfikowany

(continued)

(continued)

English	Polish	English	Polish
concentration	koncentracja	modify	modyfikować
conference	konferencja	monitor	monitorować
conflict	konflikt	monitoring	monitoring
consequence	konsekwencja	motivate	motywować
consistency	konsystencja	motivated	zmotywowany
constitution	konstytucja	motivating	motywujący
constitutional	konstytucyjny	motivation	motywacja
construct	konstruować	motive	motyw
construction	konstrukcja	negative	negatywny
constructive	konstruktywny	neutral	neutralny
consult	konsultować	neutrality	neutralność
consultant	konsultant	neutralize	neutralizować
consultation	konsultacja	nonconformist	nonkonformistyczny
consulting	konsulting	nonconformity	nonkonformizm
consume	konsumować	normal	normalny
consumer	konsument	normality	normalność
consumption	konsumpcja	nuclear	nuklearny
contact	kontaktować	objective	obiektywny
context	kontekst	orientation	orientacja
contract	kontrakt	participate	partycypować
contrast	kontrastować	partner	partner
contrasting	kontrastowy	partnership	partnerstwo
convention	konwencja	passive	pasywny
conventional	konwencjonalny	percent	procent
converse	konwersować	perception	percepcja
conversion	konwersja	periodic	periodyczny
corporation	korporacja	phenomenal	fenomenalny
correspond	koreserpondować	philosopher	filozof
correspondence	korespondencja	philosophical	filozoficzny
create	kreować	philosophize	filozofować
creation	kreacja	philosophy	filozofia
creative	kreatywny	physical	fizyczny
creativity	kreatywność	plus	plus
creator	kreator	positive	pozytywny
credit	kredyt	potential	potencjalny
criterion	kryterium	precise	precyzyjny
cultural	kulturalny	precision	precyzja

(continued)

(continued)

English	Polish	English	Polish
culture	kultura	procedure	procedura
cycle	cykl	process	proces
cyclical	cykliczny	professional	profesjonalny
debate	debatować	professionalism	profesjonalizm
decade	dekada	project	projekt
deduce	dedukować	promote	promować
deduction	dedukcja	promotion	promocja
define	definiować	proportion	proporcja
definite	definitywny	proportional	proporcjonalny
definition	definicja	psychological	psychologiczny
demonstrate	demonstrować	psychologist	psycholog
demonstration	demonstracja	psychology	psychologia
demonstrator	demonstrant	publication	publikacja
depressing	depresyjny	publish	publikować
depression	depresja	radical	radykalny
designer	dizajner	rational	racjonalny
detective	detektyw	rationality	racjonalność
detector	detektor	rationalization	racjonalizacja
deviation	dewiacja	rationalize	racjonalizować
discretion	dyskrecja	react	reagować
discriminate	dyskryminiować	reaction	reakcja
discriminated	dyskryminowany	reactionary	reakcyjny
discrimination	dyskryminacja	reactive	reakcyjny
document	dokument	reactor	reactor
domain	domena	reconstruct	rekonstruować
dominance	dominacja	reconstruction	rekonstrukcja
dominant	dominujący	regime	reżim
dominate	dominować	region	region
domination	dominacja	register	rejestrować
drama	dramat	registered	zarejestrowany
dramatic	dramatyczny	registration	rejestracja
dramatist	dramaturg	regulate	regulować
dramatize	dramatyzoać	relax	relaksować
dynamic	dynamiczny	relaxation	relaks
economical	ekonomiczny	relaxing	relaksujący
economics	ekonomia	reside	rezydować
edit	edytować	residence	rezydencja

(continued)

(continued)

edition	edycja	resident	rezydent
element	element	restriction	restrykcja
eliminate	eliminować	restrictive	restrykcyjny
elimination	eliminacja	revelation	rewelacja
empirical	empiryczny	revolution	rewolucja
energetic	energiczny	revolutionary	rewolucyjny
energy	energia	revolutionize	rewolucjonizować
erode	erodować	role	rola
erosion	erozja	scenario	scenariusz
ethical	etyczny	section	sekcja
ethics	etyka	selective	selektywny
evolution	ewolucja	series	serial
evolutionary	ewolucujny	sex	seks
exclusive	ekskluzywny	sexism	seksizm
expansion	ekspansja	sexual	seksualny
expert	ekspert	sexuality	seksualność
exploit	eksploatować	simulate	symulować
export	eksportować	simulation	symulacja
exporter	eksporter	specification	specyfikacja
federal	federalny	specificity	specyficzność
federation	federacja	sphere	Sfera
final	finalny	spherical	sferyczny
finalize	finalizować	statistical	statystyczny
finance	finansować	statistics	ststystyka
financial	finansowy	status	Status
fluctuate	fluktuować	strategic	strategiczny
fluctuation	fluktuacja	strategy	Strategia
formula	formuła	stress	Stress
formulate	formułować	stressful	stresujący
foundation	fundacja	style	Styl
function	funkcjonować	succession	Sukcesja
functional	funkcjonalny	supplement	supplement
fundamental	fundamentalny	survival	Surwiwal
generate	generować	symbol	Symbol
generation	generacja	symbolic	symboliczny
global	globalny	symbolism	symbolizm
globe	glob	symbolize	symbolizować
goal	gol	technical	techniczny

(continued)

(continued)

guarantee	gwarancja	technique	Technika
hierarchical	hierarchiczny	technological	technologiczny
hierarchy	hierarchia	technology	technologia
hypothesis	hipoteza	terminal	Terminal
hypothetical	hipotetyczny	text	Tekst
identical	identyczny	theoretical	teoretyczny
identification	identyfikacja	theorist	Teoretyk
identify	identyfikować	theory	Teoria
ideological	ideologiczny	tradition	Tradycja
ideology	ideologia	traditionalist	tradycjonalista
ignorance	ignorancja	transfer	transferować
ignorant	ignorancki	transform	transformować
ignore	ignorować	transformation	transformacja
illustrate	ilustrować	transit	Tranzyt
illustration	ilustracja	transport	transportować
image	image	trend	Trend
immigrant	imigrant	uniform	Uniform
immigrate	imigrować	variation	Wariacja
immigration	imigracja	version	Wersja
individual	inywidualny	vision	Wizja
individuality	indywidualność	visual	Wizualny
infrastructure	infrastruktura	visualization	wizualizacja
initial	inicjał	visualize	wizualizować
initiate	inicjować		

Appendix B

B.1 The VST (Study 1; Chap. 5)

Source: https://my.vocabularysize.com//

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Vocabulary Size Test[1]

Circle the letter a-d with the closest meaning to the key word in the question.

1. SEE: They **saw** it.
   a. cut
   b. waited for
x c. looked at
   d. started

2. TIME: They have a lot of **time**.
   a. money
   b. food
x c. hours
   d. friends

3. PERIOD: It was a difficult **period**.
   a. question
x b. time
   c. thing to do
   d. book

4. FIGURE: Is this the right **figure**?
   a. answer
   b. place
   c. time
x d. number

5. POOR: We are **poor**.
x a. have no money
   b. feel happy
   c. are very interested
   d. do not like to work hard

6. DRIVE: He **drives** fast.
   a. swims
   b. learns
   c. throws balls
x d. uses a car

7. JUMP: She tried to **jump**.
   a. lie on top of the water
x b. get off the ground suddenly
   c. stop the car at the edge of the road
   d. move very fast

8. SHOE: Where is your **shoe**?
   a. the person who looks after you
   b. the thing you keep your money in
   c. the thing you use for writing
x d. the thing you wear on your foot

9. STANDARD: Her **standards** are very high.
   a. the bits at the back under her shoes
   b. the marks she gets in school
   c. the money she asks for
x d. the levels she reaches in everything

10. BASIS: This was used as the **basis**.
    a. answer
    b. place to take a rest
    c. next step
x   d. main part

Second 1000

1. MAINTAIN: Can they **maintain** it?
x a. keep it as it is
   b. make it larger
   c. get a better one than it
   d. get it

2. STONE: He sat on a **stone**.
x a. hard thing
   b. kind of chair
   c. soft thing on the floor
   d. part of a tree

3. UPSET: I am **upset**.
   a. tired
   b. famous
   c. rich
x d. unhappy

4. DRAWER: The **drawer** was empty.
x a. sliding box
   b. place where cars are kept
   c. cupboard to keep things cold
   d. animal house

5. PATIENCE: He has no **patience**.
x a. will not wait happily
   b. has no free time
   c. has no faith
   d. does not know what is fair

6. NIL: His mark for that question was **nil**.
   a. very bad
x b. nothing
   c. very good
   d. in the middle

7. PUB: They went to the **pub**.
x a. place where people drink and talk
   b. place that looks after money
   c. large building with many shops
   d. building for swimming

8. CIRCLE: Make a **circle**.
   a. rough picture
   b. space with nothing in it
x c. round shape
   d. large hole

9. MICROPONE: Please use the **microphone**.
   a. machine for making food hot
x b. machine that makes sounds louder
   c. machine that makes things look bigger
   d. small telephone that can be carried around

10. PRO: He's a **pro**.
    a. someone who is employed to find out important secrets
    b. a stupid person
    c. someone who writes for a newspaper
x   d. someone who is paid for playing sport etc

Third 1000

1. SOLDIER: He is a **soldier**.
   a. person in a business
   b. student
   c. person who uses metal
x d. person in the army

2. RESTORE: It has been **restored**.
   a. said again
   b. given to a different person
   c. given a lower price
x d. made like new again

3. JUG: He was holding a **jug**.
x a. A container for pouring liquids
   b. an informal discussion
   c. A soft cap
   d. A weapon that explodes

4. SCRUB: He is **scrubbing** it.
   a. cutting shallow lines into it
   b. repairing it
x c. rubbing it hard to clean it
   d. drawing simple pictures of it

5. DINOSAUR: The children were pretending to be **dinosaurs**.
   a. robbers who work at sea
   b. very small creatures with human form but with wings
   c. large creatures with wings that breathe fire
x d. animals that lived a long time ago

6. STRAP: He broke the **strap**.
   a. promise
   b. top cover
   c. shallow dish for food
x d. strip of material for holding things together

7. PAVE: It was **paved**.
   a. prevented from going through
   b. divided
   c. given gold edges
x d. covered with a hard surface

8. DASH: They **dashed** over it.
x a. moved quickly
   b. moved slowly
   c. fought
   d. looked quickly

9. ROVE: He couldn't stop **roving**.
   a. getting drunk
x b. travelling around
   c. making a musical sound through closed lips
   d. working hard

10. LONESOME: He felt **lonesome**.
    a. ungrateful
    b. very tired
x c. lonely
    d. full of energy

Fourth 1000

1. COMPOUND: They made a new **compound**.
   a. agreement
x b. thing made of two or more parts
   c. group of people forming a business
   d. guess based on past experience

2. LATTER: I agree with the **latter**.
   a. man from the church
   b. reason given
x c. last one
   d. answer

3. CANDID: Please be **candid**.
   a. be careful
   b. show sympathy
   c. show fairness to both sides
x d. say what you really think

4. TUMMY: Look at my **tummy**.
   a. cloth to cover the head
x b. stomach
   c. small furry animal
   d. thumb

5. QUIZ: We made a **quiz**.
   a. thing to hold arrows
   b. serious mistake
x c. set of questions
   d. box for birds to make nests in

6. INPUT: We need more **input**.
x a. information, power, etc. put into something
   b. workers
   c. artificial filling for a hole in wood
   d. money

7. CRAB: Do you like **crabs**?
x a. sea creatures that walk sideways
   b. very thin small cakes
   c. tight, hard collars
   d. large black insects that sing at night

8. VOCABULARY: You will need more **vocabulary**.
x a. words
   b. skill
   c. money
   d. guns

9. REMEDY: We found a good **remedy**.
x a. way to fix a problem
   b. place to eat in public
   c. way to prepare food
   d. rule about numbers

10. ALLEGE: They **alleged** it.
x a. claimed it without proof
   b. stole the ideas for it from someone else
   c. provided facts to prove it
   d. argued against the facts that supported it

Fifth 1000

1. DEFICIT: The company had a large **deficit**.
   x a. spent a lot more money than it earned
   b. went down a lot in value
   c. had a plan for its spending that used a lot of money
   d. had a lot of money in the bank

2. WEEP: He **wept**.
   a. finished his course
   x b. cried
   c. died
   d. worried

3. NUN: We saw a **nun**.
   a. long thin creature that lives in the earth
   b. terrible accident
   x c. woman following a strict religious life
   d. unexplained bright light in the sky

4. HAUNT: The house is **haunted**.
   a. full of ornaments
   b. rented
   c. empty
   x d. full of ghosts

5. COMPOST: We need some **compost**.
   a. strong support
   b. help to feel better
   c. hard stuff made of stones and sand stuck together
   x d. rotted plant material

6. CUBE: I need one more **cube**.
   a. sharp thing used for joining things
   x b. solid square block
   c. tall cup with no saucer
   d. piece of stiff paper folded in half

7. MINIATURE: It is a **miniature**.
   x a. a very small thing of its kind
   b. an instrument to look at small objects
   c. a very small living creature
   d. a small line to join letters in handwriting

8. PEEL: Shall I **peel** it?
   a. let it sit in water for a long time
   x b. take the skin off it
   c. make it white
   d. cut it into thin pieces

9. FRACTURE: They found a **fracture**.
   x a. break
   b. small piece
   c. short coat
   d. rare jewel

10. BACTERIUM: They didn't find a single **bacterium**.
    x a. small living thing causing disease
    b. plant with red or orange flowers
    c. animal that carries water on its back
    d. thing that has been stolen and sold

Sixth 1000

1. DEVIOUS: Your plans are **devious**.
   x a. tricky
   b. well-developed
   c. not well thought out
   d. more expensive than necessary

2. PREMIER: The **premier** spoke for an hour.
   a. person who works in a law court
   b. university teacher
   c. adventurer
   x d. head of the government

3. BUTLER: They have a **butler**.
   x a. man servant
   b. machine for cutting up trees
   c. private teacher
   d. cool dark room under the house

4. ACCESSORY: They gave us some **accessories**.
   a. papers allowing us to enter a country
   b. official orders
   c. ideas to choose between
   x d. extra pieces

5. THRESHOLD: They raised the **threshold**.
   a. flag
   x b. point or line where something changes
   c. roof inside a building
   d. cost of borrowing money

6. THESIS: She has completed her **thesis**.
   x a. long written report of study carried out for a university degree
   b. talk given by a judge at the end of a trial
   c. first year of employment after becoming a teacher
   d. extended course of hospital treatment

7. STRANGLE: He **strangled** her.
   x a. killed her by pressing her throat
   b. gave her all the things she wanted
   c. took her away by force
   d. admired her greatly

8. CAVALIER: He treated her in a **cavalier** manner.
   x a. without care
   b. politely
   c. awkwardly
   d. as a brother would

9. MALIGN: His **malign** influence is still felt.
   x a. evil
   b. good
   c. very important
   d. secret

10. VEER: The car **veered**.
    x a. went suddenly in another direction
    b. moved shakily
    c. made a very loud noise
    d. slid sideways without the wheels turning

Seventh 1000

1. OLIVE: We bought **olives**.
x a. oily fruit
b. scented pink or red flowers
c. men's clothes for swimming
d. tools for digging up weeds

2. QUILT: They made a **quilt**.
a. statement about who should get their property when they die
b. firm agreement
x c. thick warm cover for a bed
d. feather pen

3. STEALTH: They did it by **stealth**.
a. spending a large amount of money
b. hurting someone so much that they agreed to their demands
x c. moving secretly with extreme care and quietness
d. taking no notice of problems they met

4. SHUDDER: The boy **shuddered**.
a. spoke with a low voice
b. almost fell
x c. shook
d. called out loudly

5. BRISTLE: The **bristles** are too hard.
a. questions
x b. short stiff hairs
c. folding beds
d. bottoms of the shoes

6. BLOC: They have joined this **bloc**.
a. musical group
b. band of thieves
c. small group of soldiers who are sent ahead of others
x d. group of countries sharing a purpose

7. DEMOGRAPHY: This book is about **demography**.
a. the study of patterns of land use
b. the study of the use of pictures to show facts about numbers
c. the study of the movement of water
x d. the study of population

8. GIMMICK: That's a good **gimmick**.
a. thing for standing on to work high above the ground
b. small thing with pockets to hold money
x c. attention-getting action or thing
d. clever plan or trick

9. AZALEA: This **azalea** is very pretty.
x a. small tree with many flowers growing in groups
b. light material made from natural threads
c. long piece of material worn by women in India
d. sea shell shaped like a fan

10. YOGHURT: This **yoghurt** is disgusting.
a. grey mud found at the bottom of rivers
b. unhealthy, open sore
x c. thick, soured milk, often with sugar and flavouring
d. large purple fruit with soft flesh

Eighth 1000

1. ERRATIC: He was **erratic**.
a. without fault
b. very bad
c. very polite
x d. unsteady

2. PALETTE: He lost his **palette**.
a. basket for carrying fish
b. wish to eat food
c. young female companion
x d. artist's board for mixing paints

3. NULL: His influence was **null**.
a. had good results
b. was unhelpful
x c. had no effect
d. was long-lasting

4. KINDERGARTEN: This is a good **kindergarten**.
a. activity that allows you to forget your worries
x b. place of learning for children too young for school
c. strong, deep bag carried on the back
d. place where you may borrow books

5. ECLIPSE: There was an **eclipse**.
a. a strong wind
b. a loud noise of something hitting the water
c. The killing of a large number of people
x d. The sun hidden by a planet

6. MARROW: This is the **marrow**.
a. symbol that brings good luck to a team
x b. Soft centre of a bone
c. control for guiding a plane
d. increase in salary

7. LOCUST: There were hundreds of **locusts**.
x a. insects with wings
b. unpaid helpers
c. people who do not eat meat
d. brightly coloured wild flowers

8. AUTHENTIC: It is **authentic**.
x a. real
b. very noisy
c. Old
d. Like a desert

9. CABARET: We saw the **cabaret**.
a. painting covering a whole wall
x b. song and dance performance
c. small crawling insect
d. person who is half fish, half woman

10. MUMBLE: He started to **mumble**.
a. think deeply
b. shake uncontrollably
c. stay further behind the others
x d. speak in an unclear way

Ninth 1000

1. HALLMARK: Does it have a **hallmark**?
   a. stamp to show when to use it by
x b. stamp to show the quality
   c. mark to show it is approved by the royal family
   d. Mark or stain to prevent copying

2. PURITAN: He is a **puritan**.
   a. person who likes attention
x b. person with strict morals
   c. person with a moving home
   d. person who hates spending money

3. MONOLOGUE: Now he has a **monologue**.
   a. single piece of glass to hold over his eye to help him to see better
x b. long turn at talking without being interrupted
   c. position with all the power
   d. picture made by joining letters together in interesting ways

4. WEIR: We looked at the **weir**.
   a. person who behaves strangely
   b. wet, muddy place with water plants
   c. old metal musical instrument played by blowing
x d. thing built across a river to control the water

5. WHIM: He had lots of **whims**.
   a. old gold coins
   b. female horses
x c. strange ideas with no motive
   d. sore red lumps

6. PERTURB: I was **perturbed**.
   a. made to agree
x b. Worried
   c. very puzzled
   d. very wet

7. REGENT: They chose a **regent**.
   a. an irresponsible person
   b. a person to run a meeting for a time
x c. a ruler acting in place of the king
   d. a person to represent them

8. OCTOPUS: They saw an **octopus**.
   a. a large bird that hunts at night
   b. a ship that can go under water
   c. a machine that flies by means of turning blades
x d. a sea creature with eight legs

9. FEN: The story is set in the **fens**.
x a. low land partly covered by water
   b. a piece of high land with few trees
   c. a block of poor-quality houses in a city
   d. a time long ago

10. LINTEL: He painted the **lintel**.
x a. Beam over the top of a door or window
   b. small boat used for getting to land from a big boat
   c. beautiful tree with spreading branches and green fruit
   d. board showing the scene in a theatre

Tenth 1000

1. AWE: They looked at the mountain with **awe**.
   a. worry
   b. interest
x c. wonder
   d. respect

2. PEASANTRY: He did a lot for the **peasantry**.
   a. local people
   b. place of worship
   c. businessmen's club
x d. poor farmers

3. EGALITARIAN: This organization is **egalitarian**.
   a. does not provide much information about itself to the public
   b. dislikes change
   c. frequently asks a court of law for a judgement
x d. treats everyone who works for it as if they are equal

4. MYSTIQUE: He has lost his **mystique**.
   a. his healthy body
x b. the secret way he makes other people think he has special power or skill
   c. the woman who has been his lover while he is married to someone else
   d. the hair on his top lip

5. UPBEAT: I'm feeling really **upbeat** about it.
   a. upset
x b. good
   c. hurt
   d. confused

6. CRANNY: We found it in the **cranny**!
   a. sale of unwanted objects
x b. narrow opening
   c. space for storing things under the roof of a house
   d. large wooden box

7. PIGTAIL: Does she have a **pigtail**?
x a. a rope of hair made by twisting bits together
   b. a lot of cloth hanging behind a dress
   c. a plant with pale pink flowers that hang down in short bunches
   d. a lover

8. CROWBAR: He used a **crowbar**.
x a. heavy iron pole with a curved end
   b. false name
   c. sharp tool for making holes in leather
   d. light metal walking stick

9. RUCK: He got hurt in the **ruck**.
   a. hollow between the stomach and the top of the leg
   b. pushing and shoving
x c. group of players gathered round the ball in some ball games
   d. race across a field of snow

10. LECTERN: He stood at the **lectern**.
x a. desk to hold a book at a height for reading
   b. table or block used for church sacrifices
   c. place where you buy drinks
   d. very edge

Eleventh 1000

1. EXCRETE: This was **excreted** recently.
   x a. pushed or sent out
   b. made clear
   c. discovered by a science experiment
   d. put on a list of illegal things

2. MUSSEL: They bought **mussels**.
   a. small glass balls for playing a game
   x b. shellfish
   c. large purple fruits
   d. pieces of soft paper to keep the clothes clean when eating

3. YOGA: She has started **yoga**.
   a. handwork done by knotting thread
   x b. a form of exercise for body and mind
   c. a game where a cork stuck with feathers is hit between two players
   d. a type of dance from eastern countries

4. COUNTERCLAIM: They made a **counterclaim**.
   x a. a demand made by one side in a law case to match the other side's demand
   b. a request for a shop to take back things with faults
   c. An agreement between two companies to exchange work
   d. a top cover for a bed

5. PUMA: They saw a **puma**.
   a. small house made of mud bricks
   b. tree from hot, dry countries
   c. very strong wind that sucks up anything in its path
   x d. large wild cat

6. PALLOR: His **pallor** caused them concern.
   a. his unusually high temperature
   b. his lack of interest in anything
   c. his group of friends
   x d. the paleness of his skin

7. APERITIF: She had an **aperitif**.
   a. a long chair for lying on with just one place to rest an arm
   b. a private singing teacher
   c. a large hat with tall feathers
   x d. a drink taken before a meal

8. HUTCH: Please clean the **hutch**.
   a. thing with metal bars to keep dirt out of water pipes
   b. space in the back of a car for bags
   c. metal piece in the middle of a bicycle wheel
   x d. cage for small animals

9. EMIR: We saw the **emir**.
   a. bird with long curved tail feathers
   b. woman who cares for other people's children in Eastern countries
   x c. Middle Eastern chief with power in his land
   d. house made from blocks of ice

10. HESSIAN: She bought some **hessian**.
    a. oily pinkish fish
    b. stuff producing a happy state of mind
    x c. coarse cloth
    d. strong-tasting root for flavouring food

Twelfth 1000

1. HAZE: We looked through the **haze**.
   a. small round window in a ship
   x b. unclear air
   c. strips of wood or plastic to cover a window
   d. list of names

2. SPLEEN: His **spleen** was damaged.
   a. knee bone
   x b. organ found near the stomach
   c. pipe taking waste water from a house
   d. respect for himself

3. SOLILOQUY: That was an excellent **soliloquy**!
   a. song for six people
   b. short clever saying with a deep meaning
   c. entertainment using lights and music
   x d. speech in the theatre by a character who is alone

4. REPTILE: She looked at the **reptile**.
   a. old hand-written book
   x b. animal with cold blood and a hard outside
   c. person who sells things by knocking on doors
   d. picture made by sticking many small pieces of different colours together

5. ALUM: This contains **alum**.
   a. a poisonous substance from a common plant
   b. a soft material made of artificial threads
   c. a tobacco powder once put in the nose
   x d. a chemical compound usually involving aluminium

6. REFECTORY: We met in the **refectory**.
   x a. room for eating
   b. office where legal papers can be signed
   c. room for several people to sleep in
   d. room with glass walls for growing plants

7. CAFFEINE: This contains a lot of **caffeine**.
   a. a substance that makes you sleepy
   b. threads from very tough leaves
   c. ideas that are not correct
   x d. a substance that makes you excited

8. IMPALE: He nearly got **impaled**.
   a. charged with a serious offence
   b. put in prison
   x c. stuck through with a sharp instrument
   d. involved in a dispute

9. COVEN: She is the leader of a **coven**.
   a. a small singing group
   b. a business that is owned by the workers
   x c. a secret society
   d. a group of church women who follow a strict religious life

10. TRILL: He practised the **trill**.
    x a. ornament in a piece of music
    b. type of stringed instrument
    c. Way of throwing a ball
    d. dance step of turning round very fast

on the toes

Thirteenth 1000

1. UBIQUITOUS: Many weeds are **ubiquitous**.
   a. are difficult to get rid of
   b. have long, strong roots
x c. are found in most countries
   d. die away in the winter

2. TALON: Just look at those **talons**!
   a. high points of mountains
x b. sharp hooks on the feet of a hunting bird
   c. heavy metal coats to protect against weapons
   d. people who make fools of themselves without realizing it

3. ROUBLE: He had a lot of **roubles**.
   a. very precious red stones
   b. distant members of his family
x c. Russian money
   d. moral or other difficulties in the mind

4. JOVIAL: He was very **jovial**.
   a. low on the social scale
   b. likely to criticize others
x c. full of fun
   d. friendly

5. COMMUNIQUE: I saw their **communiqué**.
   a. critical report about an organization
   b. garden owned by many members of a community
   c. printed material used for advertising
x d. official announcement

6. PLANKTON: We saw a lot of **plankton**.
   a. poisonous weeds that spread very quickly
x b. very small plants or animals found in water
   c. trees producing hard wood
   d. grey clay that often causes land to slip

7. SKYLARK: We watched a **skylark**.
   a. show with aeroplanes flying in patterns
   b. man-made object going round the earth
   c. person who does funny tricks
x d. small bird that flies high as it sings

8. BEAGLE: He owns two **beagles**.
   a. fast cars with roofs that fold down
   b. large guns that can shoot many people quickly
x c. small dogs with long ears
   d. houses built at holiday places

9. ATOLL: The **atoll** was beautiful.
x a. low island made of coral round a sea-water lake
   b. work of art created by weaving pictures from fine thread
   c. small crown with many precious jewels worn in the evening by women
   d. place where a river flows through a narrow place full of large rocks

10. DIDACTIC: The story is very **didactic**.
x a. tries hard to teach something
   b. is very difficult to believe
   c. deals with exciting actions
   d. is written in a way which makes the reader unsure of the meaning

Fourteenth 1000

1. CANONICAL: These are **canonical** examples.
   a. examples which break the usual rules
   b. examples taken from a religious book
x c. regular and widely accepted examples
   d. examples discovered very recently

2. ATOP: He was **atop** the hill.
   a. at the bottom of
x b. at the top of
   c. on this side of
   d. on the far side of

3. MARSUPIAL: It is a **marsupial**.
   a. an animal with hard feet
   b. a plant that grows for several years
   c. a plant with flowers that turn to face the sun
x d. an animal with a pocket for babies

4. AUGUR: It **augured** well.
x a. promised good things for the future
   b. agreed well with what was expected
   c. had a colour that looked good with something else
   d. rang with a clear, beautiful sound

5. BAWDY: It was very **bawdy**.
   a. unpredictable
   b. enjoyable
   c. rushed
x d. rude

6. GAUCHE: He was **gauche**.
   a. talkative
   b. flexible
x c. awkward
   d. determined

7. THESAURUS: She used a **thesaurus**.
x a. a kind of dictionary
   b. a chemical compound
   c. a special way of speaking
   d. an injection just under the skin

8. ERYTHROCYTE: It is an **erythrocyte**.
   a. a medicine to reduce pain
x b. a red part of the blood
   c. a reddish white metal
   d. a member of the whale family

9. CORDILLERA: They were stopped by the **cordillera**.
   a. a special law
   b. an armed ship
x c. a line of mountains
   d. the eldest son of the king

10. LIMPID: He looked into her **limpid** eyes.
x a. clear
   b. tearful
   c. deep brown
   d. beautiful

Appendix C

C.1 The Three Versions of the AVT (Study 1; Chap. 5)

Test 1

Student's ID: _____

- **Please tick (✓) the words that you KNOW at least one meaning of (and that you are SURE you can TRANSLATE or DEFINE)**
- **Please be as honest as possible. DO NOT TICK WORDS you think you DO NOT KNOW. There are many words in the list that do not exist, and ticking too many of them means I will not be able to use your data.**

adair		precise		automatic		text	
retrogradient		distribute		contrivial		bance	
detect		strategy		chapter		cottonwool	
creative		final		acknowledge		research	
stimulcrate		detailoring		awareness		community	
buttle		specifically		computer		analyze	
analysis		almanical		appropriately		consequence	
aistrope		clarification		approach		coherent	
consistently		nickling		complementary		crucial	
contextualize		adequately		despite		apparent	
minimize		input		register		function	
affective		stace		coding		automatically	
bastionate		bodelate		quorant		diversity	
distinctive		degate		demonstrate		cambule	
conclusion		snell		accurate		condimented	

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

benevolate		constantly		colleague		abrogative	
available		clause		contribute		accurately	
berrow		baldock		strategic		benefit	
series		assessment		overend		fundamental	
charlett		litholect		criteria		charactal	
achievement		menstruable		combustulate		procedure	
style		semaphrodite		batcock		process	
misabrogate		decade		identification		minimal	
lauder		dimension		consistent		rudge	
dramatically		corresponding		design		conceptual	
achieve		pauling		descript		isolation	
approximately		cantileen		core		nonagrate	
acquisition		loveridge		acklon		reservory	
scudamore		display		access		opie	
classic		analytic		attitude		dowrick	
create		comprehensive		beneficial		complexity	
constrained		oestrogeny		circumstance		generate	
derive		kiley		interaction		considerably	
distribution		moffat		dogmatile		challenge	
unique		concept		analyses		confirm	
adjust		aware		capacity		ambiguous	
ralling		brief		scurrilize		oxylate	
conduct		considerable		contortal		interpret	
empirical		area		constraint		assumption	
channing		constant		appendix		tradition	
mundy		traditional		potential		aid	
capable		functional		distinct		category	
contribution		lannery		pring		pocock	
clarify		acquire		visual		ridout	
journal		adult		section		lapidoscope	
affect		limidate		balfour		mode	
diverse		differentiate		constitute		cite	
pernicate		context		assistance		motivation	
minimum		recenticle		assign		contextual	
oligation		briefly		connery		accuracy	

Background information check

A) Age: _____ **B)** Gender: M F

C) Please note below all languages you have any knowledge of. Write these languages at the order of their acquisition and try to estimate your proficiency.

L1 (native language(s)) _____

L2 _____ A1 A2 B1 B2 C1 C2 L6 _____ A1 A2 B1 B2 C1 C2

L3 _____ A1 A2 B1 B2 C1 C2 L7 _____ A1 A2 B1 B2 C1 C2

L4 _____ A1 A2 B1 B2 C1 C2 L8 _____ A1 A2 B1 B2 C1 C2

L5 _____ A1 A2 B1 B2 C1 C2 L9 _____ A1 A2 B1 B2 C1 C2

Test 2

Student's ID: _____

- **Please tick (✓) the words that you KNOW at least one meaning of (and that you are SURE you can TRANSLATE or DEFINE)**
- **Please be as honest as possible. DO NOT TICK WORDS you think you DO NOT KNOW. There are many words in the list that do not exist, and ticking too many of them means I will not be able to use your data.**

pauling		statistical		ralling		equivalent	
aistrope		quorant		location		emerge	
condimented		infer		technology		label	
job		stace		insufficient		contrast	
conflict		menstruable		occur		scudamore	
emphasize		bastionate		nonagrate		justification	
theory		rudge		cultural		hence	
evaluation		stimulcrate		baldock		external	
cambule		emphasis		interactive		theoretical	
impact		oligation		abrogative		comment	
bodelate		cantileen		snell		error	
encounter		evolve		plus		orient	
overall		outcome		overend		definition	
alternative		pocock		indicate		bance	
methodology		ensure		connery		channing	

contrivial		inherent		mediate		technical	
method		misabrogate		hypothesis		duration	
initial		oxylate		berrow		charactal	
code		factor		recenticle		debate	
moffat		pring		interpretation		implicitly	
internal		issue		network		insight	
involve		gender		establish		format	
limidate		expose		maintain		initially	
dowrick		image		exclusively		imply	
exposure		balfour		item		instance	
buttle		perception		semaphrodite		mental	
construct		illustrate		project		intervention	
nevertheless		lauder		medium		involvement	
acklon		charlett		guideline		indication	
degate		major		focus		interact	
illustration		obtain		loveridge		evaluate	
communicate		finally		psychological		inadequate	
estimate		extract		combustulate		contemporary	
adair		majority		detailoring		culture	
lapidoscope		almanical		guarantee		incorporate	
irrelevant		contact		environment		mundy	
dogmatile		conclude		institutional		obvious	
reservory		identity		exhibit		integrate	
nickling		investigation		primarily		global	
batcock		occurrence		ridout		intermediate	
positive		element		lannery		notion	
grade		explicitly		furthermore		ongoing	
normal		explicit		scurrilize		option	
evident		investigate		cottonwool		stable	
descript		similarly		framework		contortal	
benevolate		appropriate		implicit		inappropriate	
interpretive		oestrogeny		highlight		retrogradient	
adequate		assist		link		norm	
dominant		litholect		opie		pernicate	
feature		expertise		normally		kiley	

Background information check

A) Age: _____ **B)** Gender: M F

C) Please note below all languages you have any knowledge of. Write these languages at the order of their acquisition and try to estimate your proficiency.

L1 (native language(s)) _____

L2 _____ A1 A2 B1 B2 C1 C2 L6 _____ A1 A2 B1 B2 C1 C2

L3 _____ A1 A2 B1 B2 C1 C2 L7 _____ A1 A2 B1 B2 C1 C2

L4 _____ A1 A2 B1 B2 C1 C2 L8 _____ A1 A2 B1 B2 C1 C2

L5 _____ A1 A2 B1 B2 C1 C2 L9 _____ A1 A2 B1 B2 C1 C2

Test 3

Student's ID: _____

- **Please tick (✓) the words that you KNOW at least one meaning of (and that you are SURE you can TRANSLATE or DEFINE)**
- **Please be as honest as possible. DO NOT TICK WORDS you think you DO NOT KNOW. There are many words in the list that do not exist, and ticking too many of them means I will not be able to use your data.**

abstract		scope		vary		topic	
tense		primary		site		task	
status		facilitate		similarity		whereby	
oestrogeny		orientation		survey		enable	
recenticle		portion		ethical		misabrogate	
berrow		sum		range		descript	
adair		underlying		role		nickling	
expert		random		summarize		media	
physical		practitioner		valid		overlap	
topical		pocock		reveal		contrivial	
prime		tension		author		response	
construction		reject		paradigm		phenomenon	
reservory		whereas		condimented		negative	
monitor		channing		aistrope		benevolate	
temporary		variable		assess		restricted	
perspective		acklon		seek		principle	
cottonwool		previous		moffat		intensive	
somewhat		instruction		require		rudge	
opie		participant		pernicate		limidate	

perceive		professional		overend		kiley	
degate		technique		previously		identify	
bance		aspect		stimulcrate		target	
stress		domain		requirement		relevant	
specific		bodelate		consult		percentage	
quorant		charactal		charlett		define	
output		almanical		parallel		scudamore	
conventional		assume		straightforward		implication	
complex		abrogative		prior		widespread	
ignore		reaction		buttle		dowrick	
shift		pring		similar		selection	
cambule		validity		stace		litholect	
reinforce		oligation		predict		resolve	
reliable		dynamic		bastionate		snell	
randomly		oxylate		textual		subsequent	
relevance		significantly		select		scurrilize	
consist		predictable		sufficiently		lapidoscope	
volume		qualitative		loveridge		significance	
reliably		baldock		contortal		sequence	
significant		lauder		component		sufficient	
detailoring		ralling		resource		respond	
structural		ridout		phenomena		balfour	
identical		mundy		traditionally		pauling	
phase		modified		instructor		semaphrodite	
structure		participation		mechanism		summary	
communicative		ultimately		potentially		medical	
enhance		nonagrate		connery		publish	
researcher		dogmatile		psychology		source	
menstruable		batcock		variability		retrogradient	
individual		period		cantileen		combustulate	
revise		rely		via		lannery	

Background information check

A) Age: _____ **B)** Gender: M F

C) Please note below all languages you have any knowledge of. Write these languages at the order of their acquisition and try to estimate your proficiency.

L1 (native language(s)) _____

L2 _____ A1 A2 B1 B2 C1 C2 L6 _____ A1 A2 B1 B2 C1 C2

L3 _____ A1 A2 B1 B2 C1 C2 L7 _____ A1 A2 B1 B2 C1 C2

L4 _____ A1 A2 B1 B2 C1 C2 L8 _____ A1 A2 B1 B2 C1 C2

L5 _____ A1 A2 B1 B2 C1 C2 L9 _____ A1 A2 B1 B2 C1 C2

Appendix D

D.1 The 46 Cognates and 44 Noncognates in the VST (Study 1; Chap. 5)

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Band 1:

Cognates	**Noncognates**
Figure	Period
Standard	Drive
Basis	Shoe

Band 2:

Cognates	**Noncognates**
Microphone	Patience
Pub	Nil

Band 3:

Cognates	**Noncognates**
Dinosaur	Restore

Band 4:

Cognates	**Noncognates**
Quiz	Input
Crab	Candid
Remedy	Tummy

Band 5:

Cognates	**Noncognates**
Deficit	Fracture
Compost	Haunt
Miniature	Weep
Bacterium	Peel

Band 6:

Cognates	**Noncognates**
Premier	Strangle
Accessory	Butler
Thesis	Malign

Band 7:

Cognates	**Noncognates**
Olive	Quilt
Bloc	Gimmick
Demography	Stealth
Azalea	Bristle
Yoghurt	Shudder

Band 8:

Cognates	**Noncognates**
Erratic	Locust
Palette	Marrow
Authentic	Mumble
Cabaret	Null

Band 9:

Cognates	**Noncognates**
Puritan	Octopus
Monologue	Hallmark
Perturbed	Lintel

Band 10:

Cognates	**Noncognates**
Egalitarian	Crowbar
Mystique	Peasantry

Band 11:

Cognates	**Noncognates**
Yoga	Hutch
Puma	Pallor
Aperitif	Excrete
Emir	Hessian

Band 12:

Cognates	**Noncognates**
Refectory	Impale
Caffeine	Spleen

Band 13:

Cognates	**Noncognates**
Talon	Ubiquitous
Jovial	Communiqué
Plankton	Roubles
Beagle	Skylark
Atoll	
Didactic	

Band 14:

Cognates	**Noncognates**
Canonical	Marsupial
Thesaurus	Augur
Erythrocyte	Baudy
Cordillera	Gauche

Appendix E

E.1 The Questionnaire (Studies 2 and 3; Chaps. 6 and 7)

Background information check

A) Full name: _____

B) Age: _____

C) Gender: M F

D) Please note below all languages you have any knowledge of. Write these languages at the order of their acquisition and try to estimate your proficiency.

A1 – beginner
A2 – elementary/pre-intermediate
B1 – intermediate
B2 – upper intermediate
C1 – advanced
C2 – proficient

L1 (native language(s)) _____

L2_____ A1 A2 B1 B2 C1 C2
L3 _____ A1 A2 B1 B2 C1 C2
L4 _____ A1 A2 B1 B2 C1 C2
L5 _____ A1 A2 B1 B2 C1 C2

L6 _____ A1 A2 B1 B2 C1 C2
L7 _____ A1 A2 B1 B2 C1 C2
L8_____ A1 A2 B1 B2 C1 C2
L9_____ A1 A2 B1 B2 C1 C2

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Please, give honest answers to the following questions. The answers will only be used for my research purposes and will not affect your course grades in any way.

1) Did you know what the purpose of the study was?
   a) I didn't know.
   b) I thought it was about (be as specific as possible):

_____

_____

2) Did you study the words after the classes? How did you remember them?

_____

_____

Appendix F

F.1 LexTALE (Studies 2 and 3; Chaps. 6 and 7).

Source: http://lextale.com/downloads/ExperimenterInstructionsEnglish.pdf

Below you find the items for the English version of the LexTALE test. You can implement the test in any experimental software, or as a paper and pencil test.

The columns contain the following information:

- First column: Item number. (Note that the first three items are dummies.)
- Second column: Item.
- Third column: Word status; 0 = nonword, 1 = word.

0	platery	0
0	denial	1
0	generic	1
1	mensible	0
2	scornful	1
3	stoutly	1
4	ablaze	1
5	kermshaw	0
6	moonlit	1
7	lofty	1
8	hurricane	1
9	flaw	1
10	alberation	0
11	unkempt	1
12	breeding	1
13	festivity	1
14	screech	1
15	savoury	1
16	plaudate	0
17	shin	1
18	fluid	1
19	spaunch	0
20	allied	1
21	slain	1
22	recipient	1
23	exprate	0
24	eloquence	1
25	cleanliness	1
26	dispatch	1
27	rebondicate	0
28	ingenious	1
29	bewitch	1
30	skave	0
31	plaintively	1
32	kilp	0
33	interfate	0
34	hasty	1
35	lengthy	1
36	fray	1
37	crumper	0
38	upkeep	1
39	majestic	1
40	magrity	0
41	nourishment	1
42	abergy	0
43	proom	0
44	turmoil	1
45	carbohydrate	1
46	scholar	1
47	turtle	1
48	fellick	0
49	destription	0
50	cylinder	1
51	censorship	1
52	celestial	1
53	rascal	1
54	purrage	0
55	pulsh	0
56	muddy	1
57	quirty	0
58	pudour	0
59	listless	1
60	wrought	1

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Scoring

The LexTALE score consists of the percentage of correct responses, corrected for the unequal proportion of words and nonwords in the test by averaging the percentages correct for these two item types. We call this measure % $correct_{av}$ (averaged % correct). It is calculated as follows:

((number of words correct/40 * 100) + (number of nonwords correct/20 * 100))/2

Note that the first three items are dummies; responses to those items should not be taken into account for the calculation of the score!

See also www.lextale.com/scoring.php.

Appendix G

G.1 Control and Treatment Essays (Studies 2 and 3; Chaps. 6 and 7)

Control essay (This title was omitted in the real task)

> Do you agree or disagree with the following statement?
>
> Technology has made the world a better place to live.
>
> Use specific reasons and examples to support your opinion.
>
> *An effective essay will usually contain a minimum of 300 words; however, you may write more if you wish.* (between 300 and 400 words)

Write as fast as you can. You have 60 min. When you finish, **write below the essay the time you took to write it**.

Send to my personal email: brenotesol@gmail.com. **Please use your full name as the name of the file!** For example:

My full name: Breno Silva

File name: Breno Silva.docx

Unstructured essay (This title was omitted in the real task)

Using all 10 words given in the glossary, write an essay answering the question below. Remember the following:

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

- U**se each word once**, only! **You may use them in any order you want**.
- Do not **change the part of speech of the words provided in the glossary**. You may derive the word, but not change the part of speech. For example, the verb "believe":
  - It's **OK** to use "believed", "believes" or "believing".
  - It's **NOT OK** to use the noun "belief". Don't change the part of speech.
- Read the "IMPORTANT" message below carefully.

Question: Do you agree or disagree with the following statement?

People should sometimes do things that they do not enjoy doing.

Use specific reasons and examples to support your opinion.

An effective essay will usually contain a minimum of 300 words; however, you may write more if you wish. (Write between 300 and 400 words)

You have 60 min

IMPORTANT:

- The **most important thing** in this task is to **use the 10 words** in your essay. You need to write the essay (300–400 words), of course, but **focus on using the words appropriately**. Remember you have 60 min to use all the words in your essay (each word once, only!).
- **Use the words correctly** in your text. For example, use the correct spelling, correct prepositions and suchlike.
- Make sure all words are used **once only** (with the appropriate part of speech).

When you finish, write below the essay the time (in minutes) you took to write it. Please also write if you used glossary A or B to write the essay.

Send to my personal email: brenotesol@gmail.com.

Structured essay (This title was omitted in the real task)

Using all 10 words given in the glossary, write an essay answering the question below. Remember the following:

- Use each word once, only! You may use them in any order you want.

- Do not **change the part of speech of the words provided in the glossary**. You may derive the word, but not change the part of speech. For example, the verb "believe":
  - It's **OK** to use "believed", "believes" or "believing".
  - It's NOT OK to use the noun "belief". Don't change the part of speech.
- Read the "IMPORTANT" message below carefully.

> Question: Do you agree or disagree with the following statement?
>
> Parents are the best teachers.
>
> Use specific reasons and examples to support your opinion.
>
> *An effective essay will usually contain a minimum of 300 words; however, you may write more if you wish.* (Write between 300 and 400 words)
>
> *You have 60 min*

IMPORTANT:

The **focus is on the quality** of the essay. Remember the following:

- **Plan** what to write carefully.
- Make sure to have a **good introduction and conclusion**, and to have your ideas clearly divided into **paragraphs**.
- Each paragraph must start with **a topic sentence**. Also, remember to **link your sentences clearly**, so the whole essay is very easy to read and understand.
- In general, remember to write clearly and to answer the question as well as possible. **This essay must be clear and well structured**.
- Don't forget to use each **word only once**. However, remember you have 60 min to write your essay and you should **focus on the quality of your text**.

When you finish, write below the essay the time you took to write it. Please also write if you used glossary A or B to write the essay.

Send to my personal email: brenotesol@gmail.com.

Appendix H

H.1 Glossaries (Studies 2 and 3; Chaps. 6 and 7)

Glossary A

There are **10 words** below. Read carefully and **understand the meaning of the words and how to use them** You can **use each word once only in your writing**. As you use them, tick the words in the space provided. You may use any of the meanings provided.

( ) Apparent (adjective):
(1) Able to be seen or understood:
- *[that + clause] It was becoming increasingly apparent **that** he could no longer look after himself.*

(2) Seeming to exist or be true:
- *She has this apparent innocence which, I suspect, she uses to her advantage.*

( ) Constitute (verb):
(1) To be or be considered as something:
- *The rise in crime constitutes a threat to society.*

(2) To be the parts that form it:
- *We must redefine what constitutes a family.*

( ) Insight (noun): (The ability to have) a clear, deep, and sometimes sudden understanding of a complicated problem or situation:
- *It was an interesting book, full of fascinating insights **into** human relationships.*
- *His book offers some fresh insights **into** the events leading up to the war.*

( ) Implicit (adjective): Suggested but not communicated directly:
- *He interpreted her comments as an implicit criticism of the government.*
- *Her words contained an implicit threat.*

( ) Differentiate (verb): To show or find the difference between things that are compared:
- *We do not differentiate **between** our employees on the basis of their race, religion, or national origin.*
- *It's sometimes hard to differentiate one sample from another*

( ) Acquisition (noun): The process of getting something:
- *The acquisition **of** huge amounts of data has helped our research enormously.*
- *Language acquisition starts at a very young age.*

( ) Ongoing (adjective): Continuing to exist or develop, or happening at the present moment:
- *No agreement has yet been reached and the negotiations are still ongoing.*
- *There are a number of ongoing difficulties with the project.*

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

( ) **Incorporate** (verb): To include something as part of something larger:
- *Suggestions from the survey have been incorporated **into/in** the final design.*
- *This aircraft incorporates several new safety features.*

( ) **Paradigm** (noun):
(1) A model of something, a very clear and typical example of something.
- *Some of these educators are hoping to produce a change in the current cultural paradigm.*

(2) A set of theories that explain the way a particular subject is understood.
- *His account of the effects of globalization does not fit into either of the economic paradigms that are dominant today.*

( ) **Constraint** (noun): Something that limits your freedom to do what you want
- *Among those conditions was a five-year time constraint. (= time limit).*
- *There have been financial and political constraints **on** development.*

Glossary B

There are **10 words** below. Read carefully and **understand the meaning of the words and how to use them** You can **use each word once only in your writing** . As you use them, tick the words in the space provided. You may use any of the meanings provided.

( ) **Qualitative** (adjective): Based on information that cannot be easily measured, such as people's opinions and feelings, rather than on information that can be shown in numbers:
- *The research involves qualitative analysis of students' performance .*
- *It may be necessary to use a more varied form of assessment, i.e. one that generates both qualitative and quantitative information*

( ) **Reinforce** (verb):
(1) If something reinforces an idea or opinion, it provides more proof or support for it:
- *The film reinforces the idea that women should be pretty and dumb.*

(2) To make a structure or group of people (e.g., the military), stronger.
- *The sea wall at Southend is being reinforced with tons of cement*

( ) **Assessment** (noun):
(1) The act of judging or deciding the amount, value, quality, or importance of something:
- *Would you say that is a fair assessment **of** the situation?*

(2) The judgment or decision that is made:
- *Both their assessments **of** production costs were hopelessly inaccurate.*

( ) **Inherent** (adjective): Existing as a natural or basic part of something that cannot be separated from it
- *Every business has its own inherent risks.*
- *There are risks inherent **in** almost every sport.*

( ) **Derive** (verb): To get something from something else:
- *He derives an enormous amount of satisfaction **from** restoring old houses.*
- *She derives great pleasure **from** playing the violin.*

( ) **Variability** (noun): The quality or fact of being variable (= likely to change often):
- *I've spent 10 years researching how much genetic variability there is **between** populations.*
- *Wide variability exists **between** women in the age at which menopause occurs.*

( ) **Affective** (adjective): Connected with, or having an effect on, the emotions:
- *He has no affective ties to his family.*
- *Of most importance to teachers were affective aims relating to the personal development of children.*

( ) **Facilitate** (verb): To make something possible or easier to happen:
- *Computers can be used to facilitate language learning.*
- *Both centers are electronically linked to facilitate communication.*

( ) Phenomenon (noun; plural:**phenomena**): Something that exists and can be seen, felt, tasted, etc., especially something unusual or interesting.

- *The Beatles were a phenomenon - nobody had heard anything like them before.*
- *Do you believe in the paranormal and other psychic phenomena?*

( ) Validity (noun):

(1) The quality of being based on truth or reason, or of being able to be accepted.

- *This research seems to give some validity to the theory that the drug might cause cancer.*

(2) When a document is legally accepted

- *There is no requirement for the content provider to check the validity **of** the credit card number.*

Appendix I

I.1 The VKS and Association Test (Studies 2 and 3; Chaps. 6 and 7)

Instructions (SET A):

Options I to V—choose the most suitable option

- The higher the number you choose, the more points you make (if the answer is correct).
- If you choose III or IV, please do V too. If you decide you can do V, please also do IV.
- Note that the part of speech (adjective, noun, verb) is given after each word. Please provide your translation and/or write your sentence based on the part of speech given.

Option VI—write down as many as four words that you associate with the word given. If you can't think of 4 words, it's not necessary to write all of them. For example:

1. *Relationship* (*noun*)
VI. I associate this word with: friendship, love, trust, _____.

Please **write your FULL NAME:** _____

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Test:

1. *Apparent* (*adjective*)

I. I don't remember having seen this word before. ( )

II. I have seen this word before, but I don't know what it means. ( )

III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).

IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).

V. I can use this word correctly in a sentence in English. Write your sentence here:

_____

VI. I associate this word with _____, _____, _____, _____.

2. *Constraint* (*noun*)

I. I don't remember having seen this word before. ( )

II. I have seen this word before, but I don't know what it means. ( )

III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).

IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).

V. I can use this word correctly in a sentence in English. Write your sentence here:

_____

VI. I associate this word with _____, _____, _____, _____.

3. *Insight* (*noun*)

I. I don't remember having seen this word before. ( )

II. I have seen this word before, but I don't know what it means. ( )

III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).

IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).

V. I can use this word correctly in a sentence in English. Write your sentence here:

_____

VI. I associate this word with _____, _____, _____, _____.

4. *Implicit* (*adjective*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

5. *Constitute* (*verb*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

6. *Differentiate* (*verb*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

7. *Ongoing* (*adjective*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).

V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

8. *Incorporate* (*verb*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

9. *Acquisition* (*noun*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

10. *Paradigm* (*noun*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

Please **write your FULL NAME:** _____

Instructions (SET B):

Options I to V—choose the most suitable option

- The higher the number you choose, the more points you make (if the answer is correct).

- If you choose III or IV, please do V too. If you decide you can do V, please also do IV.
- Note that the part of speech (adjective, noun, verb) is given after each word. Please provide your translation and/or write your sentence based on the part of speech given.

Option VI—write down as many as four words that you associate with the word given. If you can't think of 4 words, it's not necessary to write all of them. For example:

2. *Relationship* (*noun*)
VI. I associate this word with: friendship, love, trust, _____.

Please **write your FULL NAME:** _____

Test:
1. *Qualitative* (*adjective*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
VI. I associate this word with _____, _____, _____, _____.

2. *Affective* (*adjective*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
VI. I associate this word with _____, _____, _____, _____.

3. *Assessment* (*noun*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

4. *Variability* (*noun*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

5. *Derive* (*verb*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

6. *Reinforce* (*verb*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).

V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

7. *Facilitate* (*verb*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

8. *Validity* (*noun*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

9. *Phenomenon* (*noun*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

10. *Inherent* (*adjective*)
I. I don't remember having seen this word before. ( )
II. I have seen this word before, but I don't know what it means. ( )
III. I have seen this word before. I *think* it means_____ (synonym, translation, or brief explanation).
IV. I *know* this word. It means _____ (synonym, translation, or brief explanation).
V. I can use this word correctly in a sentence in English. Write your sentence here:
_____
_____
VI. I associate this word with _____, _____, _____, _____.

Please **write your FULL NAME:** _____

Appendix J

J.1 Two Sample Essays (Studies 2 and 3; Chaps. 6 and 7)

Essay 1: Unstructured Timed CW (original from student)
Vocabulary Set B
Mark: 2.5

Topic: People should sometimes do things that they do not enjoy doing

First of all, I want to say that our lives is a sequence of actions. It consists of your actions and society which are you in. Some people says that only you can change own life, but it is not true. You just can do all possible to **facilitate** your being. You have to choose more rational solution and accept challenges, which **reinforce** character. So, yes, people should sometimes do things that they do not enjoy doing. It may be hard or unpleasant but they are **inherent** things in everyday life.

I separate two kind of things we must to do. First category is vital things for normal functionality. For example, visiting doctors, especially when you are ill, paying taxes, buying food and other. It seems easy but I know many people who hate or scary hospitals and doctors. But they have no choice, they should ask for a **qualitative** help to be alive. Second category is more usual things like a home duties, reading a books which you need for your study, communication with strangers in banks, shops, railway stations. They are less important but if we do not do it our lives become full of chaos. Actually I have **validity** of my words with fact. I saw in TV show that there is s psychic **phenomenon** in the world when people refuse to clean their house and their home becomes look like a landfill.

So, I think our **assessments** of what we do should be provided by good sense. Of course, it exists **variability** between different ages, because how older we are more responsibilities we have. **Affective** distribution of responsibilities it is the best decision in this case. As for me, I **derive** harmony in my life from orderliness and it is reinforce my opinion.

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Essay 2: Structured Timed CW and Untimed CW (original from student)
Vocabulary Set A
Mark: 5

Topic: Parents are the best teachers

It is clearly **apparent** that parents take an important part in everyone's lives. They take care of us from the very beginnings of our existence. During the time they spend with us they want to give much information they consider as valuable **insight**s into the life. There are many things they can teach us but we can't say they are the best teachers at everything.

Parents teach us the very first things we need to survive in this world. They are for sure good at that part. It is easier to put simple knowledge in a still simple little person. They helped us put our first steps, say our first words. Most of the words children use are due to the vocabulary **acquisition** they undergo at home. Thanks to them we have been able to move around and communicate with others in the early stages of living and develop it through the years. We can also learn from parents to **differentiate** between good and bad. During the part of upbringing a child they try to **incorporate** some teaching about morality. They do that by simple actions, for example tell their child not to laugh at somebody. This is a great **paradigm** of parental teaching including the basics, which is still and **ongoing** process as they pour the knowledge into us even when we're adults.

The examples above **constitute** to parents being good at teaching the valuable basics of living. But when it comes to other things, the fact that they are that close to us becomes a **constraint** in the teaching process. Parents cannot really be objective and consider their child just a student. They can see their children as best at doing something and just compliment them, not really correcting their work. For example when a daughter gives her mom a homework paper to correct, the mother may just say it's perfect. But there might have actually been some mistakes she hadn't seen because the emotions clouded her judgement. Another thing is that parents can get impatient teaching things that seem obvious to themselves. A dad is teaching his teenage daughter or son basics of driving a car. He doesn't have the patience for that and the way he speaks and behaves is interpreted by his children as an **implicit** anger, criticism. This is not what a teacher should do.

Considering the arguments above I can say that parents can be good teachers at some subjects of life, but not all of them. They are able to pour the basic knowledge during the time of growing up and even some moral instructions later in life. But their emotions and closeness their share with children is actually an obstruction to objective teaching. All in all, parents can teach us much but some things they should leave for others to teach.

Appendix K

K.1 (Study 2; Chap. 6): VKS_6 Model

Parameters	Model 1	Model 2[a]	Model 3	Model 4	Model 5	Model 6	Model 7
Fixed effects							
Group							
Time							
Condition							
Group * Time							
Condition * Group * Time							
Random effects							
Participants (intercept)	0.50*** (0.14)	0.30 (0.15)	0.49*** (0.15)	0.54*** (0.14)	0.56*** (0.14)	0.56*** (0.14)	0.53*** (0.16)
Time | Participants slope		0.50** (0.16)	0.50** (0.16)				0.50*** (0.16)
Time | Participants (correlation intercept-slope)				−0.32 (0.32)	−0.29 (0.32)	−0.29 (0.33)	
Group | Participants slope		– [b]					
Group | Participants (correlation intercept-slope)							
Condition | Participants slope		0.07 (0.08)	0.07 (0.08)	0.08 (0.09)			

(continued)

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

(continued)

Parameters	Model 1	Model 2[a]	Model 3	Model 4	Model 5	Model 6	Model 7
Condition | Participants (correlation intercept-slope)							
Items (intercept)	0.52*** (0.08)	0.37*** (0.11)	0.57*** (0.11)	0.58*** (0.11)	0.60*** (0.11)	0.69*** (0.09)	0.69*** (0.09)
Time | Items slope		– [b]					
Time | Items (correlation intercept-slope)							
Group | Items slope		– [b]					
Group | Items (correlation intercept-slope)							
Condition | Items slope		0.22 (0.17)	0.22 (0.17)	0.22 (0.17)	0.20 (18)		
Condition | Items (correlation intercept-slope)							
Model summary							
Deviance statistic (AIC corrected)	10,535.87	11,881.16	11,875.11	11,957.15	11,839.70	11,739.26	11,675.14
Covariance structure	VC	VC	VC	AR1	AR1	AR1	VC
				VC Condition slopes	VC Condition | items	VC (items)	
Next step:	Add slopes	Remove redundant		Remove condition | participants	Remove condition | items	Remove time | participant correlation	Back to model 1 (best fit). Add fixed effects

Note Parameter estimate standard error listed in parentheses. Number of data points = 1560; items = 40; participants = 39. Probability distribution: multinomial; link function: cumulative negative log–log. VC = variance components; AR1 = first-order autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Redundant

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Parameters	Model 8	Model 9[c]	Model 10	Final model	Model 13[a]
Fixed effects					
Group	0.09 (0.27)	0.09 (0.27)	0.09 (0.27)	*0.08 (0.26)*	0.07 (0.23)
Time	−0.98*** (0.12)	−0.98*** (0.12)	−0.98*** (0.12)	*−1.01*** *(0.12)*	−0.85*** (0.12)
Condition	−0.03 (0.12)	−0.03 (0.12)	−0.03 (0.12)	*−0.03 (0.12)*	−0.04 (0.10)
Group * Time	0.35* (0.15)	0.35* (0.15)	0.35* (0.15)	*0.38* (0.1)*	0.38* (0.15)
Group * Condition	0.05 (0.22)	0.05 (0.22)	0.05 (0.22)	*0.04 (0.21)*	0.06 (0.18)
Time * Condition	0.26 (0.15)	0.26 (0.15)	0.26 (0.15)	*0.26 (0.15)*	0.26 (0.15)
Condition * Group * Time	−0.23 (0.21)	−0.23 (0.21)	−0.23 (0.21)	*−0.21 (0.21)*	−0.21 (0.19)
Random effects					
Participants (intercept)	0.60*** (0.17)	0.60*** (0.17)	0.60*** (0.17)	*0.29*** *(0.07)*	0.26*** (0.06)
Time | Participants slope		0.00 (0.04)	0.00 (0.04)		
Time | Participants (correlation intercept-slope)				*0.30 (0.31)*	0.49 (0.30)
Group | Participants slope					
Group | Participants (correlation intercept-slope)					
Condition | Participants slope					
Condition | Participants (correlation intercept-slope)					
Items (intercept)	0.80*** (0.10)	0.80*** (0.10)	0.80*** (0.10)	*0.77*** *(0.10)*	0.21*** (0.03)
Time | Items slope		– [b]			
Time | Items (correlation intercept-slope)					– [b]
Group | Items slope					
Group | Items (correlation intercept-slope)					

(continued)

(continued)

Parameters	Model 8	Model 9[c]	Model 10	Final model	Model 13[a]
Condition | Items slope					
Condition | Items (correlation intercept-slope)					
Model summary					
Deviance statistic (AIC corrected)	12,728.79	12,728.82	12,726.81	*12,029.92*	10,598.78
Covariance structure	VC	VC	VC	*AR1*	AR1
				VC (items)	
Next step:	Add slopes	Remove time | items	Add time | participant correlation	*Add time | items (AR1)*	

Note Parameter estimate standard error listed in parentheses. Number of data points = 1560; items = 40; participants = 39. Probability distribution: multinomial; link function: cumulative negative log–log. VC = variance components; AR1 = first-order autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Redundant

[c]Error = convergence not achieved

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

K.2 (Study 2): VKS_3 Model

Parameters	Model 1	Model 2[a,b]	Model 3	Model 4	Model 5	Model 6	Model 7
Fixed effects							
Group							0.08 (0.26)
Time							−0.86*** (0.10)
Condition							−0.09 (0.11)
Group * Time							0.42** (0.13)
Group * Condition							0.07 (0.21)
Time * Condition							0.25* (0.13)
Condition * Group * Time							−0.19 (0.18)
Random effects							
Participants (intercept)	0.48*** (0.13)	0.16 (0.13)	0.41*** (0.13)	0.41*** (0.13)	0.37*** (0.10)	0.62*** (0.11)	0.52*** (0.15)
Time | Participants slope		0.31** (0.11)	0.31** (0.11)	0.31** (0.11)		0.58*** (0.11)	
Time | Participants (correlation intercept-slope)					0.71* (0.33)	0.72* (0.34)	
Group | Participants slope		– [c]					
Group | Participants (correlation intercept-slope)							
Condition | Participants slope		0.03 (0.06)	0.03 (0.06)				
Condition | Participants (correlation intercept-slope)							
Items (intercept)	0.43*** (0.07)	0.27*** (0.16)	0.46*** (0.08)	0.46*** (0.07)	0.48*** (0.08)	0.48*** (0.08)	0.54*** (0.08)
Time | Items slope		– [c]					

(continued)

(continued)

Parameters	Model 1	Model 2[a,b]	Model 3	Model 4	Model 5	Model 6	Model 7
Time | Items (correlation intercept-slope)							
Group | Items slope		– [c]					
Group | Items (correlation intercept-slope)							
Condition | Items slope		– [c]					
Condition | Items (correlation intercept-slope)							
Model summary							
Deviance statistic (AIC corrected)	6084.76	6197.78	6189.71	6186.80	6284.35	6290.00	6407.82
Covariance structure	VC	VC	VC	VC	AR1	ARH1	VC
					VC (items)	VC (items)	
Next step:	Add slopes	Remove redundant	Remove condition | participants	Add correlation (AR1)	Add correlation and slope (ARH1)	Back to model 1 (best fit). Add fixed effects	Add time | participants slope

Note Parameter estimate standard error listed in parentheses. Number of data points = 1560; items = 40; participants = 39. Probability distribution: multinomial; link function: cumulative negative log–log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite

[a]Error = the estimated covariance matrix is not positive definite

[b]Error = Hessian matrix is not positive definite

[c]Redundant

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Parameters	Model 8[a,b]	Final model	Model 11[b]
Fixed effects			
Group	0.08 (0.26)	*0.05 (0.23)*	0.05 (0.22)
Time	−0.86*** (0.10)	*−0.91*** (0.11)*	−0.82*** (0.11)
Condition	−0.09 (0.11)	*−0.07 (0.11)*	−0.07 (0.10)
Group * Time	0.42** (0.13)	*0.50** (0.15)*	0.51** (0.15)
Group * Condition	0.07 (0.21)	*0.05 (0.19)*	0.05 (0.18)
Time * Condition	0.25* (0.13)	*0.25* (0.12)*	0.24* (0.12)
Condition * Group * Time	−0.19 (0.18)	*−0.16 (0.18)*	−0.17 (0.19)
Random effects			
Participants (intercept)	0.52*** (0.15)	*0.21*** (0.06)*	0.22*** (0.06)
Time | Participants slope	– [c]		
Time | Participants (correlation intercept-slope)		*0.98*** (0.29)*	0.96*** (0.26)
Group | Participants slope			
Group | Participants (correlation intercept-slope)			
Condition | Participants slope			
Condition | Participants (correlation intercept-slope)			
Items (intercept)	0.54*** (0.08)	*0.51*** (0.08)*	0.22*** (0.03)
Time | Items slope			
Time | Items (correlation intercept-slope)			– [c]
Group | Items slope			
Group | Items (correlation intercept-slope)			
Condition | Items slope			
Condition | Items (correlation intercept-slope)			
Model summary			
Deviance statistic (AIC corrected)	6409.83	*6290.88*	6286.32
Covariance structure	VC	*AR1*	AR1
		VC (items)	
Next step:	Add correlation (AR1); Remove slope	*Add time | items correlation*	

Note Parameter estimate standard error listed in parentheses. Number of data points = 1560; items = 40; participants = 39. Probability distribution: multinomial; link function: cumulative negative log–log. VC = variance components; AR1 = first-order autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite
[a]Error = the estimated covariance matrix is not positive definite
[b]Error = Hessian matrix is not positive definite
[c]Redundant
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

K.3 (Study 2): Association Model

Parameters	Model 1	Model 2[a]	Model 3	Model 4	Model 5	Model 6	Model 7
Fixed effects							
Intercept	−0.74*** (0.15)	−0.94*** (0.15)	−0.94*** (0.15)	−0.87*** (0.15)	−0.80*** (0.15)	−0.85*** (0.15)	−1.05*** (0.22)
Group							−0.06 (0.37)
Time							0.76*** (0.14)
Condition							0.10 (0.18)
Group * Time							−0.45* (0.20)
Group * Condition							−0.31 (0.26)
Time * Condition							−0.26 (0.14)
Condition * Group * Time							0.39* (0.20)
Random effects							
Participants (intercept)	0.79*** (0.23)	0.37 (0.23)	0.74*** (0.23)	0.64*** (0.18)	0.92*** (0.19)	0.92*** (0.17)	0.80*** (0.23)
Time | Participants slope		0.43** (0.15)	0.43** (0.15)		0.70*** (0.14)	0.69*** (0.14)	
Time | Participants (correlation intercept-slope)				−0.21 (0.33)	−0.30 (0.42)	−0.25 (0.42)	
Group | Participants slope		_ [b]					
Condition | Participants slope		0.12 (0.11)	0.12 (0.11)	0.13 (0.11)	0.13 (0.11)		

(continued)

(continued)

Parameters	Model 1	Model 2[a]	Model 3	Model 4	Model 5	Model 6	Model 7
Condition | Participants (correlation intercept-slope)							
Items (intercept)	0.51*** (0.07)	0.24*** (0.07)	0.49*** (0.07)	0.49*** (0.07)	0.49*** (0.07)	0.52*** (0.07)	0.52*** (0.07)
Time | Items slope		– [b]					
Time | Items (correlation intercept-slope)							
Group | Items slope		– [b]					
Group | Items (correlation intercept-slope)							
Condition | Items slope		– [b]					
Condition | Items (correlation intercept-slope)							
Model summary							
Deviance statistic (AIC corrected)	5701.04	5726.88	5718.82	5708.08	5716.53	5714.04	5715.18
Covariance structure	VC	VC	VC	AR1	ARH1	ARH1	VC
				VC	VC	VC (items)	
Next step:	Add slopes	Remove redundant	Add time | participant correlation (AR1)	Add time | participant correlation and slope (ARH1)	Remove condition | participant slope	Return to model 1 (better fit). Add fixed effects	Add time | participants slope

Note Parameter estimate standard error listed in parentheses. Number of data points = 1560; items = 40; participants = 39. Probability distribution: Poisson; link function: log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Predictor reference categories (descending) = SW, pretest. Target reference category = ascending. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Redundant

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Parameters	Model 8	Model 9[a]	Model 10	Final model	Model 12	Model 13
Fixed effects						
Intercept	−1.07*** (0.24)	−1.07*** (0.24)	−1.11*** (0.24)	*−1.12*** (0.24)*	−1.09*** (0.24)	−1.09*** (0.24)

(continued)

(continued)

Parameters	Model 8	Model 9[a]	Model 10	Final model	Model 12	Model 13
Group	−0.07 (0.37)	−0.07 (0.37)	−0.12 (0.36)	*−0.14 (0.36)*	−0.08 (0.37)	−0.08 (0.37)
Time	0.76*** (0.15)	0.79*** (0.15)	0.87*** (0.16)	*0.92*** *(0.16)*	0.87*** (0.14)	0.86*** (0.14)
Condition	0.10 (0.18)	0.10 (0.18)	0.11 (0.19)	*0.13 (0.19)*	0.12 (0.18)	0.12 (0.18)
Group * Time	−0.47* (0.21)	−0.47* (0.20)	−0.52* (0.23)	*−0.52** *(0.22)*	−0.45* (0.20)	−0.45* (0.20)
Group * Condition	−0.30 (0.26)	−0.30 (0.26)	−0.30 (0.26)	*−0.31 (0.26)*	−0.31 (0.26)	−0.31 (0.26)
Time * Condition	−0.27 (0.14)	−0.27 (0.14)	−0.27 (0.15)	*−0.27 (0.15)*	−0.26 (0.14)	−0.26 (0.14)
Condition * Group * Time	0.39* (0.19)	0.39* (0.19)	0.37* (0.19)	*0.36 (0.19)*	0.36 (0.20)	0.36 (0.20)
Random effects						
Participants (intercept)	0.85*** (0.25)	0.85*** (0.25)	0.59*** (0.17)	*0.57*** *(0.16)*	0.82*** (0.24)	0.81*** (0.23)
Time | Participants slope	0.07 (0.06)	0.07 (0.06)			0.03 (0.05)	
Time | Participants (correlation intercept-slope)			−0.39 (0.22)	*−0.32 (0.25)*		
Group | Participants slope						
Group | Participants (correlation intercept-slope)						
Condition | Participants slope						
Condition | Participants (correlation intercept-slope)						
Items (intercept)	0.52*** (0.07)	0.52*** (0.07)	0.52*** (0.07)	*0.61*** *(0.11)*	0.62*** (0.11)	0.63*** (0.11)
Time | Items slope		– [b]				

(continued)

(continued)

Parameters	Model 8	Model 9[a]	Model 10	Final model	Model 12	Model 13
Time | Items (correlation intercept-slope)				*−0.73*** (0.08)*	−0.73*** (0.08)	−0.73*** (0.08)
Group | Items slope						
Group | Items (correlation intercept-slope)						
Condition | Items slope						
Condition | Items (correlation intercept-slope)						
Model summary						
Deviance statistic (AIC corrected)	5723.84	5724.43	5709.68	*5667.63*	5683.06	5678.66
Covariance structure	VC	VC	AR1	*AR1*	VC	
			VC (items)		AR1 (items)	
Next step:	Add time | item slope	Add time | participant correlation (AR1)	Add time | item correlation (AR1)	*Remove time | participants correlation. Add slope*	Remove time | participants slope	

Note Parameter estimate standard error listed in parentheses. Number of data points = 1560; items = 40; participants = 39. Probability distribution: Poisson; link function: log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Predictor reference categories (descending) = SW, pretest. Target reference category = ascending. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Redundant

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Appendix L

L.1 Self-rating Scale to Measure Cognitive Load (Study 3; Chap. 7)

CW Tasks

Participant name: _____

Please, give honest answers to the following questions. The answers will only be used for my research purposes and will not affect your course grades in any way.

1. How difficult was it to write the essay using the 10 words given? Please assess it on a scale of 1 to 6, where 1 means "very easy" and 6 means "very difficult". ( )
2. When completing this task, I needed to keep many things in mind at the same time. Please assess this statement on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree". ( )
3. The need to use the 10 words made this essay more mentally demanding than the first essay, where I did not need to use the 10 words. Please assess this statement on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree". ( )
4. How stressed, annoyed or frustrated were you while writing this essay? Please assess it on a scale of 1 to 6, where 1 means "very low" and 6 means "very high". ( )
5. Writing this essay with 10 words made me more stressed, annoyed or frustrated than writing the first essay (without the 10 words). Please assess it on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree". ( )

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

SW Task

Participant's name: _____

Please, give honest answers to the following questions. The answers will only be used for my research purposes and will not affect your course grades in any way.

1. How difficult was it to write the sentences using the 10 words given? Please assess it on a scale of 1 to 6, where 1 means "very easy" and 6 means "very difficult". ( )
2. When completing this task (sentence writing), I needed to keep many things in mind at the same time. Please assess this statement on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree". ( )
3. The need to use the 10 words made writing these sentences more mentally demanding than writing the first essay, where I did not need to use the 10 words. Please assess this statement on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree". ( )
4. How stressed, annoyed, or frustrated were you while writing the sentences? Please assess it on a scale of 1 to 6, where 1 means "very low" and 6 means "very high". ( )
5. Writing these sentences with 10 words made me more stressed, annoyed, or frustrated than writing the first essay (without the 10 words). Please assess it on a scale of 1 to 6, where 1 means "completely disagree" and 6 means "completely agree". ( )

Appendix M

M.1 Descriptive Statistics for Imputation Procedure for WM_Scores (Study 3; Chap. 7)

Data	Imputation	*N*	Mean	*SD*	Min	Max
Original data		1160	6.7328	1.37529	4.0000	10.5000
Imputed values	1	640	6.6143	1.37088	4.0239	10.2873
	2	640	6.8345	1.31645	4.0151	10.4180
	3	640	6.6444	1.34578	4.0087	10.3981
	4	640	6.7025	1.32378	4.0131	10.1843
	5	640	6.7266	1.33364	4.0249	10.3884
Complete data after imputation	1	1800	6.7137	1.33721	4.0000	10.5000
	2	1800	6.6811	1.36622	4.0000	10.5000
	3	1800	6.6916	1.34322	4.0000	10.5000
	4	1800	6.6869	1.36467	4.0000	10.5000
	5	1800	6.7808	1.35792	4.0000	10.5000
Complete data	Pooled	1800	6.7339	1.36063	4.0000	10.5000

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

Appendix N

N.1 (Study 3; Chap. 7): VKS_6 Model

Parameters	Model 1	Model 2[a]	Model 3[a]	Model 4[a]	Model 5	Model 6[a]	Model 7
Fixed effects							
Group							
Time							
Working memory (WM)							
Group * Time							
Group * WM							
Random effects							
Participants (intercept)	0.43*** (0.09)	0.41*** (0.10)	0.30** (0.10)	0.41*** (0.10)	0.56*** (0.11)	0.36*** (0.07)	0.65*** (0.09)
Time | Participants slope		0.66*** (0.16)	0.66*** (0.16)	0.66*** (0.16)			0.81*** (0.10)
Time | Participants (correlation intercept-slope)					−0.28 (0.24)	−0.22 (0.32)	−0.09 (0.31)
Group | Participants slope			– [b]				

(continued)

B. B. Silva, *Writing to Learn Academic Words*, Second Language Learning and Teaching, https://doi.org/10.1007/978-3-031-06505-7

(continued)

Parameters	Model 1	Model 2[a]	Model 3[a]	Model 4[a]	Model 5	Model 6[a]	Model 7
Group | Participants (correlation intercept-slope)							
Items (intercept)	0.54*** (0.07)	0.73*** (0.09)	0.73*** (0.09)	0.47*** (0.09)	0.72*** (0.09)	0.20*** (0.03)	0.73*** (0.09)
Time | Items slope		– [b]					
Time | Items (correlation intercept-slope)						– [b]	
Group | Items slope				– [b]			
Group | Items (correlation intercept-slope)							
Class (intercept)	0.01 (0.04)						
Model summary							
Deviance statistic (AIC corrected)	12,990.54	14,239.78	14,239.80	14,239.81	14,285.81	13,030.80	14,249.34
Covariance structure	VC	VC	VC	VC	AR1	AR1	ARH1
					VC (items)		
Next step:	Remove class intercept. Add time slopes	Remove redundant. Add group | participants slope	Remove redundant. Add group | items slope	Remove redundant. Add time | participants correlation	Add time | items correlation	Remove redundant. Add slope	Back to model 1 without class intercept

Note Parameter estimate standard error listed in parentheses. Number of data points = 1800; items = 20; participants = 90. Probability distribution: multinomial; link function: cumulative negative log–log. VC = Variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite
[a]Error = Hessian matrix is not positive definite
[b]Redundant
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Parameters	Model 8[c]	Model 9	Model 10[e]	Final model	Model 12[a]	Model 13[a,d]
Fixed effects						
Group untimed (GU)	−0.11 (0.19)	−0.11 (0.19)	−0.13 (0.18)	*−0.13 (0.21)*	−0.12 (0.19)	−0.13 (0.22)

(continued)

(continued)

Parameters	Model 8[c]	Model 9	Model 10[e]	Final model	Model 12[a]	Model 13[a,d]
Group timed (GT)	0.14 (0.15)	0.13 (0.15)	0.10 (0.14)	*0.09 (0.20)*	0.08 (0.18)	0.09 (0.21)
Time	−0.78*** (0.05)	−0.79*** (0.05)	−0.83*** (0.05)	*−0.83*** *(0.10)*	−0.69*** (0.10)	−0.81*** (0.09)
Working memory (WM)	−0.001 (0.02)	0.001 (0.02)	−0.004 (0.01)	*−0.004 (0.03)*	−0.003 (0.03)	0.002 (0.03)
GU * Time	−0.14 (0.09)	−0.14 (0.09)	−0.12 (0.10)	*−0.12 (0.18)*	−0.12 (0.17)	−0.12 (0.17)
GT * Time	−0.09 (0.13)	−0.09 (0.14)	−0.05 (0.16)	*−0.05 (0.15)*	−0.04 (0.15)	−0.07 (0.14)
Random effects						
Participants (intercept)	0.52*** (0.12)	0.51*** (0.12)	0.24*** (0.06)	*0.24*** *(0.06)*	0.22*** (0.05)	0.65*** (0.08)
Time | Participants slope		0.05 (0.06)				0.19* (0.08)
Time | Participants (correlation intercept-slope)			0.68* (0.32)	*0.68** *(0.31)*	0.86** (0.28)	– [b]
Group | Participants slope						
Group | Participants (correlation intercept-slope)						
Items (intercept)	0.88*** (0.10)	0.88*** (0.10)	0.87*** (0.10)	*0.87*** *(0.10)*	0.26*** (0.03)	0.88*** (0.10)
Time | Items slope						
Time | Items (correlation intercept-slope)					– [b]	
Group | Items slope						
Group | Items (correlation intercept-slope)						
Class (intercept)	0.02 (0.05)	0.02 (0.05)	0.01 (0.04)			
Model summary						

(continued)

(continued)

Parameters	Model 8[c]	Model 9	Model 10[e]	Final model	Model 12[a]	Model 13[a,d]
Deviance statistic (AIC corrected)	15,750.98	15,717.69	15,105.14	*15,092.14*	13,587.50	15,583.49
Covariance structure	VC	VC	AR1		AR1	ARH1
			VC (items/class)			VC (items)
Next step:	Add time | participants slope	Remove slope. Add time | participants correlation	Remove class intercept	*Add AR1 for items*	Back to model 10. Add time | participants correlation	Back to model 10. Add time | items slope

Note Parameter estimate standard error listed in parentheses. Number of data points = 1800; items = 20; participants = 90. Probability distribution: multinomial; link function: cumulative negative log–log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Redundant

[c]Error = convergence was not achieved

[d]Error = the estimated covariance matrix is not positive definite

[e]Error = unable to compute confidence intervals for estimates

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

N.2 (Study 3): VKS_3 Model

Parameters	Model 1	Model 2[a]	Model 3	Model 4[a]	Model 5[a]	Model 6	Model 7[a]	Model 8
Fixed effects								
Group								
Time								
Working memory (WM)								
Group * Time								
Group * WM								
Random effects								
Participants (intercept)	0.38*** (0.09)	0.31*** (0.08)	0.31*** (0.08)	0.06 (0.08)	0.31*** (0.08)	0.29*** (0.06)	0.26*** (0.06)	0.48*** (0.08)
Time | Participants slope		0.37** (0.12)	0.37** (0.12)	0.37** (0.12)	0.37** (0.12)			0.62*** (0.10)

(continued)

(continued)

Parameters	Model 1	Model 2[a]	Model 3	Model 4[a]	Model 5[a]	Model 6	Model 7[a]	Model 8
Time | Participants (correlation intercept-slope)						0.73 (0.41)	0.94** (0.32)	0.85** (0.35)
Group | Participants slope				– [b]				
Group | Participants (correlation intercept-slope)								
Items (intercept)	0.47*** (0.07)	0.50*** (0.07)	0.50*** (0.07)	0.50*** (0.07)	0.38*** (0.07)	0.52*** (0.07)	0.23*** (0.03)	0.53*** (0.07)
Time | Items slope		– [b]						
Time | Items (correlation intercept-slope)							– [b]	
Group | Items slope					– [b]			
Group | Items (correlation intercept-slope)								
Class (intercept)	0.01 (0.04)							
Model summary								
Deviance statistic (AIC corrected)	7244.20	7324.11	7322.10	7324.11	7324.11	7413.14	7442.19	7445.35
Covariance structure	VC	VC	VC	VC	VC	AR1	AR1	ARH1
						VC (items)		VC (items)
Next step:	Remove class intercept; Add time slopes	Remove redundant	Add group | participants slope	Remove redundant. Add group | items slope	Remove redundant. Add time | participants correlation	Add time | items correlation	Remove redundant. Add slope	Back to model 1

Note Parameter estimate standard error listed in parentheses. Number of data points = 1800; items = 20; participants = 90. Probability distribution: multinomial; link function: cumulative negative log–log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Reference categories (descending) = SW, pretest, and VKS score 1. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Redundant

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Parameters	Final model	Model 10[a,c]	Model 11[a]	Model 12[a]	Model 13[a,c]
Fixed effects					
Group untimed (GU)	*0.13 (0.11)*	0.13 (0.11)	0.13 (0.11)	0.13 (0.19)	0.13 (0.11)
Group timed (GT)	*−0.14 (0.13)*	−0.14 (0.13)	−0.10 (0.12)	−0.12 (0.19)	−0.10 (0.12)
Time	*0.66*** (0.04)*	0.66*** (0.04)	0.70*** (0.05)	0.66*** (0.09)	0.69*** (0.05)
Working memory (WM)	*0.02 (0.02)*	0.02 (0.02)	0.02 (0.02)	0.02 (0.03)	0.02 (0.03)
GU * Time	*0.12* (0.06)*	0.12* (0.06)	0.11 (0.08)	0.12 (0.15)	0.11 (0.07)
GT * Time	*0.007 (0.14)*	0.007 (0.14)	−0.04 (0.17)	0.006 (0.13)	−0.22 (0.13)
Random effects					
Participants (intercept)	*0.42*** (0.10)*	0.42*** (0.09)	0.18*** (0.04)	0.42*** (0.09)	0.56*** (0.08)
Time | Participants slope		– [b]			0.22** (0.08)
Time | Participants (correlation intercept-slope)			– [b]		– [b]
Group | Participants slope					
Group | Participants (correlation intercept-slope)					
Items (intercept)	*0.60*** (0.08)*	0.60*** (0.08)	0.59*** (0.08)	0.60*** (0.08)	0.59*** (0.08)
Time | Items slope					
Time | Items (correlation intercept-slope)				– [b]	
Group | Items slope					
Group | Items (correlation intercept-slope)					
Class (intercept)	*0.02 (0.04)*	0.01 (0.04)	– [b]	– [b]	– [b]

(continued)

(continued)

Parameters	Final model	Model 10[a,c]	Model 11[a]	Model 12[a]	Model 13[a,c]
Model summary					
Deviance statistic (AIC corrected)	*7618.09*	7620.10	7534.80	7615.91	7592.19
Covariance structure	*VC*	VC	AR1	VC	ARH1
			VC (items)	AR1 (items)	VC (items)
Next step:	*Add time | participants slope*	Remove slope. Add time | participants correlation	Remove redundant. Add AR1 for items	Back to model 9. Add time | participants ARH1	Final model is model 9

Note Parameter estimate standard error listed in parentheses. Number of data points = 1800; items = 20; participants = 90. Probability distribution: multinomial; link function: cumulative complementary log–log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Reference categories (descending) = SW, pretest. Reference target: VKS score 3. Degrees of freedom estimation = Satterthwaite
[a]Error = Hessian matrix is not positive definite
[b]Redundant
[c]Error = the estimated covariance matrix is not positive definite
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

N.3 (Study 3): Association Model

Parameters	Model 1	Model 2[a,b]	Model 3	Model 4[a]	Model 5	Model 6	Model 7
Fixed effects							
Intercept	−0.41*** (0.14)	−0.53*** (0.14)	−0.53*** (0.13)	−0.53** (0.14)	−0.48** (0.12)	−0.50** (0.13)	−0.50** (0.13)
Group untimed (GU)							
Group timed (GT)							
Time							
Working memory (WM)							
GU * Time							
GT * Time							
Random effects							

(continued)

(continued)

Parameters	Model 1	Model 2[a,b]	Model 3	Model 4[a]	Model 5	Model 6	Model 7
Participants (intercept)	0.41*** (0.10)	0.38*** (0.10)	0.38*** (0.10)	0.20* (0.10)	0.28*** (0.06)	0.27*** (0.06)	0.63*** (0.08)
Time | Participants slope		0.22*** (0.07)	0.22*** (0.07)	0.22*** (0.07)			0.30*** (0.09)
Time | Participants (correlation intercept-slope)					0.16 (0.30)	0.12 (0.29)	0.24 (0.41)
Group | Participants slope				– [c]			
Group | Participants (correlation intercept-slope)							
Items (intercept)	0.61*** (0.07)	0.61*** (0.06)	0.61*** (0.06)	0.61*** (0.06)	0.61*** (0.06)	0.33*** (0.06)	0.28*** (0.06)
Time | Items slope		– [c]					
Time | Items (correlation intercept-slope)						0.02 (0.23)	0.38 (0.31)
Group | Items slope							
Group | Items (correlation intercept-slope)							
Class (intercept)	0.15 (0.11)	0.13 (0.10)	0.13 (0.10)	0.13 (0.10)	0.12 (0.09)	0.13 (0.09)	0.13 (0.09)
Model summary							
Deviance statistic (AIC corrected)	6267.87	6266.78	6264.77	6266.78	6230.63	6232.46	6256.46
Covariance structure	VC	VC	VC	VC	AR1	AR1	ARH1
					VC (items) VC (class)	VC (class)	AR1 (items) VC (class)
Next step:	Add time slopes	Remove redundant	Add time | participants correlation	Remove redundant. Add time | participants correlation	Add time | items correlation	Add time | participant slope	Add time | items slope

Note Parameter estimate standard error listed in parentheses. Number of data points = 1800; items = 20; participants = 90. Probability distribution: Poisson; link function: log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Predictor reference categories (descending) = SW, pretest. Target reference category = ascending. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite
[b]Error = the estimated covariance matrix is not positive definite
[c]Redundant
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Parameters	Model 8	Model 9	Final model	Model 12[a]	Model 13[a,b]
Fixed effects					
Intercept	−0.59*** (0.13)	−0.46* (0.19)	*−0.39 (0.19)*	−0.41 (0.26)	−0.50* (0.25)
Group untimed (GU)		0.43 (0.23)	*0.44 (0.23)*	0.55* (0.23)	0.47* (0.22)
Group timed (GT)		−0.22 (0.26)	*−0.21 (0.26)*	−0.11 (0.24)	−0.12 (0.23)
Time		0.46*** (0.03)	*0.47*** (0.03)*	0.50*** (0.09)	0.46*** (0.09)
Working memory (WM)		−0.04 (0.023)	*−0.04** (0.02)*	−0.05 (0.03)	−0.03 (0.03)
GU * Time		0.07 (0.13)	*0.09 (0.13)*	0.05 (0.13)	0.08 (0.13)
GT * Time		0.01 (0.06)	*0.01 (0.06)*	0.04 (0.12)	0.01 (0.12)
Random effects					
Participants (intercept)	0.60*** (0.08)	0.24*** (0.06)	*0.25*** (0.06)*	0.72*** (0.07)	0.27*** (0.06)
Time | Participants slope	0.45*** (0.07)			– [c]	
Time | Participants (correlation intercept-slope)	0.24 (0.34)	−0.30 (0.22)	*−0.18 (0.25)*	– [c]	−0.25 (0.21)
Group | Participants slope					
Group | Participants (correlation intercept-slope)					
Items (intercept)	0.78*** (0.04)	0.62*** (0.06)	*0.51*** (0.08)*	0.51*** (0.08)	0.79*** (0.04)
Time | Items slope	– [c]				– [c]
Time | Items (correlation intercept-slope)	– [c]		*−0.53*** (0.11)*	−0.53*** (0.11)	– [c]

(continued)

(continued)

Parameters	Model 8	Model 9	Final model	Model 12[a]	Model 13[a,b]
Group | Items slope					
Group | Items (correlation intercept-slope)					
Class (intercept)	0.12 (0.09)	0.10 (0.08)	*0.10 (0.08)*	0.10 (0.08)	0.10 (0.08)
Model summary					
Deviance statistic (AIC corrected)	6276.71	6238.12	*6200.22*	6224.32	6239.67
Covariance structure	ARH1	AR1	*AR1*	ARH1	AR1
	VC (class)	VC (items) VC (class)	*VC (class)*	AR1 (items)	VC stage | items
Next step:	Back to model 5 (best fit). Add fixed effects		*Add Time | Participants slope*	Back to model 11. Add time | items slope	Remove stage | items

Note Parameter estimate standard error listed in parentheses. Number of data points = 1800; items = 20; participants = 90. Probability distribution: Poisson; link function: log. VC = variance components; AR1 = first-order autoregressive; ARH1 = heterogenous autoregressive. Predictor reference categories (descending) = SW, pretest. Target reference category = ascending. Degrees of freedom estimation = Satterthwaite

[a]Error = Hessian matrix is not positive definite

[b]Error = the estimated covariance matrix is not positive definite

[c]Redundant

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$